Interactions: Do Teacher Behaviors Predict Achievement, Executive Function, and Non-Cognitive Outcomes in Elementary School?

Alejandra Campos*
Pedro Carneiro+
Yyannu Cruz-Aguayo*
Norbert Schady*

August 15, 2020

Abstract

Knowing what teacher characteristics or behaviors consistently predict child outcomes is critical for policy design. We use an experiment with seven consecutive rounds of random assignment and almost perfect compliance to estimate how the quality of teacher-child interactions affects child outcomes in elementary school. We find modest impacts of teacher behaviors on math and language achievement, with somewhat larger effects for children in kindergarten and 1st grade. The impact of teacher behaviors on achievement fades out, but can still be detected seven years later. Better teachers also improve child executive function and reduce the incidence of behavioral problems, but the magnitude of the impacts is small. We do not find clear evidence that better teacher-child interactions improve depression, self-esteem, growth mindset, or grit in a sustained way. Children do not exert more effort in response to differences in teacher quality, and teachers do not exhibit different behaviors when, by chance, they are randomly assigned to children of varying characteristics. We discuss policy implications and avenues for future research.

* Inter-American Development Bank
+University College London

A. Introduction

What makes some teachers more effective than others? Decades of research in economics and other disciplines has established that some teachers have larger impacts on learning than others—even within the same grade and school. Teachers also affect long-term outcomes, including high-school graduation, the probability of going to college, the quality of college attended, and earnings (Chetty et al. 2011; Chetty, Friedman, and Rockoff 2014). However, attempts to identify teacher characteristics that consistently predict their effectiveness have largely been disappointing (Rivkin, Hanushek, and Kain 2005; Staiger and Rockoff 2010).

In this paper, we use data from Ecuador, a middle-income country in South America, to analyze whether in-class teacher behaviors that educators and educational psychologists have argued are good measures of teacher quality predict test scores, executive function, and non-cognitive outcomes in elementary school.[1]

Our results are based on an experiment we carried out in 200 schools. In these schools, children and teachers were randomly assigned to classrooms in kindergarten, and were then randomly reassigned to classrooms in every grade between 1st and 6th grades.[2] There were at least two kindergarten classrooms in every school, and compliance with the random assignment in kindergarten through 6th grade was 98.5 percent, on average.[3] Thus, every child was exposed to seven exogenous, orthogonal shocks to classroom quality.

At the end of each grade we tested children in math and language. Between kindergarten and 4th grade, we also collected data on executive function (EF). EF refers to a set of skills that allow individuals to plan, focus attention, remember instructions, and juggle multiple tasks.[4] It includes working memory, inhibitory control, and cognitive flexibility (Center for the Developing Child 2019). In 1st grade, we asked children how much effort they put into their work in school, and whether they read at home. Finally, we collected data on a rich set of non-cognitive outcomes. In every grade, we asked teachers to list the worst-behaved children in their classroom and, at the end of 6th grade, we collected data on child

[1] One recent review (Pianta, Downer, and Hamre 2017, p. 123) states: "There is a growing consensus that teachers' daily interactions with students are among the most important ways to foster child development in prekindergarten through third grade."

[2] We refer to our assignment as "random" as shorthand, although technically random assignment occurred only in 3rd through 6th grades. In the other grades, the assignment rules were as-good-as-random. We describe the assignment rules below, and provide a number of randomization checks in Appendix A.

[3] Although all 200 schools in the sample had at least two kindergarten classrooms in kindergarten, some schools closed classrooms, or added new ones, over the 7 years of the study. As a result, there were 10, 7, 14, 18, 21, and 15 schools that had only one classroom in 1st, 2nd, 3rd, 4th, 5th, and 6th grade, respectively.

[4] Executive function has both cognitive and non-cognitive dimensions. The frontal lobes of the brain, in particular the prefrontal cortex, play a critical role in EF.

depression, self-esteem, grit, and growth mindset. We note that the outcomes we measure have all been shown to predict educational attainment and wages in adulthood.[5]

In every grade between kindergarten and 4th grade we filmed teachers teaching for at least 4 hours, and coded the resulting video using a rubric known as the Classroom Assessment Scoring System (CLASS; Pianta, LaParo, and Hamre 2007). The CLASS is one the most oft-used classroom observation tools in the U.S., and it shares features with other instruments used by school districts to evaluate teachers. It is not specific to any given subject or grade (although different versions have been developed for children in different grades). Rather, the CLASS focuses on the interactions between teachers and students, in particular on the quality of the *Instructional Support*, *Socio-Emotional Support*, and *Classroom Management* that a teacher provides.[6]

We analyze the impact of being in a classroom taught by a teacher with high CLASS scores on child achievement, executive function, and non-cognitive outcomes. Identification of these effects comes from the random assignment of teachers and children to classrooms within schools.

We first focus on cross-classroom differences in achievement. We show that children randomly assigned to teachers with higher CLASS scores have higher test scores in math and language at the end of that grade. The effect of teacher behaviors on child achievement is essentially unchanged when we control for teacher experience, tenure, and gender. We also find that the CLASS effects are significantly larger in "earlier" grades (kindergarten and 1st grades) than in "later" grades (2nd through 4th grades). Finally, we show that there is substantial fade-out of the effects of teacher behaviors over time.

Our analysis then turns to outcomes other than achievement. We begin by analyzing the effect of teacher behaviors on child executive function. We show that children randomly assigned to higher-CLASS teachers have better EF outcomes at the end of that grade. The effect of being randomly assigned to teachers with better interactions is largest on working memory, and smallest on inhibitory control.

---

[5] There are dozens of studies in economics that show how early achievement predicts outcomes in adulthood, including (but not limited to) those in the labor market. Inhibitory control in early childhood, one of the components of EF, has been shown to predict a variety of outcomes in adulthood, including educational attainment, labor market outcomes, health, and criminal behavior, even after controlling for socioeconomic status in childhood (Moffitt et al. 2011). Classroom misbehavior in 8th grade is negatively associated with own earnings in adulthood (Segal 2013), and children with behavioral problems in elementary school reduce the earnings of their classmates in their mid-20s (Carrell, Hoekstra, and Kuka 2018). Mental health in childhood has been shown to predict socioeconomic status in adulthood in long-term panels (Goodman et al. 2011). Interventions that build growth mindset have been shown to raise academic achievement (Blackwell, Trzesniewski, and Dweck 2007; Yeager et al. 2019), and "grittier" children may be more likely to succeed in school and thereafter (Duckworth and Seligman 2005; Duckworth et al. 2007).

[6] The CLASS implementation guide explicitly states what the CLASS does *not* measure: "The CLASS focuses on the quality of classroom interactional processes. This differs from other measurement tools that focus on the content of the physical environment, available materials, or a specific curriculum. For CLASS, the physical environment (including materials) and curriculum matter in the context of how teachers put them to use in their interactions with children." See Hamre, Goffin, and Kraft-Sayre (2009, p. 5).

Next, we test whether students randomly assigned to teachers who have higher CLASS scores have better non-cognitive outcomes. We show that high-quality teachers appear to reduce the incidence of behavioral problems. We also test whether the CLASS scores of teachers that children were randomly assigned to between kindergarten and 4th grade predict child depression, self-esteem, grit, and growth mindset at the end of 6th grade, the only grade in which these outcomes are available. There is a hint in our data that high- CLASS teachers may improve these outcomes, but our results are imprecise.

Finally, we test whether 1st grade children adjust their effort (trying harder in class, reading at home) in response to differences in teacher quality, and whether teachers exhibit different behaviors in response to random fluctuations in student characteristics (lagged achievement, executive function, or behavioral problems). In neither case do we find evidence of behavioral responses.

In sum, children randomly assigned to teachers with better CLASS scores have higher achievement, better executive function, and a lower incidence of behavioral problems. However, as we discuss below, the effects are very modest in magnitude: They imply that moving a child from the 25th to the 75th percentile of teacher quality raises achievement by 0.05 SDs, increases executive function by 0.04 SDs, and reduces the probability that a child is reported to have behavioral problems by 5 percent. Most of the effects fade out quickly.

Our paper contributes to an understanding of how children acquire skills in elementary school. In discussing this, we stress that random assignment, with essentially perfect compliance, ensures that our results are not confounded by any unobserved characteristics of children. We also underline that, to the best of our knowledge, this is the first paper that uses multiple (seven) rounds of random assignment to assess how the effects of teacher behaviors on achievement, executive function, and non-cognitive outcomes evolve over time.

First, we add to the literature on teacher evaluation. Classroom observations, using standards-based observation protocols, are a central component of how teachers are assessed in most school districts in the U.S,[7] as well as in low- and middle-income countries.[8] We show that teacher behaviors predict child achievement. This finding has important policy implications, in particular for teacher assessment in grades in which the calculation of value added is arguably not feasible at scale (like

[7] Steinberg and Donaldson (2016) review teacher evaluation systems in all 50 U.S. states, the 25 largest school districts, and Washington D.C. They write (pp. 346-47): "We find that the classroom observation score is the most frequently used measure of teacher performance.... Of the 46 states and 23 districts implementing new teacher evaluation systems, all incorporate classroom observation as a component of a teacher's summative evaluation rating. Classroom observation scores also represent the largest share of a teacher's summative rating. Across state and district settings, on average, 54 percent and 52 percent, respectively, of a teacher's rating is based on observation scores". Previous research by economists on classroom practices includes Kane et al. (2011); Kane and Staiger (2012).

[8] Bruns and Luque (2014), and Cruz-Aguayo, Hincapie, and Rodriguez (2020) discuss the evidence from Latin America.

kindergarten).[9] That said, because the CLASS effects we estimate are small in magnitude, evaluating teachers only by this metric would involve considerable mis-classification of effective and ineffective teachers.[10]

Second, we add to the literature on inputs into the production of human capital. Cunha and Heckman (2007) develop a model in which inputs provided earlier have larger effects than those provided later, and different inputs interact. Consistent with their model, we show that the impact on achievement of being randomly assigned to a high- CLASS teacher is larger for children in kindergarten and 1st grade. The fact that children do not change their behaviors (exerting more effort in class, reading more at home) in response to differences in teacher quality, and that teachers do not change their behaviors in response to random fluctuations in child characteristics (lagged achievement, executive function, or behavior problems) also provides insights into the production function of skills in elementary school.[11]

Third, we contribute to a literature on the importance of "soft" or "interpersonal" skills in various occupations.[12] The CLASS is a measure of teacher *behaviors*, not *skills*, but teachers need a variety of interpersonal skills to carry out these behaviors.[13] Building a warm, supportive environment for students, effectively managing time in the classroom, and providing appropriate scaffolding for learning, all of which are scored on the CLASS, require empathy, emotional awareness, and organization, among many other soft skills. As we show, some teachers in our sample are more likely to engage in these behaviors than others. These differences in behaviors have (modest) effects on child outcomes, suggesting that soft or interpersonal skills may be important in teaching, especially for teachers of young children.

[9] To calculate value added, one needs measures of achievement in grades *g* and *g+1*. At older ages, children can be given in-class tests, applied in a group setting, but this is not feasible for young children who have to be tested individually.
[10] To get a sense of this, we calculate classroom value added (as in Araujo et al. 2016), divide the sample into classrooms with above- and below-average value added, and classrooms with teachers with above- and below-average CLASS scores. Forty-five percent of teachers are off-diagonal (above-average by one measure and below-average by the other). When we calculate quintiles of value-added and the CLASS, 12 percent of teachers who are in the highest quintile of the CLASS are in the lowest quintile of value added and, conversely, 13 percent of teachers in the lowest quintile of the CLASS are in the highest quintile of value added.
[11] Our analysis of fade-out complements earlier work using teacher value added with U.S. data (Chetty et al. 2014; Jacob, Lefgren, and Sims 2010).
[12] See, among many references, Heckman and Rubinstein (2001); Heckman, Stixrud, and Urzua (2006); Heckman and Kautz (2012); Kautz et al. (2017).
[13] For a discussion of the relationship between traits, personality, skills, and behaviors see Borghans et al. (2014), Kautz et al. (2017) and the many references therein.

Finally, we add to a small literature on the impact of schools and teachers on outcomes other than test scores.[14] We find some evidence that the in-class behaviors of teachers affect executive function and the incidence of behavioral problems, but these effects fade out quickly. We also find some evidence consistent with CLASS effects on depression, self-esteem, grit, and growth mindset, but our results are too imprecise to draw definitive conclusions about these outcomes.

The rest of the paper proceeds as follows. In Section 2 we describe the setting and our data. We discuss our estimation approach in Section 3, and present results in Section 4. Section 5 concludes.

2. Setting and Data

A. Setting

Ecuador is a middle-income country in South America. Schooling is compulsory from 5 to 14 years of age. The elementary school cycle runs from kindergarten to 6th grade, middle school from 7th through 9th grades, and high school from 10th through 12th grades. There are 3.6 million children in the education system in Ecuador, 80 percent of whom attend public schools; the remaining 20 percent attend private schools. There are more than 150,000 public sector teachers. The teacher salary scale is overwhelmingly determined by seniority.

Ecuador has made considerable progress expanding the coverage of the education system. However, many children, especially the poor, appear to learn little in school. On an international test of 3rd graders, 38.1 percent of children in Ecuador had the lowest of the four levels of performance on math, very similar to the average for the 15 countries in Latin America that participated in the test (39.5 percent), but substantially more than higher-performing countries like Costa Rica (17.6 percent) or Chile (10.0 percent) (Berlinski and Schady 2015). As is the case in many other countries in Latin America, quality, not access, appears to be the key education challenge in Ecuador.

B. Experimental set-up

The multi-grade experiment we conducted included 200 schools in the coastal region of the country.[15] An incoming cohort of children was randomly assigned to kindergarten classrooms within schools in the 2012 school year. These children were reassigned to 1st grade classrooms in 2013, 2nd grade classrooms in 2014, 3rd grade classrooms in 2015, 4th grade classrooms in 2016, 5th grade classrooms in 2017, and 6th

[14] Jackson (2018) and Liu and Loeb (2019) find teacher effects on a variety of student behaviors in the U.S., including absences, suspensions, class grades, and grade repetition. Kraft (2019) estimates teacher effects on grit and growth mindset.

[15] As we discuss in Araujo et al. (2016), these schools are a random sample of all elementary schools in the coast with at least two kindergarten classrooms.

grade classrooms in 2018. Compliance with the assignment rules was very high—98.5 percent on average.

We refer to our assignment as "random" as shorthand, although technically random assignment occurred only in 3rd through 6th grades. In the other grades, the assignment rules were as-good-as-random. Specifically, the assignment rules we implemented were as follows: In kindergarten, all children in each school were ordered by their last name and first name, and were then assigned to teachers in alternating order; in 1st grade, they were ordered by their date of birth, from oldest to youngest, and were then assigned to teachers in alternating order; in 2nd grade, they were divided by gender, ordered by their first name and last name, and then assigned in alternating order; in 3rd through 6th grades, they were divided by gender and then randomly assigned to one or another classroom.

We provide a number of randomization checks in Appendix A. These checks show that random assignment worked as expected: Differences in the baseline characteristics of children assigned to teachers with higher (lower) CLASS scores are very small, and the CLASS scores of teachers that children were assigned to in grades *g*, *g+1* …*g+4* are uncorrelated.

C. Data

i. Teacher data

We use the CLASS (Pianta, LaParo, and Hamre 2007) to measure teacher behaviors. The CLASS measures teacher behaviors in three broad domains: *Emotional Support*, *Classroom Organization*, and *Instructional Support*. Within each of these domains, there are a number of CLASS dimensions.[16] The behaviors that coders are looking for in each dimension are quite specific—Appendix Table B1 gives an example. For each of these behaviors, the CLASS protocol gives coders concrete guidance on whether the score given should be "low" (scores of 1–2), "medium" (3–5), or "high" (6–7).

The CLASS has been widely used both for research and policy purposes in the U.S., especially in preschool settings. For example, in the U.S., Head Start grantees need a minimum score on the CLASS to be re-certified for funding. The CLASS has also been used as a measure of teacher quality in Latin America, including in our earlier work in Ecuador (Araujo et al. 2016), in Chile (Bassi, Meghir, and Reynoso 2019; Yoshikawa et al. 2015) and in Peru (Araujo, Dormal, and Schady 2019).

All teachers in kindergarten through 4th grade were filmed for a full day (from approximately 8 in the morning until 1 in the afternoon); they did not know on what day they would be filmed until the day

[16] Within *Socio-Emotional Support*, these dimensions are positive climate, negative climate, teacher sensitivity, and regard for student perspectives; within *Classroom Organization*, the dimensions are behavior management, productivity, and instructional learning formats; and within *Instructional Support*, they are concept development, quality of feedback, and language modeling.

itself. We closely followed CLASS protocols to code the film. Specifically, each day of video recording was cut into 20-minute segments, and each segment was coded by two separate coders.

Figure 1 graphs univariate densities of the distribution of CLASS scores, by domain.[17] The figure shows that CLASS scores are on average highest in *Classroom Organization*, with teachers distributed in the "medium" and "high" parts of the distribution; somewhat lower in *Socio-Emotional Support*, with most teachers in the "medium" range; and lowest in *Instructional Support*, where all teachers have "low" CLASS scores.

Table 1, Panel A, shows that differences in mean CLASS scores by grade are small. However, the variance of CLASS scores decreases substantially by grade—it is more than twice as high in kindergarten as in 4th grade, a point we return to below. Table 1 also summarizes other characteristics of classrooms and teachers. Average class size is 38. The average teacher in the sample has 18 years of experience, and only a very small proportion, 4 percent, are "rookie" teachers with 3 years of experience or less.[18] Eighty-seven percent of teachers are women, and 78 percent are tenured.[19]

ii. Child data

Table 2 summarizes the baseline characteristics of children. Children were 5 years of age on the first day of school, and half of them are girls. Mothers were in their early 30s and fathers in their mid-30s. Both parents had on average just under 9 years of schooling, which corresponds to completed middle school. The average receptive vocabulary score of children in the sample is 1.7 SDs below the level of children that were used to norm the sample for the test.[20]

We collected data on math and language achievement at the end of each grade between kindergarten and 6th grade. For both subjects, tests were a mixture of material that teachers were meant

[17] In Araujo et al. (2016) we show that the CLASS scores of kindergarten teachers in our sample are comparable to the CLASS scores in a nationally-representative sample of schools in Ecuador, but are substantially lower than those generally observed in the U.S.

[18] We do not know why the number of rookie teachers is so small. We are not aware of any government policy to freeze teacher hiring in the years preceding our experiment. The fact that our sample is mainly drawn from urban areas (because every school had to have at least two teachers per grade to be included) may account for the skewed distribution of experience if more senior teachers have a preference to be in urban schools and have some choice in their school assignment.

[19] The proportion of teachers who are males increases substantially by grade, from 1 percent in kindergarten to 27 percent in 3rd and 4th grades, and the proportion tenured increases, from 64 percent in kindergarten to percentages in the mid- to high-80s in 2nd, 3rd, and 4th grades. The differences in proportion tenured reflect differences by grade but also secular increases in tenure, as a result of a deliberate government policy in Ecuador.

[20] To measure baseline receptive vocabulary, we use the *Test de Vocabulario en Imagenes Peabody* (TVIP) (Dunn et al 1986), the Spanish-speaking version of the much-used Peabody Picture Vocabulary Test (PPVT). The TVIP was normed on samples of Mexican and Puerto Rican children. It has been used widely to measure development among Latin American children. See Paxson and Schady (2007) for a comparison of vocabulary scores between children in Ecuador and the U.S., and Schady et al. (2015) for evidence on levels and socioeconomic gradients in the TVIP in five Latin American countries, including Ecuador.

to have explicitly covered in class—for example, in math, addition or subtraction; material that would have been covered, but probably in a somewhat different format—for example, simple word problems; and material that would not have been covered at all in class but that has been shown to predict current and future math achievement—for example, the Siegler number line task.[21] We aggregated responses in math and, separately, language, by Item Response Theory (IRT), and calculated an average achievement score that gives the same weight to math and language.[22]

In every grade between kindergarten and 4th grade, we tested child executive function. EF is generally thought of as having three domains: working memory, inhibitory control, and cognitive flexibility. Working memory measures the ability to retain and manipulate information; for example, 2nd grade child were asked to remember (increasingly long) strings of numbers and repeat them in order and then backwards. Cognitive flexibility measures the ability to shift attention between tasks and adapt to different rules; for example, 1st grade children were shown picture cards that had trucks or stars, red or blue, and were asked to first sort cards by *shape* (trucks versus stars), and then by *color* (red versus blue). Inhibitory control refers to the capacity to suppress impulsive responses; for example, kindergarten children were quickly shown a series of flash cards that had either a sun or a moon and were asked to say the word "day" when they saw the moon and "night" when they saw the sun. We calculate scores on each of the three domains in executive function, as well as an overall EF score.[23]

At the end of each grade we asked teachers who were the worst-behaved children in their classroom. Teachers could list up to five children, in order; on average, they listed four.[24] In 1st grade, we asked children whether they tried as hard as possible in class, and whether they read at home; both questions gave children the option of answering "always", "sometimes", or "never". In 6th grade, we collected data on child depression, self-esteem, growth mindset, and grit. To measure child depression, we used the Patient-Reported Outcomes Measurement Information System (PROMIS) Depression Scale

[21] The number line task works as follows. Children are shown a line with the two endpoints clearly marked—for example, in 1st grade, the left end of the line is marked with a 0, and the right end is marked with a 20. They are then asked to place various numbers on the line—for example, the number 2 or the number 18. The accuracy with which children place the numbers has been shown to be predictive of general math achievement (see Siegler and Booth 2004).
[22] Our results are very similar if, instead, we calculate a simple sum of correct responses within blocks of questions on the test, and give equal weight to each of these test sections (as in Araujo et al. 2016).
[23] Unlike test scores on math and language, it does not make sense to aggregate questions by factor analysis because some of the tests are timed. For example, in one task, children are given 2 minutes to find as many sequences of dog, house, and ball, in that exact order, on a sheet that has rows of dogs, houses, and balls in various possible sequences. The score on this test is the number of correct sequences found by the child. We calculate separate scores for working memory, inhibitory control, and cognitive flexibility, and a total EF score which gives the same weight to each dimension.
[24] We also asked teachers whether these children carried out specific disruptive behaviors, and with what frequency. In practice, in our analysis, we only use information on whether a child was said to have behavioral problems by his teacher because neither the information on the rank nor the frequency of certain behaviors adds to the explanatory power of the results we report below.

for children aged 11-17 years, developed by the American Psychiatric Association.[25] To measure self-esteem, we selected 5 questions from the National Longitudinal Study of Adolescent to Adult Health (Add Health).[26] To measure Growth Mindset, we selected 10 of the 20 questions on the Dweck "Mindset Quiz"; growth mindset refers to the belief that intelligence is malleable, rather than fixed, and can be increased with effort (Blackwell, Trzesniewski, and Dweck 2007; Dweck 2008). Finally, to measure grit, we adapted 4 questions from the 8-item Grit Scale for children (Duckworth and Quinn 2009); grit refers to the capacity of individuals to persevere at a given task. For each of these 6th grade outcomes, we aggregated responses by factor analysis. We also calculate an overall non-cognitive score that gives the same weight to each of the individual tests.

Most of the tests were applied to children individually (as opposed to in a group setting) by specially trained enumerators.[27] All tests, other than the non-cognitive tests applied in 6th grade, were applied in school. In all tests, to choose questions, we piloted the test; made changes as necessary; and selected questions that could be understood by children in our context, and which showed reasonable levels of variability in the pilot. In Appendix C, we present univariate densities of the different tests, as well as evidence on differences in outcomes by socioeconomic status and gender.

D. Cross-classroom differences

Identification in our experiment relies on there being meaningful variation in outcomes across classrooms within schools. To provide evidence of this, we first show differences in outcomes between pairs of classrooms in the same grade and school.[28] Table 3, Panel A, shows that the median difference in the CLASS is 0.18 points. This panel also shows that, consistent with the results in Table 1, cross-classroom differences are largest in kindergarten (0.22 points), and smallest in 4th grade (0.11 points). Finally, cross-classroom differences in the CLASS are largest for *Classroom Management* (0.29 points), somewhat smaller for *Socio-Emotional Support* (0.22 points), and much smaller in *Instructional Support* (0.08 points).

Panel B focuses on differences in child outcomes. It shows that the median difference across classrooms in the same grade and school is 0.18 SDs for math achievement, 0.16 SDs for language

[25] Olino et al (2013), Klein et al. (2005), and Aylward et al. (2008) argue that the PROMIS depression scale has superior qualities (greater precision, more internal reliability, and more discriminant validity) than other commonly-used depression scales, including the Beck Depression Inventory (BDI), the Children's Depression Inventory (CDI), and the Center for Epidemiologic Studies-Depression (CES-D) scale.

[26] See Harris and Udry 2018 for a description of the Add Health data. The questions on self-esteem in Add Health build on the much-used Rosenberg Self-Esteem Scale (Rosenberg 1989).

[27] The only exception is some of the achievement tests in 4th through 6th grades, which were applied in a group setting.

[28] When there are more than two classrooms in a grade and school, we select two at random.

achievement, 0.16 SDs for executive function, and 0.21 SDs for the composite measure of the non-cognitive outcomes that were collected in 6th grade.

Next, in Table 4, we carry out a variance decomposition of the CLASS and child outcomes, stacking observations when data are available for more than one grade. The first row of the table shows that 58 percent of the variance in the CLASS is explained by cross-school differences, with the remaining 42 percent accounted for by differences across classrooms within the same school. Turning to test scores, we find that 10 and 11 percent of the total variance in math and language achievement, respectively, is explained by differences across schools. The incremental contribution of classrooms within schools is modest, about 2 percent. The last column, finally, shows that child fixed effects explain about three-quarters of the variation in math and language test scores in our sample.

Similar patterns are apparent for executive function, where schools explain 7 percent of the variance, classrooms explain an additional 2 percent, and child fixed effects explain 60 percent of the total variation. The last rows of Table 4 focus on the non-cognitive outcomes collected in 6th grade. These results show that anywhere between 4 percent (for grit) and 7 percent (for depression) of the variation in outcomes is explained by differences across schools. Interestingly, classrooms explain a somewhat larger proportion of the variance of these outcomes, between 3 and 4 percent, than is the case with the measures of achievement and executive function.[29]

In sum, Tables 3 and 4 show that there is modest variation in child outcomes across classrooms within the same school. In the rest of the paper, we investigate whether these cross-classroom differences in achievement, executive function, and non-cognitive outcomes can be explained, at least in part, by differences in the behaviors of teachers.

3. Estimation Strategy

Our basic estimation approach involves running regressions of the following form:

$$Y_{sgtj} = \delta_{sg} + \rho CLASS_{sgt} + \theta T_{sgt} + \gamma X_{sgtj} + u_{sgtj} \quad (1)$$

where Y_{sgtj} is an outcome for child *j* randomly assigned to a classroom taught by teacher *t* in grade *g* and school *s*; δ_{sg} is a full set of school indicators when we run regressions separately by grade, and school-by-grade indicators in those regressions that pool observations from all grades; $CLASS_{sgt}$ is a parametrization of the CLASS score of teacher *t* in grade *g* and school *s*; T_{sgt} are controls for teacher

[29] Unsurprisingly, there is also a fixed, child-level component to behavior: children who are reported to have behavior problems in one grade are also more likely to be reported to have them in the following grade. Specifically, the correlation between behavior problems in grades *g* and *g+1* is 0.39. To give a sense of magnitude, note that roughly 10 percent of children are reported to have behavioral problems in any given grade, so, if these probabilities were not correlated, we would expect that 0.01 of children would be reported as having behavioral problems in both grades.

experience, gender and contract status (tenured or not), which we include in some specifications; X_{sgtj} are controls for child gender, age, and its square; and u_{sgtj} is the regression residual. Standard errors are clustered at the school (or school-by-grade) level throughout.

In practice, a number of estimation questions arise, and we discuss each of these in turn. First, one must decide how to parametrize the CLASS. A standard approach would be to convert the raw CLASS scores, on the 7-point scale, into grade-specific z-scores, by subtracting the grade-specific mean and dividing by the grade-specific standard deviation. However, as we show above, there is much more variability in the CLASS in earlier than in later grades. As a result, a one-unit (one-SD) difference in the CLASS would imply differences in teacher behaviors of a larger magnitude in some grades than in others.

We therefore work with two parametrizations of the CLASS. In one approach, we simply define a teacher as "high-quality" (or above-average) if her CLASS score is above the mean for her school and grade. This formulation is natural given that in most school-grade combinations in our sample (68 percent), there are exactly two classrooms per grade. In the other parametrization we use a continuous measure of the CLASS on the 1-7-point scale.[30] As we show below, results are generally similar with both approaches.

Another question that arises is what child controls should be included in X_{sgtj}. Given random assignment, (any) controls are in principle unnecessary. In practice, we include only child gender, age, and its square, because these are the only controls that are available for all children in every grade. To see why this is so, note that lagged test scores in *g-1*, *g-2* … *g-n*, would not be available for children who transferred into our sample of schools in grade *g*, and baseline characteristics would only be available for children who remained in our sample of schools since the beginning of kindergarten. Thus, more controls come at the expense of fewer observations and, if transfers in and out of schools are non-random, estimation on a sample of children that is not representative of all children in a given classroom.

To make inference on comparisons of the impacts of the CLASS in different grades, and to estimate the fade-out of CLASS effects over time, we run a system of (seemingly unrelated) regressions of the following form:

$$Y_{sgtj} = \delta_{sg} + \rho CLASS_{sktj} + \gamma X_{sgtj} + u_{sgtj} \quad (2)$$

[30] It is not ex-ante clear which of these two alternatives is preferable: using a binary indicator throws out information but, if there is measurement error in the CLASS, as is almost certainly the case, treating the CLASS as an ordinal, rather than a cardinal score, may help. Measurement error in the CLASS arises both because of coder error, and because the behaviors teachers exhibit at any time are not a perfect measure of the behaviors they engage in over the course of a day—leave alone over the course of a school year. Further details on the process of CLASS filming and coding, and a discussion of measurement error, are given in Appendix B.

where $k=2012…2016$, and $k \leq t$. There are 5 of these equations for $t=\{2018, 2017, 2016\}$ (and $k=2012…2016$), 4 equations for $t=2015$ (and $k=2012…2015$), 3 equations for $t=2014$, 2 equations for $t=2013$, and one equation for $t=2012$. All 25 equations are estimated simultaneously.[31]

Finally, we note that while our identification is based on the random assignment of teachers and children to classrooms within schools, with essentially perfect compliance, there could be teacher attributes, unmeasured but correlated with the behaviors measured in the CLASS, that complicate interpretation of the coefficient on the CLASS. We partly address this by including in some specifications regressors for teacher characteristics that are generally available in administrative data—gender, experience, and tenure—and seeing whether these controls meaningfully change the coefficient on the CLASS.

4. Results

A. CLASS effects on achievement

To motivate our analysis, we start with a simple figure. For this purpose, we first calculate deciles of the cross-classroom differences in the CLASS from Table 3, and then relate these differences to cross-classroom differences in achievement.[32] Results are in Figure 2, for math (Panel A) and language (Panel B). The scatterplots and regression lines show there is a positive association between the CLASS, on the one hand, and child achievement in math and language, on the other. This relationship appears to be stronger for math, where the regression line is steeper, and the points are closer to their predicted values.

Next, in Table 5, we report the results from estimates of (1) above. The table shows that having a teacher with an above-average CLASS score increases overall achievement by 0.044 SDs, while a 1-point increase in the CLASS raises test scores by 0.18 SDs. The implied effects are small: to put the magnitude in context, we note that a 1-point increase in the CLASS is equivalent to moving a teacher from the 3rd to the 97th percentile in our data. Table 5 also shows that effects are somewhat larger for

[31] Note also that, because some children transfer in and out of our sample of schools, the sample sizes in these regressions are smaller than those when estimating (1), especially as the number of periods over which we estimate fade-out increases. For example, for kindergarten, we can estimate the contemporaneous effect of the CLASS, as well as effects 1, 2 … 6 periods later. To estimate these effects, we work with a sample of children who were enrolled in our sample of schools in every grade between kindergarten and 6th grade. On the other hand, for 4th grade we can only estimate the contemporaneous effect of the CLASS and the effect 1 and 2 periods later. In this case, we work with a sample of children who were enrolled in our sample of schools in 4th, 5th, and 6th grades.

[32] To generate the figures, in every school and grade, we take two classrooms, label one as classroom A and the other as classroom B, and calculate the differences A-B in the CLASS and achievement. When there are more than 2 classrooms in a school and grade, we select two at random. We have 200 schools and 5 years of data, so at the end of this process we have approximately 1,000 differences. We sort observations into deciles of the cross-classroom difference in the CLASS, and calculate the median difference in the CLASS and achievement in each decile. Finally, we plot the average differences in achievement (vertical axis) as a function of the differences in the CLASS (horizontal axis) for each decile, and include a regression line for the 10 points.

math than language.[33] Teacher experience and gender do not predict achievement, but children taught by tenured (as opposed to contract) teachers have higher test scores (coefficient of 0.031, with a standard error of 0.016).[34] Importantly, adding these teacher controls to the regression has no effect on the coefficient on the CLASS.

In Table 6, we disaggregate estimates by grade. CLASS effects on overall achievement are largest in kindergarten and 1st grades, 0.26 SDs on average. We also report the p-values from two F-tests. First, we test whether the coefficients in all grades are the same; the p-value on this test is 0.06 when the CLASS is parametrized as a binary variable, 0.08 when it is parametrized as a continuous variable. Second, we test whether the effect of the CLASS in "earlier" grades (kindergarten and 1st grades) and "later" grades (2nd through 4th grades) is the same; the p-values on these tests are 0.02 and 0.01, respectively, for the two parametrizations of the CLASS.

Next, in Table 7, we analyze fade-out. A comparison of the two panels in the table shows that the *magnitude* of fade-out is sensitive to how the CLASS is parametrized, especially for kindergarten. In both panels, there appears to be substantial fade-out in the first period. With the binary parametrization of the CLASS, however, 81 percent of the original kindergarten effect is still apparent in 6th grade, while this value is much lower, 43 percent, with the specification in which the CLASS is treated as a continuous variable. Regardless of how we parametrize the CLASS, the table shows that only the kindergarten effect (not the 1st and 2nd grade effects) is significant after four years, and only the kindergarten effect (not the 1st grade effect) is significant five years later. The table also shows that, as the number of lags increases, we are more likely to reject the null hypothesis that the coefficients for all grades (for a given number of lags) are the same, and reject the null that the coefficients on "earlier" and "later" grades are the same—especially with the binary parametrization of the CLASS.

In sum, the CLASS predicts child achievement, and the estimated CLASS effects on achievement is unchanged when we include teacher controls that are generally available in administrative data. The effects of teacher quality, as measured by the CLASS, are modest overall, but somewhat larger in kindergarten and 1st grades. There is fade-out, but the estimate of how much of the original effect of the CLASS on achievement remains over time is sensitive to estimation choices. Regardless, children who were randomly assigned to kindergarten teachers with higher-quality interactions continue to have higher test scores in 6th grade.

[33] We note that this pattern of effects—larger impacts of school-based interventions on math than on language—is not uncommon in the literature (see the discussion in Fryer 2017).

[34] We also find no effect of experience when, instead of a continuous measure of experience, we use a variable for "rookie" teachers (teachers with three years of experience or less, as in Araujo et al. 2016), or when we define rookie teachers as those in the lowest quintile of experience in a school.

B. CLASS effects on executive function and non-cognitive outcomes

We now turn to possible effects of being randomly assigned to a high-quality teacher on executive function, the probability that a child is reported as having behavioral problems, and on depression, self-esteem, growth mindset and grit.

As with the results on achievement, we begin with some simple figures. Figure 3 shows that, overall, children who were randomly assigned to teachers with higher CLASS scores have higher levels of executive function and better non-cognitive outcomes. However, compared to the figure for achievement, the regression lines in both panels are flatter, and the observed values are further from their predicted values—especially for the non-cognitive outcomes collected in 6th grade.

Table 8 reports CLASS effects on executive function from equation (1) above, separately for each dimension (attention, cognitive flexibility, and inhibitory control), as well as for an aggregate measure of EF that gives the same weight to each dimension. Panel A shows that having a teacher with an above-average CLASS score increases executive function by 0.034 SDs, while a 1-point increase in the CLASS raises EF by 0.13 SDs. Disaggregating by EF dimension, we find that the CLASS effects are largest for working memory (0.12 SDs, with the continuous parametrization), somewhat smaller for cognitive flexibility (0.08 SDs), and smallest for inhibitory control (0.06 SDs).[35] Other teacher characteristics do not consistently predict executive function and, much as with achievement, including them in the regression does not appreciably change the coefficient on the CLASS.

In Table 9, we test whether the CLASS scores of teachers in grade *g-1* reduce the probability that a child is reported to have behavior problems by his teachers in grades *g* and *g+1*. The table shows that teachers with higher CLASS scores make it less likely that children have behavioral issues in the following grade. However, the effect is very small, and fades out and is not significant by the following grade.[36]

Table 10, finally, reports CLASS effects on depression, self-esteem, growth mindset and grit in 6th grade, the only grade that these data are available. In the binary parametrization of the CLASS, we regress these variables on the number of above-average teachers between kindergarten and 4th grade. We also run regressions in which the omitted category is children who had exactly 2 or 3 high- CLASS teachers (roughly 60 percent of the sample), and report the coefficients for children who had 0 or 1, and

[35] The differences in CLASS effects by EF dimension may reflect true, underlying differences in impacts, or issues related to measurement. In Appendix C, we show that working memory scores are more normally distributed than is the case for cognitive flexibility and inhibitory control, especially in earlier grades.

[36] To get a sense of the magnitude, note that the median teacher indicated that 4 children in her class had behavioral problems, and average class size is 38. Thus, the probability that a child selected at random is reported as having a behavioral problem is roughly 0.1. Having a teacher with an above-average CLASS score (using the binary parametrization of the CLASS) in the previous grade reduces this probability by 0.005, to 0.095.

4 or 5 high- CLASS teachers, respectively.[37] In the specification that uses the continuous measure of the CLASS, we average the CLASS scores of the five teachers a child was randomly assigned to between kindergarten and 4th grades.

There is a hint in Table 10 that children who were randomly assigned to more high-quality teachers between kindergarten and 4th grade may have better non-cognitive outcomes at the end of 6th grade, but only with the binary parametrization of the CLASS. In the measure that aggregates all four outcomes (depression, self-esteem, growth mindset, and grit), the difference between children with 0 or 1 above-average CLASS teachers, on the one hand, and those with 4 or 5 above-average teachers, on the other, is 0.064 SDs, and this difference is significant (p-value: 0.05). With other parametrizations, however, the picture is less clear.[38]

C. Child and teacher responses to random variations in quality

We now turn to possible behavioral responses by children or teachers. In Table 11, we report results from regressions of in-class effort or reading at home on the CLASS. We use two specifications for the effort and reading regressions: in one, the dependent variable takes on the value of one for children who say they always try as hard as they can (always read at home), zero for those who reply "sometimes" or "never"; in the other regression, which we run using an Ordered Probit, the dependent variable can take the values of 0, 1 or 2, corresponding to the response categories. Table 11 shows that, no matter how we define the dependent variables, and regardless of which parametrization of the CLASS we use, there is no evidence that children try harder in class or read more at home when they are randomly assigned to teachers with higher CLASS scores.

In Table 12, we analyze possible differences in teacher behaviors, as measured by the CLASS, in response to random fluctuations in classroom composition, Specifically, we regress *lagged* achievement, *lagged* executive function, or whether a child was reported as having behavioral issues in the *previous* year on the current CLASS. If the CLASS were purely a measure of some intrinsic teacher quality—purely a teacher trait, rather than a measure of the quality of the interactions between two sets of agents, teachers and students—then these could be seen as tests of random assignment. However, since it is possible that teachers adapt their behavior to the students they teach, and since the quality of the interactions captured

[37] There are 2^5 possible sequences of above- and below-average teachers between kindergarten and 4th grade. One sequence corresponds to having an uninterrupted string of above-average teachers, and another sequence corresponds to having an uninterrupted string of below-average teachers; 5 sequences correspond to having exactly 1 above-average teacher, and 5 correspond to having exactly 4 above-average teachers; in 10 sequences, children had 2 above-average teachers, and in another 10 sequences they had 3 above-average teachers by the end of 4th grade.

[38] In Appendix D we show that the CLASS effects on EF fade out quickly, and do not vary by grade. In this appendix we also show there is no evidence that the CLASS effects on non-cognitive outcomes vary by grade, although it is important to keep in mind that the CLASS in these estimations refers to grades that are further in the past in some cases—for example, kindergarten—than in others—for example, 4th grade.

by the CLASS can potentially reflect attributes of both teachers and students, this is also a test of whether teacher behaviors react to student characteristics. In practice, all the coefficients in the table are very small, and none are anywhere near conventional levels of significance. Thus, teacher behaviors do not appear to respond to differences in the skills of children randomly assigned to their classrooms—at least, over the relatively modest amount of within-school, cross-classroom variation in student characteristics there is in our data.

In sum, children do not change their behaviors when they are assigned to better teachers, and teachers do not change their behaviors when they are assigned to a classroom with children who have higher (or lower) skills. We conclude that the CLASS effects we report are likely to be primarily the direct effect of exogenous differences in teacher behaviors on child outcomes, rather than behavioral responses made by children, parents (as we show in Araujo et al. 2016), or teachers.[39]

5. Conclusion

In this paper, we analyze how the in-class behaviors of teachers, as measured by a widely-used classroom observation tool, affect achievement, executive function, and non-cognitive outcomes in elementary school. Our set-up is unique, with identification based on seven consecutive rounds of random assignment in 200 schools, with almost perfect compliance.

We find that the in-class behaviors of teachers predict achievement, although the effect sizes are small—moving a child from the 25th to the 75th percentile of teacher quality raises achievement by 0.05 SDs. Impacts are larger for teachers of young children (kindergarten and 1st grades) than for those of somewhat older children (2nd through 4th grades). The effect of teacher behaviors on outcomes other than achievement, including executive function, the incidence of behavior problems, depression, self-esteem, growth mindset, and grit is weaker, although there may be considerations of measurement error and power. This is an area where, we believe, future research is important.

We also show that children do not appear to respond to better teachers by working harder, at school or at home, and that teachers do not exhibit differences in their in-class behaviors when they are randomly assigned to children who have higher achievement, better executive function, or fewer behavioral problems.

We close with some general thoughts about policy implications. First, because classroom observations are an integral part of most teacher evaluation systems, our results are encouraging. This is

[39] In our earlier work (Araujo et al. 2016) we showed that, in kindergarten, parents give higher scores on a 1-5 scale to teachers with higher value added or CLASS scores, but do not change their inputs into child development and learning (including the availability of books, pencils, and toys of various kinds) or behaviors (including whether they read to, sang to, or played with their kindergarten children) in response to differences in teacher quality.

especially the case for children in the earliest grades, like kindergarten, where other means of evaluating teachers, like value added, are probably not realistic at scale. That said, the effect sizes we estimate are quite modest, so evaluation with an instrument like the CLASS would involve a great deal of mis-classification of effective and ineffective teachers, regardless of whether by "effectiveness" one means impacts on achievement or other outcomes, both in the short- and particularly in the medium-run.[40]

Second, the fact that the largest effects are found in kindergarten and 1st grade is important. Our research does not answer the question whether teacher quality, in general, is more important for younger than for older children, although some have argued that the returns to skill acquisition are generally larger at younger ages (Heckman 2013). However, our results suggest that assigning teachers who exhibit the behaviors we measure to kindergarten and 1st grade, or to try to build soft skills particularly among these teachers, may make sense.

Finally, we stress the importance of following the children from this (and other) large-scale experiment over time, ideally into adulthood. Although there is some evidence of the long-term effects of teachers in developed countries, especially the U.S., much less is known about this in poorer settings like the one we study.

[40] For this and other reasons, a number of papers have argued that it may be best to evaluate teachers by combining different metrics, including (but not limited to) classroom observation. See Kane and Staiger (2012) and Rockoff et al. (2011), among many references for the U.S. We also note that the classroom observation tool we use is not easy to apply and code reliably. There would be high returns to research that seeks to understand whether simpler observation tools have predictive power, and how best to combine these with other instruments to evaluate teachers.

References

Almlund, Mathilde, Angela L. Duckworth, James Heckman, and Tim Kautz. 2011. "Personality Psychology and Economics." In *Handbook of the Economics of Education*, Vol. 4, E. Hanushek, S. Machin, and L. Woessman, eds. (Oxford: Elsevier).

Araujo, M. Caridad, Pedro Carneiro, Yyannu Cruz-Aguayo, and Norbert Schady. 2016. "Teacher Quality and Learning Outcomes in Kindergarten." *Quarterly Journal of Economics* 131(3): 1415-53.

Araujo, M. Caridad, Marta Dormal, and Norbert Schady. 2019. "Child Care Quality and Child Development." *Journal of Human Resources* 54(3): 656-82.

Aylward, Glen P., and Terry Stancin. 2008. "Measurement and Psychometric Considerations." In Wolraich, Mark L., Dennis D. Drotar, Paul H. Dworkin, and Ellen C. Perrin, Eds., *Developmental-Behavioral Pediatrics*. Philadelphia: Mosby.

Bassi, Marina, Costas Meghir, and Ana Reynoso. 2019. "Education Quality and Teaching Practices." Cowles Foundation Discussion Paper 2181.

Berlinski, Samuel, and Norbert Schady. 2015. *The Early Years: Child Well-Being and the Role of Public Policy*. New York: Palgrave Macmillan.

Blackwell, Lisa S., Kali H. Trzesniewski, and Carol S. Dweck. 2007. "Implicit Theories of Intelligence Predict Achievement Across an Adolescent Transition: A Longitudinal Study and an Intervention." *Child Development* 78(1): 246–63.

Borghans, Lex, Angela Lee Duckworth, James J. Heckman, and Bas ter Weel. 2008. "The Economics and Psychology of Personality Traits." *Journal of Human Resources* 43(4): 972-1059.

Bruns, Barbara, and Javier Luque. 2014. *Great Teachers: How to Raise Student Learning in Latin America and the Caribbean*. Washington, D.C.: World Bank.

Carrell, Scott, Mark Hoekstra, and Elira Kuka. 2018. "The Long-Run Effects of Disruptive Peers." *American Economic Review* 108(11): 3377-3415.

Center for the Developing Child. 2020. "Executive Function & Self-Regulation." Available at https://developingchild.harvard.edu/science/key-concepts/executive-function/. Accessed on March 15, 2020.

Chetty, Raj, John Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Schanzenbach, and Danny Yagan. 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR." *Quarterly Journal of Economics* 126(4): 1593–1660.

Chetty, Raj, John Friedman, and Jonah Rockoff. 2014. "Measuring the Impact of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review* 104(9): 2633–79.

Cruz-Aguayo, Yyannu, Diana Hincapie, and Catherine Rodriguez. 2020. *Profesores a Prueba: Claves para una Evaluación Docente Exitosa*. (Washington, D.C.: Inter-American Development Bank).

Cunha, Flavio, and James Heckman. 2007. "The Technology of Skill Formation." *American Economic Review, Papers and Proceedings* 97(2): 31-47.

Duckworth, Angela, and Martin Seligman, 2005. "Self-Discipline Outdoes IQ in Predicting Academic Performance in Adolescence." *Psychological Science* 16(12): 939-44.

Duckworth, Angela, and Patrick D. Quinn. 2009. "Development and Validation of the Short Grit Scale (GritS)." *Journal of Personality Assessment* 91(2): 166-174.

Dunn, Lloyd, Delia Lugo, Eligio Padilla, and Leota Dunn. 1986. *Test de Vocabulario en Imagenes Peabody.* (Circle Pines, MN: American Guidance Service).

Dweck, Carol. 2008. *Mindset: The New Psychology of Success.* (Ballantine Books: New York).

Fryer, Roland G. 2017. "The Production of Human Capital in Developed Countries: Evidence from 196 Randomized Field Experiments." In: *Handbook of Field Experiments*, Vol. 2, Esther Duflo and Abhijit Banerjee, eds. (Amsterdam: North Holland).

Goodman, A., and J. Smith. 2011. "The Long Shadow Cast by Childhood Physical and Mental Problems on Adult Life." *Proceedings of the National Academy of Sciences* 108: 6032-37.

Hamre, Bridget, Stacie C. Goffin, and Marcia Kraft-Sayre. 2009. *Classroom Assessment Scoring System (CLASS) Implementation Guide: Measuring and Improving Classrooms in Early Childhood Settings.* Available at https://www.vbgrowsmart.com/providers/Documents/CLASSImplementationGuide.pdf , accessed on May 30, 2020.

Harris, Kathleen Mullan, and Richard J. Udry. 2018. *National Longitudinal Study of Adolescent to Adult Health (Add Health), 1994-2008* [Public Use]. Carolina Population Center, University of North Carolina-Chapel Hill, Inter-University Consortium for Political and Social Research.

Heckman, James. 2013. *Giving Kids a Fair Chance.* (MIT Press: Cambridge, MA).

Heckman, James, and Tim Kautz. 2012. "Hard Evidence on Soft Skills." *Labour Economics* 19(4): 451-74.

Heckman, James, and Yona Rubinstein. 2001. "The Importance of Noncognitive Skills: Lessons from the GED Testing Program." *American Economic Review, Papers and Proceedings* 91(2): 145-59.

Heckman, James, Jora Stixrud, and Sergio Urzua. 2006. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *Journal of Labor Economics* 24(3): 411-82.

Jackson, Kirabo. 2018. "What Do Test Scores Miss? The Importance of Teacher Effects on Non-Test Score Outcomes." *Journal of Political Economy* 126(5): 2072-2107.

Jacob, Brian, Lars Lefgren, and David P. Sims. 2010, "The Persistence of Teacher-Induced Learning." *Journal of Human Resources* 45(4): 915-43.

Kane, Thomas, and Douglas Staiger. 2012. *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains* (Seattle: Bill and Melinda Gates Foundation).

Kane, Thomas, Eric Taylor, John Tyler, and Amy Wooten. 2011. "Identifying Effective Classroom Practices Using Student Achievement Data." *Journal of Human Resources* 46(3): 587–613.

Kautz, Tim, James Heckman, Ron Diris, Bas ter Weel, and Lex Borghans. 2017. "Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success." NBER Working Paper 20749.

Klein, Daniel N., Lea R. Dougherty, and Thomas M. Olino. 2005. "Toward Guidelines for Evidence-Based Assessment of Depression in Children and Adolescents." *Journal of Clinical Child and Adolescent Psychology* 34(3): 412-32.

Kraft, Matthew A. 2019. "Teacher Effects on Complex Cognitive Skills and Social-Emotional Competencies." *Journal of Human Resources* 54(1): 1-36.

Liu, Jing, and Susanna Loeb. 2019. "Engaging Teachers: Measuring the Impact of Teachers on Student Attendance in Secondary School." *Journal of Human Resources*. Forthcoming. doi: 10.3368/jhr.56.2.1216-8430R3.

Moffitt, Terrie, Louise Arseneault, Daniel Belsky, Nigel Dickson, Robert Hancox, HonaLee Harrington, Renate Houts, Richie Poulton, Brent Roberts, Stephen Ross, Malcolm Sears, W. Murray Thomson, and Avshalom Caspi. 2011. "A Gradient of Childhood Self-Control Predicts Health, Wealth, and Public Safety." *Proceedings of the National Academy of Sciences* 108: 2693–98.

Olino, Thomas M., Lan Yu, Dana L. McMakin, Erika E. Forbes, John R. Seeley, Peter M. Lewinsohn, and Paul A. Pilkonis. 2013. "Comparisons Across Depression Assessment Instruments in Adolescence and Young Adulthood: An Item Response Theory Study Using Two Linking Methods." *Journal of Abnormal Child Psychology* 41(8): 1267–77.

Paxson, Christina, and Norbert Schady. 2007. "Cognitive Development among Young Children in Ecuador: The Roles of Wealth, Health, and Parenting." *Journal of Human Resources* 42(1): 49-84.

Pianta, Robert, Karen La Paro, and Bridget Hamre. 2007. *Classroom Assessment Scoring System—CLASS*. Baltimore: Brookes.

Pianta, Robert, Jason Downer, and Bridget Hamre. 2017. "Quality in Early Education Classrooms: Definitions, Gaps, and Systems." *The Future of Children* 26(4): 119-37.

Rivkin, Steven, Eric Hanushek, and John Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica* 73 (2): 417–458.

Rockoff, Jonah, Brian Jacob, Thomas Kane, and Douglas Staiger. 2011. "Can You Recognize an Effective Teacher when You Recruit One?" *Education Finance and Policy* 6: 43–74.

Rosenberg, Morris 1989. *Society and the Adolescent Self-Image* (Rev. ed.). Middletown, CT: Wesleyan University Press.

Schady, Norbert, Jere Behrman, M. Caridad Araujo, Rodrigo Azuero, Raquel Bernal, David Bravo, Florencia Lopez-Boo, Karen Macours, Daniela Marshall, Christina Paxson, and Renos Vakis. 2015. "Wealth Gradients in Early Childhood Cognitive Development in Five Latin American Countries." *Journal of Human Resources* 50(2): 446-63.

Segal, Carmit. 2013. "Misbehavior, Education, and Labor Market Outcomes." *Journal of the European Economic Association* 11(4): 743-79.

Siegler, Robert S., and Julie L. Booth. 2004. "Development of Numerical Estimation in Young

Children." *Child Development* 75(2): 428-44.

Staiger, Douglas and Jonah Rockoff. 2010. "Searching for Effective Teachers with Imperfect Information." *Journal of Economic Perspectives* 24 (3): 97–118.

Steinberg, Matthew P., and Morgaen L. Donaldson. 2016. "The New Educational Accountability: Understanding the Landscape of Teacher Evaluation in the Post-NCLB Era." *Education Finance and Policy* 11(3): 340-59.

Yeager, David S., Paul Hanselman, Gregory M. Walton, Jared S. Murray, Robert Crosnoe, Chandra Muller, Elizabeth Tipton, Barbara Schneider, Chris S. Hulleman, Cintia P. Hinojosa, David Paunesku, Carissa Romero, Kate Flint, Alice Roberts, Jill Trott, Ronaldo Iachan, Jenny Buontempo, Sophia Man Yang, Carlos M. Carvalho, P. Richard Hahn, Maithreyi Gopalan, Pratik Mhatre, Ronald Ferguson, Angela L. Duckworth, and Carol S. Dweck. 2019. "A National Experiment Reveals Where a Growth Mindset Improves Achievement." *Science* 573(7774): 364-69.

Yoshikawa, Hirokazu, Diana Leyva, Catherine Snow, Ernesto Treviño, M. Clara Barata, Christine Weiland, Celia J. Gomez, Lorenzo Moreno, Andrea Rolla, Nikhit D'Sa, and Mary Catherine Arbour. 2015. "Experimental Impacts of a Teacher Professional Development Program in Chile on Preschool Classroom Quality and Child Outcomes." *Developmental Psychology* 51(3): 309–322.

Table 1: Child characteristics

	Mean	Standard deviation	N
Age of child (months)	60.3	4.9	13,858
Gender of child	0.49	0.50	14,477
Receptive vocabulary score (TVIP)	83.3	16.9	13,733
Mother's years of completed schooling	8.8	3.8	13,627
Father's years of completed schooling	8.5	3.8	10,594
Mother's age	30.2	6.6	13,637
Father's age	34.6	7.9	10,620
Proportion who attended preschool	0.61	0.49	14,472
Household has piped water in home	0.83	0.38	14,407
Household has flush toilet in home	0.46	0.50	14,407
Main material of walls is brick or concrete	0.80	0.40	14,407
Main material of floors is dirt	0.06	0.24	14,407
Household has refrigerator	0.82	0.39	14,407
Household has washing machine	0.45	0.50	14,407
Household has TV	0.96	0.18	14,407
Household has computer	0.19	0.39	14,407

Notes: Table reports means and standard deviations of the characteristics of children entering kindergarten in 2012, measured at the beginning of the school year. The TVIP is the *Test de Vocabulario en Imagenes Peabody*, the Spanish version of the Peabody Picture Vocabulary Test (PPVT). The test is standardized using the tables provided by the test developers which set the mean at 100 and the standard deviation at 15 at each age.

Table 2: Characteristics of teachers, by grade

	Mean	Standard deviation	# of teachers
	Panel A: Pooled sample		
CLASS Total	3.3	0.24	2316
Socioemotional support	3.8	0.33	2316
Classroom management	4.9	0.40	2316
Instructional support	1.1	0.15	2316
Proportion female	0.87	0.34	2314
Proportion tenured	0.78	0.41	2302
Years of experience	17.9	10.5	2308
Proportion of teachers with <=3 years of experience	0.04	0.20	2308
Class size	37.6	7.4	2320
	Panel B: Sample by grade		
Kindergarten	3.4	0.28	450
1st grade	3.3	0.23	452
2nd grade	3.3	0.24	465
3rd grade	3.3	0.24	470
4th grade	3.4	0.19	479

Notes: Table reports means and standard deviations of the characteristics of teachers in our sample.

Table 3: Median differences in outcomes and CLASS across classrooms in same grade and school

	TOTAL	K	1st grade	2nd grade	3rd grade	4th grade	5th grade	6th grade
Panel A: Child outcomes								
Math	0.18	0.20	0.19	0.22	0.19	0.14	0.13	0.18
Language	0.16	0.17	0.18	0.15	0.18	0.18	0.13	0.15
Executive function	0.16	0.19	0.17	0.16	0.16	0.15		
Depression								0.23
Self-esteem								0.20
Growth mindset								0.22
Grit								0.20
Panel B: CLASS scores								
CLASS (Total)	0.18	0.22	0.20	0.16	0.15	0.11		
Classroom management	0.22	0.25	0.25	0.22	0.19	0.16		
Socio-emotional support	0.29	0.38	0.33	0.25	0.25	0.21		
Instructional support	0.08	0.08	0.04	0.08	0.08	0.13		

Notes: Table reports differences in child and teacher outcomes between classrooms in the same school and grade. When there are 3 or more classrooms, 2 are selected at random.

Table 4: Proportion of variance of child and teacher outcomes explained by schools, classrooms, and children

	Schools	Classrooms	Children
Panel A: Child outcomes			
Math	0.10	0.13	0.75
Language	0.11	0.13	0.74
Executive function	0.07	0.09	0.60
Depression	0.07	0.11	
Self-esteem	0.05	0.08	
Growth mindset	0.06	0.09	
Grit	0.04	0.07	
Panel B: Teacher outcomes			
CLASS (Total)	0.60		
Classroom Management	0.62		
Socio-emotional Support	0.62		
Instructional Support	0.51		

Notes: Table reports the R-squared from regressions of outcomes on school, classroom, or child fixed effects, respectively.

Table 5: The effects of teacher characteristics and behaviors on achievement

	Binary parametrization of CLASS					Continuous parametrization of CLASS				
Panel A: Total achievement										
CLASS	0.044***				0.046***	0.180***				0.181***
	(0.009)				(0.009)	(0.029)				(0.029)
Experience		0.000			-0.000		0.000			-0.000
		(0.001)			(0.001)		(0.001)			(0.001)
Tenured			0.031**		0.030*			0.031**		0.024
			(0.016)		(0.017)			(0.016)		(0.017)
Gender				0.004	0.003				0.004	0.003
				(0.019)	(0.019)				(0.019)	(0.019)
N	82,081	81,799	81,576	82,007	81,576	82,081	81,799	81,576	82,007	81,576
R-squared	0.14	0.14	0.14	0.14	0.14	0.136	0.14	0.14	0.14	0.14
Panel B: Math achievement										
CLASS	0.044***				0.046***	0.193***				0.196***
	(0.009)				(0.009)	(0.030)				(0.030)
Experience		0.000			-0.000		0.000			-0.000
		(0.001)			(0.001)		(0.001)			(0.001)
Tenured			0.028*		0.027			0.028*		0.020
			(0.017)		(0.018)			(0.017)		(0.018)
Gender				0.000	-0.003				0.000	-0.000
				(0.021)	(0.021)				(0.021)	(0.020)
N	82,088	81,806	81,583	82,014	81,583	82,088	81,806	81,583	82,014	81,583
R-squared	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12	0.12
Panel C: Language achievement										
CLASS	0.031***				0.032***	0.116***				0.116***
	(0.008)				(0.008)	(0.026)				(0.026)
Experience		-0.000			-0.001		-0.000			-0.001
		(0.001)			(0.001)		(0.001)			(0.001)
Tenured			0.023		0.026			0.023		0.022
			(0.016)		(0.017)			(0.016)		(0.017)
Gender				0.005	0.004				0.005	0.004
				(0.017)	(0.017)				(0.017)	(0.017)
N	82,152	81,869	81,646	82,078	81,646	82,152	81,869	81,646	82,078	81,646
R-squared	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14

Notes: Table reports results from regressions of achievement on the CLASS. All regressions include school-by-grade fixed effects, child age in months and its square, and child gender. Standard errors are clustered at the school-by-grade level. *, **, and *** indicate significance at the 10 percent, 5 percent, and 1 percent, respectively.

Table 6: Contemporaneous CLASS effects on achievement, by grade

	Binary parametrization of CLASS			Continuous parametrization of CLASS		
	Total achievement	Math	Language	Total achievement	Math	Language
Kindergarten	0.084***	0.073***	0.063***	0.242***	0.217***	0.174***
	(0.022)	(0.022)	(0.021)	(0.054)	(0.051)	(0.053)
1st grade	0.060***	0.072***	0.022	0.280***	0.306***	0.143***
	(0.022)	(0.023)	(0.018)	(0.067)	(0.071)	(0.054)
2nd grade	0.015	0.018	0.013	0.063	0.109	0.046
	(0.017)	(0.021)	(0.016)	(0.058)	(0.075)	(0.053)
3rd grade	0.050***	0.042**	0.057***	0.134**	0.134**	0.128**
	(0.019)	(0.021)	(0.017)	(0.058)	(0.061)	(0.058)
4th grade	0.017	0.021	0.005	0.136	0.167**	0.055
	(0.017)	(0.018)	(0.018)	(0.083)	(0.082)	(0.086)
F-test (1)	0.06	0.17	0.10	0.08	0.28	0.44
F-test (2)	0.02	0.02	0.31	0.01	0.04	0.13
R-squared	0.14	0.12	0.14	0.14	0.12	0.14
N	82,081	82,088	82,152	82,081	82,088	82,152

Notes: Table reports results from regressions of achievement on the CLASS. All regressions include school fixed effects, child age in months and its square, and child gender. Standard errors are clustered at the school-by-grade level. *, **, and *** indicate significance at the 10 percent, 5 percent, and 1 percent, respectively. F-test (1) is a test that the coefficient in all five grades are equal; F-test (2) is a test that the coefficients on the "earlier" grades (kindergarten and 1st grade) are the same as those on the "later" grades (2nd, 3rd, and 4th grades).

Table 7: Fade-out of CLASS effects on total achievement, by grade

	Binary parametrization of CLASS						
	t	t+1	t+2	t+3	t+4	t+5	t+6
Kindergarten	0.079***	0.044***	0.049***	0.063***	0.058***	0.058***	0.064***
	(0.027)	(0.020)	(0.019)	(0.020)	(0.021)	(0.021)	(0.020)
1st grade	0.049**	0.028	0.009	0.008	0.009	0.009	
	(0.024)	(0.020)	(0.019)	(0.018)	(0.018)	(0.018)	
2nd grade	0.016	0.01	0.002	-0.002	-0.013		
	(0.019)	(0.017)	(0.018)	(0.017)	(0.018)		
3rd grade	0.058***	0.048***	0.046***	0.030*			
	(0.020)	(0.018)	(0.017)	(0.018)			
4th grade	0.017	0.009	0.018				
	(0.018)	(0.016)	(0.015)				
F-test (1)	0.18	0.38	0.23	0.07	0.03	0.07	
F-test (2)	0.11	0.43	0.66	0.25	0.05		
	Continuous parametrization of CLASS						
	t	t+1	t+2	t+3	t+4	t+5	t+6
Kindergarten	0.221***	0.084*	0.077*	0.114**	0.090*	0.076	0.096*
	(0.065)	(0.050)	(0.046)	(0.055)	(0.051)	(0.055)	(0.055)
1st grade	0.231***	0.149***	0.065	0.066	0.072	0.031	
	(0.069)	(0.056)	(0.055)	(0.053)	(0.051)	(0.053)	
2nd grade	0.055	0.010	0.033	0.013	-0.033		
	(0.067)	(0.061)	(0.060)	(0.060)	(0.054)		
3rd grade	0.150**	0.128**	0.143**	0.092			
	(0.061)	(0.061)	(0.057)	(0.061)			
4th grade	0.163*	0.126**	0.127**				
	(0.083)	(0.063)	(0.062)				
F-test (1)	0.33	0.49	0.68	0.63	0.21	0.54	
F-test (2)	0.08	0.60	0.54	0.52	0.08		

Notes: Table reports results from regressions of achievement on the CLASS. All regressions include school-by-grade fixed effects, child age in months and its square, and child gender. Standard errors are clustered at the school-by-grade level. *, **, and *** indicate significance at the 10 percent, 5 percent, and 1 percent, respectively. Sample sizes are 9,076 (kindergarten), 11,197 (1st grade), 12,525 (2nd grade), 13,760 (3rd grade), and 14,733 (4th grade). Sample sizes are smaller in earlier grades because, to be included in the regressions for a given grade, children need to have attended the sample of schools in the study in every grade thereafter. Coefficients on the CLASS in *t* are not identical to those in Table 6 because of differences in the samples. F-test (1) is a test that the coefficient in all five grades are equal; F-test (2) is a test that the coefficients on the "earlier" grades (kindergarten and 1st grade) are the same as those on the "later" grades (2nd, 3rd, and 4th grades).

Table 8: The effects of teacher characteristics and behaviors on executive function

	Binary parametrization of CLASS					Continuous parametrization of CLASS				
Panel A: Total executive function score										
CLASS	0.034***				0.035***	0.133***				0.135***
	(0.008)				(0.008)	(0.025)				(0.025)
Experience		-0.000			-0.001		-0.000			-0.001
		(0.001)			(0.001)		(0.001)			(0.001)
Tenured			0.019		0.021			0.019		0.016
			(0.014)		(0.015)			(0.014)		(0.015)
Gender				-0.006	-0.008				-0.006	-0.008
				(0.017)	(0.017)				(0.017)	(0.016)
N	82,101	81,818	81,595	82,027	81,595	82,101	81,818	81,595	82,027	81,595
R-squared	0.08	0.08	0.08	0.08	0.08	0.076	0.075	0.075	0.075	0.076
Panel B: Cognitive flexibility										
CLASS	0.017**				0.018***	0.084***				0.090***
	(0.007)				(0.008)	(0.023)				(0.023)
Experience		-0.001			-0.001		-0.001			-0.001
		(0.001)			(0.001)		(0.001)			(0.001)
Tenured			-0.011		-0.009			-0.011		-0.013
			(0.13)		(0.014)			(0.013)		(0.014)
Gender				-0.019	-0.022				-0.019	-0.022
				(0.016)	(0.016)				(0.016)	(0.016)
N	82,101	81,818	81,595	82,027	81,595	82,101	81,818	81,595	82,027	81,595
R-squared	0.05	0.05	0.05	0.05	0.05	0.050	0.050	0.050	0.050	0.050
Panel C: Inhibitory control										
CLASS	0.006				0.007	0.062**				0.065**
	(0.009)				(0.009)	(0.028)				(0.028)
Experience		-0.000			-0.000		-0.000			-0.000
		(0.001)			(0.001)		(0.001)			(0.001)
Tenured			0.000		-0.000			0.000		-0.003
			(0.015)		(0.016)			(0.015)		(0.016)
Gender				-0.011	-0.012				-0.011	-0.013
				(0.017)	(0.017)				(0.017)	(0.017)
N	65,790	65,548	65,453	65,757	65,453	65,790	65,548	65,453	65,757	65,453
R-squared	0.04	0.04	0.04	0.04	0.04	0.042	0.042	0.042	0.042	0.042

Panel D: Working memory										
CLASS	0.035***				0.036***	0.116***				0.114***
	(0.008)				(0.008)	(0.025)				(0.025)
Experience		0.000			-0.000		0.000			-0.000
		(0.001)			(0.001)		(0.001)			(0.001)
Tenured			0.028*		0.028*			0.028*		0.024
			(0.015)		(0.016)			(0.015)		(0.016)
Gender				-0.001	-0.001				-0.001	-0.001
				(0.016)	(0.016)				(0.016)	(0.016)
N	82,101	81,818	81,595	82,027	81,595	82,101	81,818	81,595	82,027	81,595
R-squared	0.08	0.08	0.08	0.08	0.08	0.079	0.078	0.078	0.078	0.078

Notes: Table reports results from regressions of executive function on the CLASS. All regressions include school-by-grade fixed effects, child age in months and its square, and child gender. Standard errors are clustered at the school-by-grade level. *, **, and *** indicate significance at the 10 percent, 5 percent, and 1 percent, respectively. Sample sizes for the inhibitory control regressions are smaller because no inhibitory control test was applied in 1st grade.

Table 9: CLASS effects on the incidence of behavioral problems

	Binary parametrization of CLASS		**Continuous parametrization of CLASS**	
	Probability of bad behavior, *g+1*	Probability of bad behavior, *g+2*	Probability of bad behavior, *g+1*	Probability of bad behavior, *g+2*
Panel A: Pooled sample	-0.005* (0.002)	-0.003 (0.002)	-0.014* (0.007)	-0.010 (0.007)
Panel B: Effects by grade				
Kindergarten	-0.003 (0.005)	-0.007 (0.005)	-0.014 (0.013)	-0.004 (0.014)
1st grade	-0.009* (0.005)	-0.001 (0.005)	-0.024* (0.014)	-0.001 (0.014)
2nd grade	0.001 (0.006)	-0.002 (0.005)	0.007 (0.018)	-0.017 (0.015)
3rd grade	-0.005 (0.005)	-0.008 (0.005)	-0.012 (0.018)	-0.026 (0.017)
4th grade	-0.006 (0.006)	0.002 (0.005)	-0.030 (0.022)	0.001 (0.020)
F-test (1)	0.75	0.63	0.66	0.73
F-test (2)	0.59	0.71	0.62	0.42

Notes: Table reports coefficients and standard errors of regressions of an indicator variable equal to one if a teacher in grade *g+1* (*g+2*) reported that a child was among the worst-behaved in her classroom on the CLASS in grade *g*. All regressions include school-by-grade fixed effects, child age in months and its square, and child gender. Standard errors are clustered at the school-by-grade level. *, **, and *** indicate significance at the 10 percent, 5 percent, and 1 percent, respectively. F-test (1) is a test that the coefficient in all five grades are equal; F-test (2) is a test that the coefficients on the "earlier" grades (kindergarten and 1st grade) are the same as those on the "later" grades (2nd, 3rd, and 4th grades).

Table 10: CLASS effects on non-cognitive outcomes in 6th grade

	Depression	Self-esteem	Growth mindset	Grit	Aggregate
Panel A: # of above-average teachers (*continuous*)	0.007	0.001	0.014	0.002	0.007
	(0.010)	(0.010)	(0.010)	(0.010)	(0.009)
Panel B: # of above-average teachers (*discrete*)					
0 or 1 A teachers	-0.021	-0.017	-0.031	-0.043	-0.036
	(0.026)	(0.028)	(0.028)	(0.029)	(0.026)
4 or 5 A teachers	0.019	0.020	0.048*	-0.010	0.028
	(0.031)	(0.029)	(0.028)	(0.028)	(0.026)
F-test	0.30	0.30	0.02	0.37	0.05
Panel C: Average CLASS *(continuous)*	0.164	-0.118	0.135	-0.060	-0.008
	(0.159)	(0.161)	(0.162)	(0.146)	(0.154)

Notes: Table reports coefficients and standard errors of regressions of non-cognitive outcomes on the CLASS. In Panel A, the CLASS is parametrized as the number of above-average teachers between kindergarten and 4th grade; in Panel B, we report the coefficients on indicator variables for 0 or 1 above-average teachers, and 4 or 5 above-average teachers (the omitted category is children who had 2 or 3 above-average teachers between kindergarten and 4th grade); in panel C, we take the average of the CLASS scores of the teachers a child had between kindergarten and 4th grade. All regressions include school-by-grade fixed effects, child age in months and its square, and child gender. Standard errors are clustered at the school-by-grade level. *, **, and *** indicate significance at the 10 percent, 5 percent, and 1 percent, respectively. F-test is a test that the coefficient on the indicator variable for 0 or 1 above-average teachers in Panel B is equal to the corresponding indicator variable for 4 or 5 above-average teachers.

Table 11: Child responses to differences in teacher quality in 1st grade

	Effort in class	Reading at home
Panel A: Binary parametrization of CLASS		
OLS	-0.001	0.006
	(0.006)	(0.007)
Ordered Probit		
Response category: *never*	0.001	-0.004
	(0.003)	(0.005)
Response category: *sometimes*	0.001	-0.002
	(0.002)	(0.002)
Response category: a*lways*	-0.002	0.006
	(0.006)	(0.007)
F-test	0.95	0.72
Panel B: Continuous parametrization of CLASS		
OLS	-0.008	0.017
	(0.017)	(0.022)
Ordered Probit		
Response category: *never*	0.004	-0.011
	(0.010)	(0.015)
Response category: *sometimes*	0.003	-0.005
	(0.007)	(0.007)
Response category: a*lways*	-0.006	0.016
	(0.017)	(0.021)
F-test	0.94	0.75

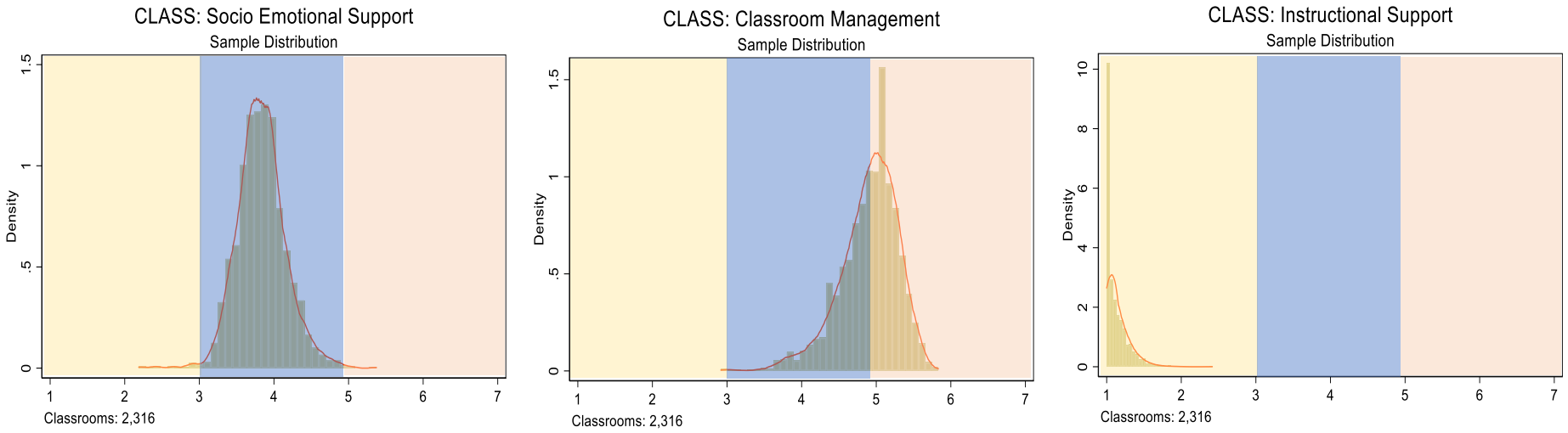
Note: Table reports coefficients and standard errors of regressions of child effort in class or reading at home on the CLASS. In the OLS regressions, the dependent variable takes on the value if the child reports that she "always" tries as hard as she can in class (always reads at home); in the ordered probit regressions, we report marginal effects of the CLASS on "never", "sometimes", or "always" try as hard as possible in school (read at home). In Panel A, the CLASS is parametrized as an indicator variable that takes on the value of one if a teacher had a CLASS score above the average for her school and grade, zero otherwise, and in Panel B the CLASS is parametrized as a continuous variable. All regressions include school-by-grade fixed effects, child age in months and its square, and child gender. Standard errors are clustered at the school-by-grade level. F-test gives p-value on test that the marginal effects for the "never", "sometimes", and "always" response categories are equal. *, **, and *** indicate significance at the 10 percent, 5 percent, and 1 percent, respectively.

Table 12: Teacher responses to random variation in *lagged* student characteristics

	Math achievement, *t-1*	Language achievement, *t-1*	Executive function, *t-1*	Behavioral problems, *t-1*
Panel A: Binary parametrization of CLASS	0.000	0.002	-0.008	-0.003
	(0.008)	(0.008)	(0.026)	(0.008)
Panel B: Continuous parametrization of CLASS	0.015	-0.001	-0.000	-0.002
	(0.025)	(0.022)	(0.007)	(0.003)
N	60,141	60,155	60,144	60,535

Notes: Table reports coefficients and standard errors of regressions of *lagged* achievement, *lagged* executive function, or an indicator variable for children with behavioral problems in the previous year on the CLASS. In Panel A, the CLASS is parametrized as an indicator variable that takes on the value of one if a teacher had a CLASS score above the average for her school and grade, zero otherwise, and in Panel B the CLASS is parametrized as a continuous variable. All regressions refer to the pooled regressions with data on child outcomes between kindergarten and 3rd grades, and data on the CLASS for teachers between 1st and 4th grades, and include school-by-grade fixed effects, child age in months and its square, and child gender. Standard errors are clustered at the school-by-grade level. *, **, and *** indicate significance at the 10 percent, 5 percent, and 1 percent, respectively.

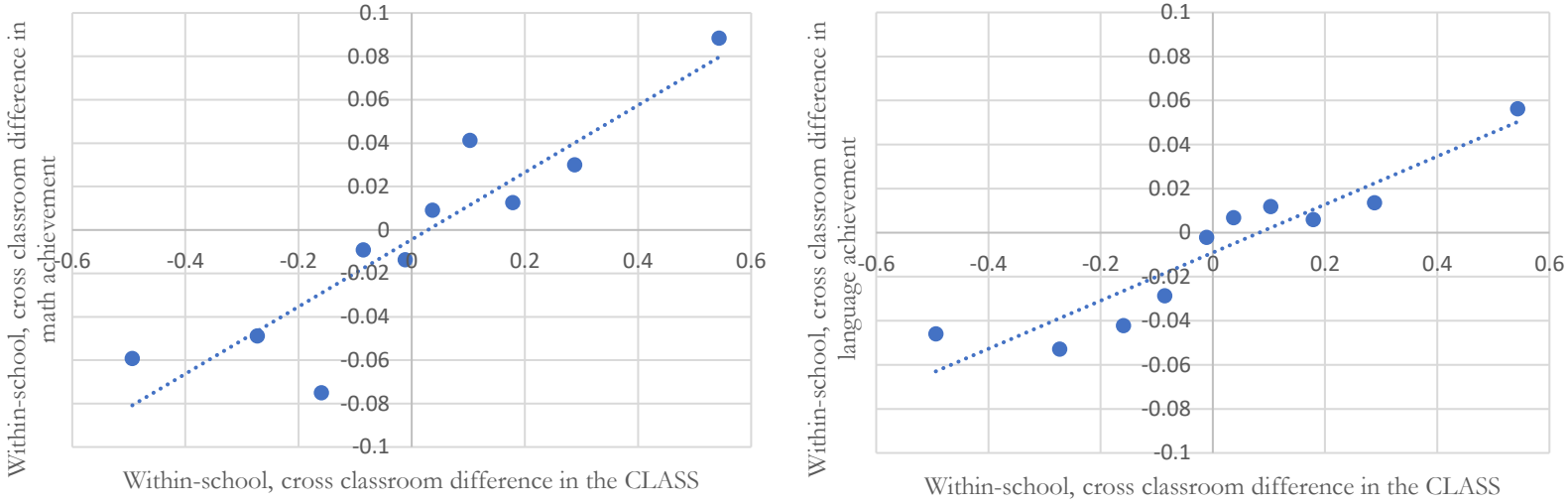
Figure 1: Distribution of CLASS scores



Notes: Figure shows univariate densities of the three CLASS domains. Scores lower than 3 are considered to be "low", those between 3 and 5 are considered to be "medium", and those higher than 5 are considered to be "high" by the CLASS developers.

Figure 2: Associations between CLASS and achievement

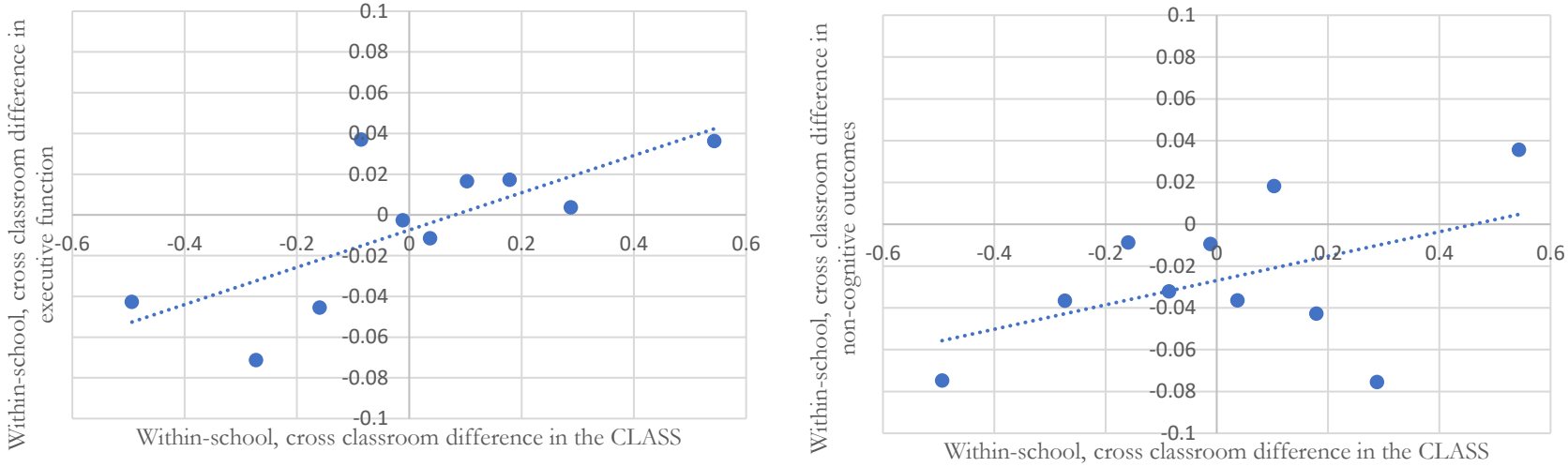
Panel A: Math Achievement **Panel B: Language Achievement**



Notes: To generate the figures, in every school and grade, we take two classrooms, label one as classroom A and the other as classroom B, and calculate the median difference A-B in the CLASS and achievement. When there are more than 2 classrooms in a school and grade, we select two at random. We have 200 schools and 5 years of data, so at the end of this process we have approximately 1,000 differences. We sort observations into deciles of the cross-classroom difference in the CLASS, and calculate the median difference in the CLASS and achievement in each decile. Finally, we plot the average differences in achievement (vertical axis) as a function of the differences in the CLASS (horizontal axis) for each decile, and include a regression line for the 10 points.

Figure 3: Associations between CLASS, executive function, and non-cognitive outcomes

Panel A: Executive Function **Panel B: Non-cognitive outcomes**



Notes: To generate the figures, in every school and grade, we take two classrooms, label one as classroom A and the other as classroom B, and calculate the median difference A-B in the CLASS and executive function (Panel A) or the aggregate measure of non-cognitive outcomes (Panel B). When there are more than 2 classrooms in a school and grade, we select two at random. We have 200 schools and 5 years of data, so at the end of this process we have approximately 1,000 differences (200 for non-cognitive outcomes). We sort observations into deciles of the cross-classroom difference in the CLASS, and calculate the median difference in the CLASS and executive function or the average of non-cognitive outcomes in each decile. Finally, we plot the average differences in child outcomes (vertical axis) as a function of the differences in the CLASS (horizontal axis) for each decile, and include a regression line for the 10 points.

Appendix A: Randomization checks

We refer to our assignment as "random" as shorthand, although technically random assignment occurred only in 3rd through 6th grades. In the other grades, the assignment rules were as-good-as-random. Specifically, the assignment rules we implemented were as follows: In kindergarten, all children in each school were ordered by their last name and first name, and were then assigned to teachers in alternating order; in 1st grade, they were ordered by their date of birth, from oldest to youngest, and were then assigned to teachers in alternating order; in 2nd grade, they were divided by gender, ordered by their first name and last name, and then assigned in alternating order; in 3rd through 6th grades, they were divided by gender and then randomly assigned to one or another classroom.

To check on randomization, we first regress each baseline characteristic of children on the CLASS and school fixed effects, separately by grade. These results are in Table A1. The coefficients on these regressions are all small in magnitude, and only 3 (out of 30) are significant at conventional levels.[41] Second, Appendix Table A2 shows that, as expected, there is no association between the CLASS scores of teachers that children were assigned to in grades *g*, *g+1* …*g+4* – nine of ten correlations are smaller than 0.01 in absolute value, and one is 0.014.

[41] The coefficients from these regressions imply that, in 1st grade, in classrooms with a 1-point-higher CLASS score, the probability that a child was a girl was 0.057 higher; in 4th grade, children in classrooms with a 1-point-higher CLASS score were 0.62 months younger; in 2nd grade, the households of children in classrooms with a 1-point-higher CLASS score had 0.08 standard deviations higher wealth (where "wealth" is given by the 8 household characteristics in the last rows of Table 1 in the main body of the paper, aggregated by factor analysis).

Table A1: Associations between CLASS and baseline characteristics of children and families

	Gender	Age	TVIP	Mother's education	Mother's Age	Wealth Index
Kindergarten	0.005	-0.144	0.019	0.006	-0.014	0.028
	(0.022)	(0.195)	(0.048)	(0.016)	(0.328)	(0.036)
1st grade	0.057**	-0.134	0.007	0.002	0.202	0.020
	(0.028)	(0.136)	(0.054)	(0.024)	(0.330)	(0.034)
2nd grade	0.033*	0.021	-0.058	0.022	0.442	0.080**
	(0.018)	(0.265)	(0.061)	(0.029)	(0.415)	(0.040)
3rd grade	0.023	0.137	-0.026	0.011	0.162	-0.031
	(0.019)	(0.287)	(0.054)	(0.027)	(0.466)	(0.040)
4th grade	-0.003	-0.616**	-0.136*	-0.048	-0.342	0.005
	(0.029)	(0.301)	(0.071)	(0.033)	(0.427)	(0.058)

Notes: Table reports results from regressions of baseline characteristics on the CLASS. All regressions include school-by-grade fixed effects, child age in months and its square, and child gender. Standard errors are clustered at the school-by-grade level. *, **, and *** indicate significance at the 10 percent, 5 percent, and 1 percent, respectively.

Table A2: Correlations between CLASS scores, kindergarten through 4th grade

	1st grade	2nd grade	3rd grade	4th grade
Kindergarten	-0.010	0.000	0.007	0.001
1st grade		-0.004	-0.008	0.006
2nd grade			0.011	-0.003
3rd grade				0.014*

Notes: Table presents the results from pairwise correlations between the CLASS a child was exposed to in grades *g*, *g+1* … *g+n*, after removing school-by-grade averages. *, **, and *** indicate significance at the 10 percent, 5 percent, and 1 percent, respectively.

Appendix B: Application of the CLASS in Ecuador

This Appendix reproduces Appendix B in Araujo et al. (2016), with the values in the tables updated to include all five grades in which we applied the CLASS. (Araujo et al. 2016 used the kindergarten data only.)

To measure teacher behaviors (or interactions), we use the CLASS (Pianta et al. 2007). The CLASS measures teacher behaviors in three broad *domains*: emotional support, classroom organization, and instructional support. Within each of these domains, there are a number of CLASS *dimensions*. Within emotional support these dimensions are positive climate, negative climate, teacher sensitivity, and regard for student perspectives; within classroom organization, the dimensions are behavior management, productivity, and instructional learning formats; and within instructional support, they are concept development, quality of feedback, and language modeling.

The *behaviors* that coders are looking for in each dimension are quite specific—see Appendix Table B1 for an example of the behaviors considered under the behavior management dimension. For this dimension, a coder scoring a particular segment would assess whether there are clear behavior rules and expectations, and whether these are applied consistently; whether a teacher is proactive in anticipating problem behavior (rather than simply reacting to it when it has escalated); how the teacher deals with instances of misbehavior, including whether misbehavior is redirected using subtle cues; whether the teacher is attentive to positive behaviors (not only misbehavior); and whether there is generally compliance by students with classroom rules or, rather, frequent defiance. For each of these behaviors, the CLASS protocol then gives a coder concrete guidance on whether the score given should be "low" (scores of 1-2), "medium" (scores of 3-5), or "high" (scores of 6-7).

In practice, in our application of the CLASS, scores across different dimensions are highly correlated with each other. In our study sample, the correlation coefficients across the three different CLASS domains range from 0.46 (for emotional support and instructional support) to 0.70 (for emotional support and classroom organization). Similar findings have been reported elsewhere. Kane et al. (2011) report high correlations between different dimensions of a classroom observation tool based on the Framework for Teaching (FFT; Danielson 1996) that is used to assess teacher performance in the Cincinnati public school system, with pairwise correlations between 0.62 and 0.81. Kane and Staiger (2012) show that scores on the FFT and the CLASS in the MET study are highly correlated with each other. Also, in an analysis based on principal components, they show that 91 percent and 73 percent of the variance in the FFT and CLASS, respectively, are accounted for by the first principal component of the teacher behaviors that are measured by each instrument (10 dimensions in the case of the CLASS, scored on a 1-7 point scale, and 8 on the FFT, scored on a 1-4 point scale).

To apply the CLASS in Ecuador, we filmed all teachers for a full school day (from approximately eight in the morning until one in the afternoon). In accordance with CLASS protocols, we then discarded the first hour of film (when teachers and students are more likely to be aware of, and responding to, the camera), as well as all times that were not instructional (for example, break, lunch) or did not involve the main teacher (for example, PE class). The remaining video was cut into usable 20-minute *segments*. We selected the first four segments per teacher, for a total of more than 9,000 segments. These segments were coded by a group of 6-8 coders who were explicitly trained for this purpose. A master CLASS coder trained, provided feedback, and supervised the coders. During the entire process, we interacted extensively with the developers of the CLASS at the University of Virginia.

One concern with any application of the CLASS is that teachers "act" for the camera. Informal observations by the study team and, in particular, the master CLASS trainer suggests that this was not the case. As a precaution, and in addition to discarding the first hour of video footage, we compare average CLASS scores for the first and fourth segments. We find that average CLASS scores are somewhat lower later in the day than earlier, but the difference is small (the mean score is 3.35 in the fourth segment, compared to 3.48 in the first segment); moreover, the change in CLASS scores between the first and fourth segment is not significantly associated with a teacher's mean CLASS scores. This pattern of results suggests that teachers are not "acting" for the camera, and that any "camera effects" are unrelated to underlying teacher quality, as measured by the CLASS.

In spite of the rigorous process we followed for coder selection, training, and supervision, and as with any other classroom observation tool, there is likely to be substantial measurement error in the CLASS. This measurement error can arise from at least two important sources: coding error, and the fact that the CLASS score is taken from a single day of teaching (from the approximately 200 days a child spends in school a year in Ecuador). There may also be filming error if the quality of the video is poor, but we do not believe that this was an important concern in our application.

To minimize coder error, all segments were coded by two separate, randomly assigned, coders. We expected there would be substantial discrepancies in scores across coders. In practice, however, as we show in Appendix Table B2, the inter-coder reliability ratio was high, 0.86, suggesting that this source of measurement error was relatively unimportant in our application of the CLASS, at least when all CLASS dimensions are taken together. We note that inter-coder reliability in our study compares favorably with that found in other studies that use the CLASS. Pianta et al. (2008) report an inter-coder correlation of 0.71, compared to 0.87 in our study; Brown et al. (2010) double-coded 12 percent of classroom observations, and report an inter-coder reliability ratio of 0.83 for this sub-sample, compared to 0.92 in our study.

Another important source of measurement error occurs because teachers are filmed on a single day. This day is a noisy measure of the quality of teacher-child interactions in that classroom over the course of the school year for a variety of reasons. Teachers may have a particularly good or bad day; a particularly troublesome student may be absent from the class on the day when filming occurred; there could be some source of external disruption (say, construction outside the classroom); some teachers may be better at teaching subject matter that is covered early or late in the year. To get a sense of the importance of this source of measurement error, in Appendix Table B2, we report the correlations between the CLASS scores a teacher received based on the video from the 1st and 4th segments. That correlation, 0.40, is substantially lower than the correlation across coders discussed above. We note that this pattern—large increases in measured relative to "true" variability with more segments per day, but smaller increases with more coders per segment—has also been found in a Generalizability Study (G-Study) of the CLASS with US data (Mashburn et al. 2012).

Further details on filming and coding are given in Filming and Coding Protocols for the CLASS in Ecuador. These are available from the authors upon request.

References

Brown, J., S. Jones, M. LaRusso and L. Aber. 2010. "Improving Classroom Quality: Teacher Influences and Experimental Impacts of the 4Rs Program." *Journal of Educational Psychology* 102(1): 153-67.

Danielson, C. 1996. *Enhancing Professional Practice: A Framework for Teaching.* Alexandria, VA: Association for Supervision and Curriculum Development.

Downer, J.T., L.M. Booren, O.K. Lima, A.E. Luckner, and R.C. Pianta. 2010. "The Individualized Classroom Assessment Scoring System (inCLASS): Preliminary Reliability and Validity of a System for Observing Preschoolers' Competence in Classroom Interactions." *Early Childhood Research Quarterly* 25(1): 1-16.

Kane, T., and D. Staiger 2012. *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains.* Bill & Melinda Gates Foundation.

Kane, T., E. Taylor, J. Tyler, and A. Wooten. 2011. "Identifying Effective Classroom Practices Using Student Achievement Data." *Journal of Human Resources* 46(3): 587-613.

Mashburn, A., J. Brown, J. Downer, K. Grimm, S. Jones, and R. Pianta. 2012. "Conducting a Generalizability Study to Understand Sources of Variation in Observational Assessments of Classroom Settings." Unpublished manuscript, University of Virginia.

Pianta, R., K. LaParo and B. Hamre. 2007. *Classroom Assessment Scoring System—CLASS*. Baltimore: Brookes.

Pianta, R., A. Mashburn, J. Downer, B. Hamre, and L. Justice. 2008. "Effects of Web-Mediated Professional Development Resources on Teacher-Child Interactions in Pre-Kindergarten Classrooms." *Early Childhood Research Quarterly* 23(4): 431-51.

Appendix Table B1: CLASS scores for Behavior Management dimension

Behavior Management Encompasses the teacher's ability to provide clear behavioral expectations and use effective methods to prevent and redirect misbehavior.	Low (1,2)	Mid (3,4,5)	High (6,7)
Clear Behavior Expectations ▪ Clear expectations ▪ Consistency ▪ Clarity of rules	Rules and expectations are absent, unclear, or inconsistently enforced.	Rules and expectations may be stated clearly, but are inconsistently enforced.	Rules and expectations for behavior are clear and are consistently enforced.
Proactive ▪ Anticipates problem behavior or escalation ▪ Rarely reactive ▪ Monitoring	Teacher is reactive and monitoring is absent or ineffective.	Teacher uses a mix of proactive and reactive responses; sometimes monitors but at other times misses early indicators of problems.	Teacher is consistently proactive and monitors effectively to prevent problems from developing.
Redirection of Misbehavior ▪ Effectively reduces misbehavior ▪ Attention to the positive ▪ Uses subtle cues to redirect ▪ Efficient	Attempts to redirect misbehavior are ineffective; teacher rarely focuses on positives or uses subtle cues. As a result, misbehavior continues/escalates and takes time away from learning.	Some attempts to redirect misbehavior are effective; teacher sometimes focuses on positives and uses subtle cues. As a result, there are few times when misbehavior continues/escalates or takes time away from learning.	Teacher effectively redirects misbehavior by focusing on positives and making use of subtle cues. Behavior management does not take time away from learning.
Student Behavior ▪ Frequent compliance ▪ Little aggression & defiance	There are frequent instances of misbehavior in the classroom.	There are periodic episodes of misbehavior in the classroom.	There are few, if any, instances of student misbehavior in the classroom.

Source: Pianta et al. (2007).

Table B2. Sources of measurement error in the CLASS

	Inter-coder correlation	Inter-segment correlation (1st and 4th segments)
Kindergarten	0.86	0.45
1st Grade	0.84	0.39
2nd Grade	0.89	0.38
3rd Grade	0.88	0.38
4th Grade	0.79	0.35
Pooled	0.86	0.40

Note: Table reports the correlation of the CLASS for different coders and different segments within a day for the same teacher

Appendix C: Univariate densities, maternal education gradients, and gender differences in child outcomes

In this appendix, we present univariate densities of child outcomes, as well as differences by maternal education and gender.

A. Univariate densities:

Figure C1 presents the univariate densities of our achievement measures, separately by grade. The figure shows that most of the distributions appear to have a reasonable spread and are generally symmetric. One clear exception is math achievement in kindergarten, which is left-censored.

Figure C2 presents comparable densities for executive function. It shows that the distributions of inhibitory control and cognitive flexibility are often highly skewed. This is not surprising given the nature of the tests. As an example, we describe the executive function tests we applied in kindergarten.

In the inhibitory control test, kindergarten children were quickly shown a series of 14 flash cards that had either a sun or a moon and were asked to say the word "day" when they saw the moon and "night" when they saw the sun. Just over half (50.8 percent) of all children made no mistake on this test, so there is a concentration of mass at the highest value, while very few children (1.6 percent) answered all prompts incorrectly.[42]

The cognitive flexibility test we applied in kindergarten worked as follows. Children were handed a series of picture cards, one by one. Cards had either a truck or a star, in red or blue. The enumerator asked the child to sort cards by *color*, or by *shape*. Specifically, in the first half of the test, the enumerator asked the child to play the "colors" game, handed her cards, indicating their color, and asked the child to place them in the correct pile ("this is a red card: where does it go?"). After 10 cards, the enumerator told the child that they would switch to the "shapes" game, and reminded the child that, in this game, trucks should be placed in one pile and stars in another. The enumerator then handed the child cards, indicating the shapes on the card, and asked her to place them in the correct pile ("this is a star: where does it go?"). In both the first and the second part of the test, if the child made three consecutive mistakes, the enumerator paused the test, reminded her what game they were playing ("remember we are playing the shapes game; in the shapes game, all trucks go in this pile, and all stars in this other pile"), and handed the child a new card with the corresponding instruction. A small proportion of children in kindergarten (7.5 percent) did not understand the game, despite repeated examples, and were given a score of 0; just under half of all children (47 percent) answered all prompts correctly in both the "colors" and "shapes" parts of the test; and just over a quarter (27.3 percent) of all children made no mistakes in the first part of the test (the "colors" game), but incorrectly classified every card in the second part of the test (the "shapes" game). These children were unable to switch rules, despite repeated promptings from the enumerator. The distribution of scores for this test therefore has a concentration of mass at two points, with much less mass at other points.

[42] We did not apply an inhibitory control test in 1st grade because, during the pilot, we found that virtually all 1st graders got a perfect (or close to perfect) score on the "Day-Night" test, but only a minority of children could carry out the inhibitory control test we applied in 2nd grade. In that test, children were shown words that correspond to a color, written in ink of a different color (for example, the word "green" written in red ink). They were then asked to say the name of the color of the ink, thus suppressing the natural reaction, which is to read the word written on the page. The test favors children who cannot read, or can read only very imperfectly, which is why we did not apply it in 1st grade.

The working memory test had two parts. In the first part, children were given 2 minutes to find as many sequences of dog, house, and ball, in that order, on a sheet that has rows of dogs, houses, and balls in various possible sequences. The score on this part of the test is the number of correct sequences found by the child. In the second part of the test, the enumerator recited strings of numbers, and asked the child to repeat them, in the same order or backwards. Figure C2 shows that the aggregate working memory score is distributed smoothly, with little evidence of a concentration of mass at particular values.

In practice the correlations of the scores across the three dimensions in our sample are low—in the range of 0.21 to 0.32 between cognitive flexibility and working memory, between 0.17 and 0.33 between working memory and inhibitory control, and in the range of 0.12 to 0.15 between cognitive flexibility and inhibitory control—see Appendix Table C1.[43] When the scores across the three dimensions are averaged, the distributions of the total executive function score are generally smooth and symmetric.

Figure C3, finally, shows univariate densities of the four non-cognitive measures we applied in 6th grade. The figure shows that the distribution of the depression and grit scores appear to be right-censored. The distribution for the aggregate measure of non-cognitive outcomes, on the other hand, is smooth and symmetric. Table C2 shows that the different non-cognitive outcomes are positively correlated, although the correlations are far from unity—they range from 0.20 (between depression and grit) to 0.49 (between growth mindset and self-esteem).

B. Maternal education gradients and differences by gender

Table C3 shows there are differences in almost every outcome by maternal education.[44] In achievement, differences between the highest- and lowest-education groups are 0.41 SDs and 0.48 SDs for math and language, respectively; for executive function, these differences are 0.16 SDs for inhibitory control, 0.20 SDs for cognitive flexibility, and 0.31 SDs for working memory; for the non-cognitive outcomes the differences are 0.30 SDs for depression, 0.18 SDs for self-esteem, 0.10 SDs for grit, and 0.25 SDs for growth mindset.[45] All but one difference are significant at the 1 percent level or higher. The only exception is the incidence of behavioral problems, which is very similar for children of mothers in the three education categories.

Table C3 also shows there are notable differences by gender in many outcomes. Boys have higher math scores than girls (difference of 0.11 SDs), but lower language scores (0.10 SDs); differences in executive function by gender are generally small; gender differences in non-cognitive outcomes, on the other hand, are large, and consistently favor girls—boys have worse depression scores (0.12 SDs), less self-esteem (0.16 SDs), less grit (0.16 SDs), and lower scores on growth mindset (0.10 SDs). By far the biggest differences are in the incidence of behavioral problems: Consistent with what has been found elsewhere (see Bertrand and Pan 2013), 86 percent of the children who are reported as having behavioral problems by their teachers are boys.

[43] The fact that these correlations are very low is likely to be a result of both measurement error and differences across the constructs that each domain measures.

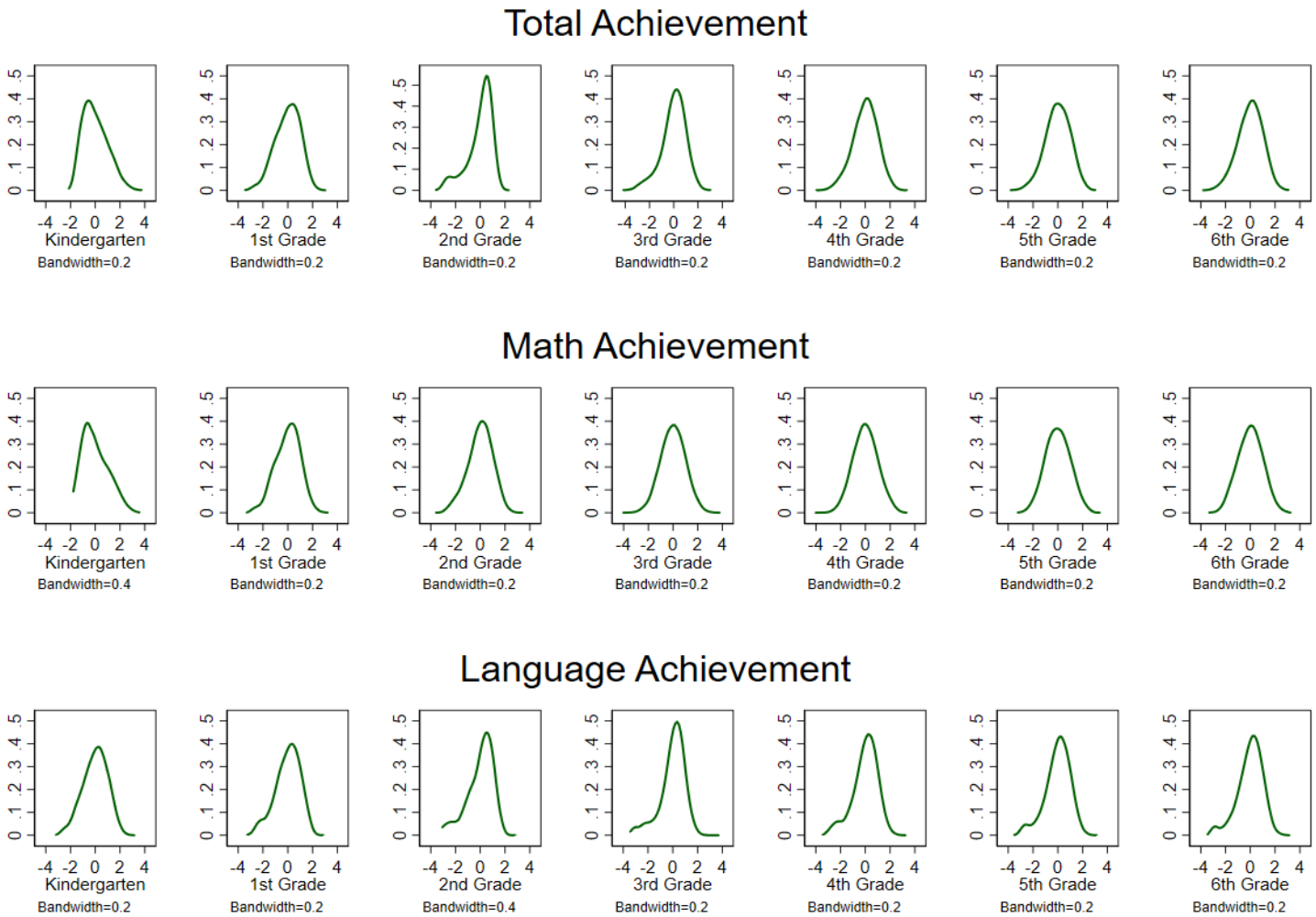
[44] When outcomes are available for more than one grade, we report the average across all grades.

[45] All outcomes have been rescaled so that a positive value is better than a negative value. For example, a higher depression score means that children are *less* likely to be depressed.

References

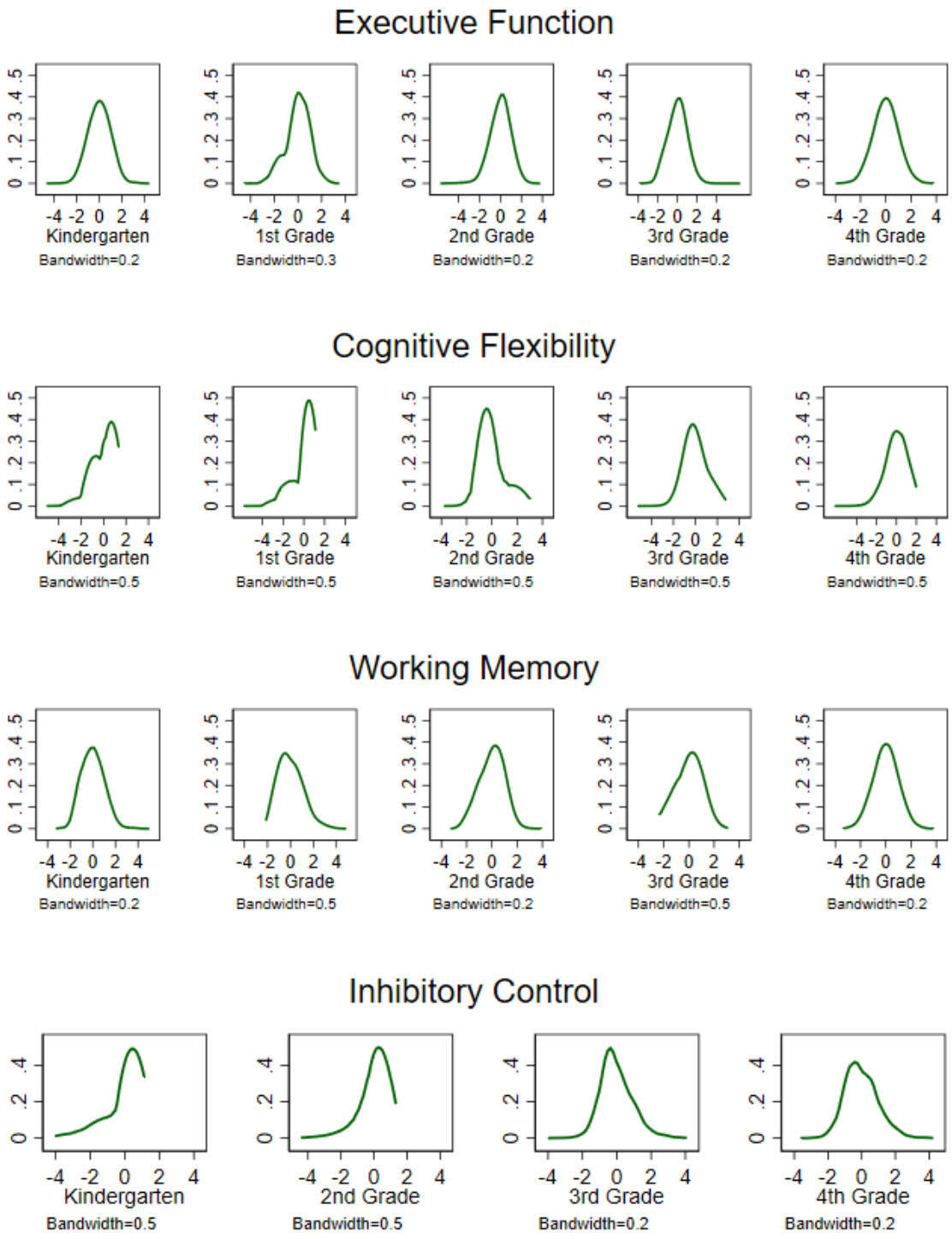
Bertrand, Marianne, and Jessica Pan. 2013. "The Trouble with Boys: Social Influences and the Gender Gap in Disruptive Behavior." *American Economic Journal: Applied Economics* 5(1): 32-64.

Figure C1: Distributions of achievement, by grade



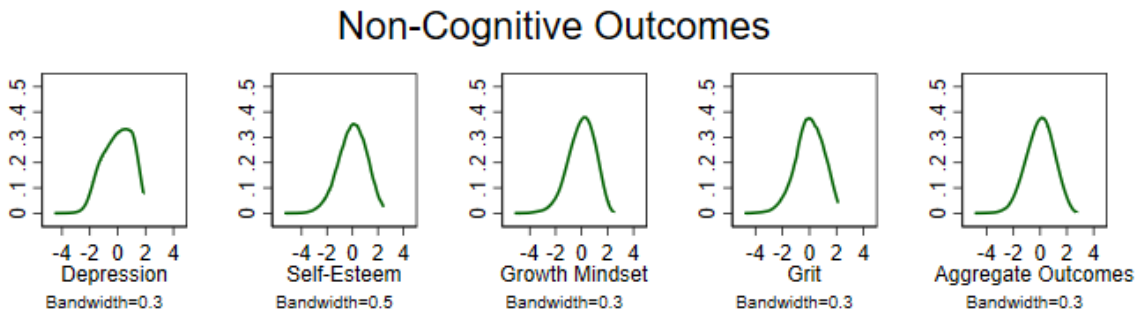
Note: The figure shows univariate densities of achievement, in z-scores, by grade.

Figure C2: Distributions of executive function, by grade



Note: The figure shows univariate densities of executive function, in z-scores, by grade.

Figure C3: Distributions of non-cognitive outcomes, by grade



Note: The figure shows univariate densities of non-cognitive outcomes, in z-scores, by grade.

Table C1: Correlations across dimensions in executive function

	Inhibitory Control	Cognitive Flexibility
	Kindergarten	
Cognitive Flexibility	0.13	
Working Memory	0.22	0.29
	1st Grade	
Working Memory		0.23
	2nd Grade	
Cognitive Flexibility	0.15	
Working Memory	0.25	0.24
	3rd Grade	
Cognitive Flexibility	0.12	
Working Memory	0.17	0.21
	4th Grade	
Cognitive Flexibility	0.15	
Working Memory	0.33	0.32
	Pooled	
Cognitive Flexibility	0.14	
Working Memory	0.24	0.26

Note: The table reports the pairwise correlations between executive function dimensions. All the correlations are significant at the 1 percent level.

Table C2: Correlations across non-cognitive outcomes

	Depression	Self- Esteem	Growth Mindset
Self- Esteem	0.24		
Growth Mindset	0.26	0.49	
Grit	0.20	0.45	0.38

Note: Table presents the results from pairwise correlations between non-cognitive outcomes collected in 6th grade. All the correlations are significant at the 1 percent level.

Table C3: Maternal education gradients and gender differences

	Mothers with complete or incomplete primary school education	Mothers with incomplete secondary school education	Mothers with complete secondary school education or higher
Total Achievement	-0.24	-0.04	0.33
Math	-0.20	-0.04	0.28
Language	-0.25	-0.02	0.33
Executive Function	-0.18	0.00	0.23
Inhibitory Control	-0.07	0.00	0.12
Cognitive Flexibility	-0.10	-0.01	0.14
Working Memory	-0.17	0.00	0.21
Aggregate non-cognitive	-0.13	-0.02	0.18
Depression	-0.14	-0.04	0.20
Self Esteem	-0.09	-0.01	0.12
Grit	-0.05	-0.01	0.07
Growth Mindset	-0.13	-0.02	0.17
Incidence of behavioral problems	0.10	0.11	0.09
	Boys	**Girls**	
Total Achievement	0.01	-0.01	
Math	0.06	-0.06	
Language	-0.05	0.05	
Executive function	-0.01	0.01	
Inhibitory control	0.01	-0.01	
Cognitive flexibility	-0.01	0.01	
Working memory	-0.02	0.02	
Aggregate non-cognitive	-0.09	0.09	
Depression	-0.06	0.06	
Self-esteem	-0.08	0.08	
Grit	-0.08	0.08	
Growth mindset	-0.05	0.05	
Incidence of behavioral problems	0.19	0.03	

Note: Table reports the means by mother's education and gender. All the differences are significant at the 1 percent level, other than the difference in the incidence of behavioral problems by maternal education. 39.5 percent of the sample are children with mothers with complete or incomplete primary school, 29 percent with mothers with incomplete secondary school education and 31,5 percent with complete secondary school education or higher. 51 percent of the children are boys.

Appendix D: Fade-out and differences by grade in executive function and non-cognitive outcomes

This Appendix presents results on fade-out and differences by grade in the effects of the CLASS on executive function and non-cognitive outcomes.

Table D1 has a format comparable to Table 6 in the main body of the text. Table D1 shows that, unlike the results for achievement, there is no evidence that the CLASS effects are larger (or smaller) in the "earlier" grades (kindergarten and 1st grade) than in the "later" grades (2nd grade through 4th grade).

Table D2 analyzes fade-out of CLASS effects on executive function, and has a format comparable to Table 7 in the main body of the text. The table shows that, by and large, the coefficients on the CLASS fade out quickly. In no case are the CLASS effects on executive function significant after two lags.

Table D3, finally, analyzes CLASS effects on the non-cognitive outcome collected in 6th grade, by the grade in which children were exposed to higher- CLASS teachers. These results are very imprecise—there are 40 coefficients in the table, and none of them is significant. There is no pattern whereby the effects of the CLASS are larger (or smaller) in "earlier" than in "later" grades, and we can never reject that the coefficients are the same. By way of caution, we underline that that the CLASS in these estimations refers to grades that are further in the past in some cases—for example, kindergarten—than in others—for example, 4th grade. This is an important consideration to keep in mind if there is substantial fade-out of CLASS effects on depression, self-esteem, growth mindset, and grit.

In sum, the evidence in this Appendix indicates that the effects of the CLASS on executive function and non-cognitive outcomes do not vary with the grade in which children were exposed to higher-quality teachers. In the case of EF, we also show that there is quick fade-out of effects over time.

Table D1: Contemporaneous CLASS effects, by grade

	Binary parametrization of CLASS				Continuous parametrization of CLASS			
	Executive Function	Inhibitory control	Cognitive Flexibility	Working Memory	Executive Function	Inhibitory control	Cognitive Flexibility	Working Memory
Kindergarten	0.044**	0.003	0.008	0.051***	0.128***	0.036	0.051	0.131***
	(0.019)	(0.017)	(0.018)	(0.019)	(0.047)	(0.045)	(0.051)	(0.047)
1st grade	0.038**		0.029*	0.030*	0.147***		0.122***	0.110**
	(0.017)		(0.016)	(0.017)	(0.047)		(0.043)	(0.051)
2nd grade	0.007	0.011	-0.000	0.005	0.066	0.063	0.072	0.031
	(0.017)	(0.017)	(0.016)	(0.018)	(0.050)	(0.053)	(0.048)	(0.051)
3rd grade	0.065***	-0.004	0.034*	0.074***	0.163**	0.055	0.090	0.154**
	(0.019)	(0.018)	(0.018)	(0.018)	(0.063)	(0.056)	(0.055)	(0.060)
4th grade	0.016	0.016	0.011	0.014	0.167**	0.122	0.089	0.154**
	(0.017)	(0.018)	(0.016)	(0.017)	(0.080)	(0.079)	(0.062)	-0.076
F-test (1)	0.16	0.85	0.53	0.04	0.69	0.82	0.87	0.49
F-test (2)	0.48	0.81	0.81	0.55	0.91	0.45	0.95	0.88
R-squared	0.08	0.05	0.05	0.08	0.08	0.05	0.05	0.08
N	82,101	65,790	82,101	82,101	82,101	65,790	82,101	82,101

Notes: Table reports results from regressions of executive on the CLASS. All regressions include school-by-grade fixed effects, child age in months and its square, and child gender. Standard errors are clustered at the school-by-grade level. *, **, and *** indicate significance at the 10 percent, 5 percent, and 1 percent, respectively. F-test (1) is a test that the coefficient in all five grades are equal; F-test (2) is a test that the coefficients on the "earlier" grades (kindergarten and 1st grade) are the same as those on the "later" grades (2nd, 3rd, and 4th grades).

Table D2: Fade-out of CLASS effects on executive function, by grade

	Binary parametrization of CLASS				
	t	t+1	t+2	t+3	t+4
Kindergarten	0.047**	0.021	0.035	0.048**	0.006
	(0.022)	(0.021)	(0.022)	(0.020)	(0.020)
1st grade	0.035*	0.009	-0.016	0.004	
	(0.019)	(0.018)	(0.020)	(0.017)	
2nd grade	0.010	-0.006	-0.022		
	(0.019)	(0.019)	(0.018)		
3rd grade	0.063***	0.040**			
	(0.020)	(0.017)			
4th grade	0.016				
	(0.018)				
F-test (1)	0.32	0.30	0.09	0.08	
F-test (2)	0.52	0.92	0.17		
	Continuous parametrization of CLASS				
	t	t+1	t+2	t+3	t+4
Kindergarten	0.095*	0.005	0.047	0.107**	0.027
	(0.055)	(0.048)	(0.049)	(0.050)	(0.048)
1st grade	0.131**	0.049	-0.017	0.012	
	(0.052)	(0.061)	(0.058)	(0.048)	
2nd grade	0.056	0.013	0.001		
	(0.060)	(0.057)	(0.064)		
3rd grade	0.146**	0.093			
	(0.067)	(0.061)			
4th grade	0.161**				
	(0.080)				
F-test (1)	0.79	0.67	0.67	0.15	
F-test (2)	0.87	0.66	0.85		

Notes: Table reports results from regressions of executive function on the CLASS. All regressions include school-by-grade fixed effects, child age in months and its square, and child gender. Standard errors are clustered at the school-by-grade level. *, **, and *** indicate significance at the 10 percent, 5 percent, and 1 percent, respectively. Sample sizes are 9,076 (kindergarten), 11,197 (1st grade), 12,525 (2nd grade), 13,760 (3rd grade), and 14,733 (4th grade). Sample sizes are smaller in earlier grades because, to be included in the regressions for a given grade, children need to have attended the sample of schools in the study in every grade thereafter. Coefficients on the CLASS in t are not identical to those in Table 6 because of differences in the samples. F-test (1) is a test that the coefficient in all five grades are equal; F-test (2) is a test that the coefficients on the "early" grades (kindergarten and 1st grade) are the same as those on the "late" grades (3rd, 4th and 5h grades)

Table D3: CLASS effects on non-cognitive outcomes in 6th grade

	Depression	Self-esteem	Growth mindset	Grit	Aggregate
Binary parametrization of CLASS					
Kindergarten	0.016	-0.007	0.030	0.006	0.011
	(0.022)	(0.022)	(0.022)	(0.021)	(0.022)
1st grade	-0.004	-0.017	-0.036	-0.005	-0.024
	(0.024)	(0.023)	(0.023)	(0.024)	(0.022)
2nd grade	0.000	0.019	0.027	0.029	0.028
	(0.026)	(0.022)	(0.024)	(0.024)	(0.024)
3rd grade	0.012	-0.025	0.020	-0.006	-0.006
	(0.022)	(0.025)	(0.021)	(0.024)	(0.022)
4th grade	0.008	0.029	0.028	-0.013	0.023
	(0.023)	(0.022)	(0.024)	(0.020)	(0.022)
F-test (1)	0.97	0.38	0.21	0.73	0.45
F-test (2)	0.98	0.34	0.18	0.88	0.29
Continuous parametrization of CLASS					
Kindergarten	0.042	-0.005	0.035	0.049	0.030
	(0.062)	(0.064)	(0.059)	(0.051)	(0.061)
1st grade	-0.019	-0.104	-0.101	-0.057	-0.111
	(0.071)	(0.067)	(0.066)	(0.067)	(0.068)
2nd grade	0.010	0.038	0.092	0.028	0.064
	(0.080)	(0.073)	(0.085)	(0.076)	(0.078)
3rd grade	0.090	-0.100	0.061	-0.079	-0.041
	(0.079)	(0.086)	(0.080)	(0.068)	(0.084)
4th grade	0.033	0.061	0.082	-0.031	0.058
	(0.084)	(0.076)	(0.089)	(0.085)	(0.081)
F-test (1)	0.89	0.38	0.30	0.53	0.37
F-test (2)	0.63	0.41	0.09	0.70	0.31

Notes: Table reports coefficients and standard errors of regressions of non-cognitive outcomes on the CLASS scores. All regressions include school-by-grade fixed effects, child age in months and its square, and child gender. Standard errors are clustered at the school-by-grade level. *, **, and *** indicate significance at the 10 percent, 5 percent, and 1 percent, respectively. F-test (1) is a test that the coefficient in all five grades are equal; F-test (2) is a test that the coefficients on the "early" grades (kindergarten and 1st grade) are the same as those on the "late" grades (3rd, 4th and 5th grades)