Caridad Araujo, Maria; Carneiro, Pedro; Cruz-Aguayo, Yyannú; Schady, Norbert

Working Paper

Teacher Quality and Learning Outcomes in Kindergarten

IZA Discussion Papers, No. 9796

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Caridad Araujo, Maria; Carneiro, Pedro; Cruz-Aguayo, Yyannú; Schady, Norbert (2016) : Teacher Quality and Learning Outcomes in Kindergarten, IZA Discussion Papers, No. 9796, Institute for the Study of Labor (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/141555

DISCUSSION PAPER SERIES

IZA DP No. 9796

Teacher Quality and Learning Outcomes in Kindergarten

M. Caridad Araujo
Pedro Carneiro
Yyannú Cruz-Aguayo
Norbert Schady

March 2016

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Teacher Quality and Learning Outcomes in Kindergarten

M. Caridad Araujo
Inter-American Development Bank

Pedro Carneiro
University College London, IFS, CEMMAP and IZA

Yyannú Cruz-Aguayo
Inter-American Development Bank

Norbert Schady
Inter-American Development Bank

Discussion Paper No. 9796
March 2016

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 9796
March 2016

ABSTRACT

Teacher Quality and Learning Outcomes in Kindergarten[1]

We assigned two cohorts of kindergarten students, totaling more than 24,000 children, to teachers within schools with a rule that is as-good-as-random. We collected data on children at the beginning of the school year, and applied 12 tests of math, language and executive function (EF) at the end of the year. All teachers were filmed teaching for a full day, and the videos were coded using a well-known classroom observation tool, the Classroom Assessment Scoring System (or CLASS). We find substantial classroom effects: A one-standard deviation increase in classroom quality results in 0.11, 0.11, and 0.07 standard deviation higher test scores in language, math, and EF, respectively. Teacher behaviors, as measured by the CLASS, are associated with higher test scores. Parents recognize better teachers, but do not change their behaviors appreciably to take account of differences in teacher quality.

JEL Classification: I24, I25

Keywords: teacher quality, learning, test scores

Corresponding author:

Pedro Carneiro
Department of Economics
University College London
Gower Street
WC1E 6BT, London
United Kingdom
E-mail: p.carneiro@ucl.ac.uk

[1] This project was financed by the Inter-American Development Bank and the Government of Japan. We thank the Ministry of Education in Ecuador for their steady collaboration; Jere Behrman, Raj Chetty, Ariel Fiszbein, Brian Jacob, Tom Kane, Larry Katz, Santiago Levy, Jennifer LoCasale-Crouch, Karthik Muralidharan, Dick Murnane, Hugo Ñopo, Bob Pianta, Hector Salazar, Emiliana Vegas, Hiro Yoshikawa, five anonymous referees, and seminar participants at the Universities of Harvard, Michigan, Pennsylvania, Virginia, Yale, UCL, and the Latin American Impact Evaluation Network for their comments; Sara Schodt for tirelessly overseeing the process of training and supervision of CLASS coders; and Rodrigo Azuero, Jorge Luis Castañeda, Maria Adelaida Martínez, and Cynthia van der Werf for outstanding research assistance, including training and supervision of the enumerators. Carneiro thanks the financial support from the Economic and Social Research Council for the ESRC Centre for Microdata Methods and Practice (grant reference RES-589-28-0001), the support of the European Research Council through ERC-2009-StG-240910 and ERC-2009-AdG-249612. All remaining errors are our own. The findings, interpretations, and conclusions expressed in this paper are those of the authors and do not necessarily represent the views of the Inter-American Development Bank, its Executive Directors, or the governments they represent.

1. Introduction

Teacher quality has been at the center of the academic and policy discussion on education in recent years. It is generally accepted that teachers, even teachers within the same schools, vary widely in the impact they have on student learning. However, there is still considerable debate about how best to measure teacher effectiveness and its correlates.

A substantial amount of research has shown that readily observable teacher characteristics—experience, education, and contractual status, among others—explain very little of the differences in teacher quality (Hanushek and Rivkin 2012; Rivkin et al. 2005). This finding has led researchers (and policy-makers) to take one of two approaches. The main approach in economics has been to measure value added. In this approach, the quality of a teacher is equated with the increase in learning in her classroom, measured by gains in test scores. In contrast, most recent work in education and psychology has focused on what happens inside the classroom. In this approach, the quality of a teacher is measured by the quality of the interactions between teachers and children.

In spite of the increase in the number of studies of teacher effectiveness in recent years, much remains to be learned. With one important exception, the influential Measuring Effective Teaching (MET) project (Kane and Staiger 2012), no prior study has combined the measurement of teacher value added with systematic classroom observation of teachers.

Moreover, most of the evidence on the short-term effects of teachers has focused on test scores in math and language. Little is known about the extent to which good teachers can also affect non-cognitive outcomes, or improve the ability of a child to focus on, stick with, and carry out a given task. This is important because these traits are necessary to succeed in school, and are well-rewarded in the labor market thereafter (Heckman and Kautz 2012).

The vast majority of studies of teacher value added have focused on developed countries generally, and on the United States specifically. However, concerns about teacher quality are likely to be just as important in developing countries, where learning outcomes are frequently dismal. In India 31 percent of 3rd grade children could not recognize simple words (Kremer et al. 2013), and in the Dominican Republic 75 percent of 3rd grade children could not solve simple addition problems (Berlinski and Schady 2015). In settings such as these, establishing how much teachers vary in their effectiveness, and why this is so, are critical policy questions.

Finally, little is known about the extent to which other agents, notably parents, respond to differences in teacher quality by changing investments in their children. This is important as it affects the interpretation of "teacher effects" (Todd and Wolpin 2003). For example, if parents

attempt to compensate for poor teachers by investing more in their children, one might incorrectly conclude that teachers do not matter very much.[2]

In this paper we study the impact of teachers using unusually rich data from Ecuador, a middle-income country. We focus on children entering kindergarten. The first years of formal education are particularly important because they lay the foundation for, and may determine the returns to, all subsequent schooling investments (Cunha and Heckman 2007; Shonkoff and Phillips 2000).

We assigned two cohorts of children entering kindergarten, totaling more than 24,000 students, to different classrooms using a rule that is as-good-as-random. Compliance with the assignment rule was almost perfect. Random assignment means we can convincingly deal with the identification challenges that have been the subject of much controversy in this literature (for example, Chetty et al. 2014a, and Rothstein 2010, among many other contributions).

We collected very rich data on students. At the end of the school year, we tested children in math and language (eight separate tests). We also tested children's inhibitory control, working memory, capacity to pay attention, and cognitive flexibility. These processes, jointly known as "executive function" (EF), measure a child's ability to regulate her thoughts, actions, and emotions, all of which are central to the learning process (Anderson 2002; Espy 2004; Senn et al. 2004).

We also collected very rich data on teachers. In addition to standard information on years of experience, education, and contract status (tenured or not), we measured teacher IQ, personality, attention and inhibitory control, and parental education. Moreover, we filmed teachers teaching a class for an entire school day. We coded these videos to measure the interaction of teachers and students, using a protocol known as the Classroom Assessment Scoring System (CLASS, Pianta et al. 2007). The CLASS is a measure of a series of teacher behaviors that can collectively be described as "Responsive Teaching" (Hamre et al. 2014).

Finally, we collected household data for the children in our study, which include a parental assessment of teacher quality (on a Likert-like 5-point scale), inputs into child development and learning (including the availability of books, pencils, and toys of various kinds), and parental behaviors (including whether parents read to, sang to, or played with, their children). These data were collected towards the end of the school year. They allow us to test whether parents alter their

[2] Pop-Eleches and Urquiola (2013) use a regression discontinuity design to show that, in Romania, parents whose children have been accepted at a more selective school are less likely to help their children with their homework, suggesting that they view their own efforts and school quality as substitutes.

investments and behaviors in ways that reinforce the effect of a good teacher, or compensate for a bad one.

The strength of our identification strategy and the data we collected allow us to make several important contributions to the literature on teacher effectiveness. The first set of results in our paper focuses on teacher value added. We provide the first experimental estimates of teacher value added on math and language in a developing country. These estimates suggest that teachers vary considerably in their effectiveness, in a magnitude that is comparable to what is observed in the United States (see, for example, Chetty et al. 2014a, and Jackson et al. 2014).

We also provide the first estimates of classroom (as opposed to teacher) effects on executive function, in a developed or developing country. EF in young children strongly predicts learning trajectories and long-term outcomes, including in the labor market (Moffitt et al. 2011; Séguin and Zelazo 2005). A number of papers have shown that teacher effects on test scores depreciate quickly, before stabilizing at around one-quarter of their original value (Jackson et al. 2014; Jacob et al. 2010; Rothstein 2010). Despite this fade-out, there is strong evidence of teacher effects on long-term outcomes, including college attendance, earnings, and the likelihood of becoming parents as teenagers (Chetty et al. 2011; Chetty et al. 2014b). If EF has effects on long-term outcomes that are not captured by test scores, then the classroom effects on EF we estimate may help reconcile the apparent paradox of the long-term importance of teachers in spite of the rapid depreciation of teacher effects on test scores.

The second set of results focuses on the correlates of teacher effectiveness. As has been found in many other settings, children assigned to "rookie" teachers (those who have three years of experience or less) learn less on average. All the other characteristics of teachers, including tenure status, teacher IQ, the Big Five personality traits, and inhibitory control and attention, do not consistently predict test scores.

We then turn our attention to the quality of student-teacher interactions. We show that teacher behaviors, as measured by the CLASS, are significantly associated with learning in math, language, and executive function. These findings support the conclusions from two recent reviews (Kremer et al. 2013; Murnane and Gaminian 2014), both of which argue that changing pedagogical practices is the key to improving learning outcomes in the developing world. Our results also complement estimates from the MET project, which show that students randomly assigned to

teachers with higher-quality interactions learn more in the United States (Kane and Staiger 2012),[3] and fixed effects estimates that show that teaching practices—in particular, the extent to which teaching is "horizontal" (students working in groups), rather than "vertical" (lectures by teachers)—predict differences in social capital across countries, across schools within countries, and across classrooms within schools (Algan et al. 2013). Taken together, these results and ours support the notion that *how* children are taught may have implications for a variety of outcomes that are economically important.

Finally, we show that in Ecuador parents can discriminate between good and bad teachers: On average, they give higher scores to teachers who produce more learning, teachers with better CLASS scores, and teachers with more experience. However, parents do not appear to substantially change their inputs and behaviors, at least the ones we observe, in response to differences in teacher quality. This suggests that the "teacher effects" we estimate are largely the *direct* effect of teachers on learning (although we cannot rule out that there are changes in other unobserved parental inputs).

2. Setting and data

A. Setting

Ecuador is a middle-income country. It is one of the smaller countries in South America, roughly equal to Colorado in size, and has a population of 16.3 million. GDP per capita (in PPP US dollars) is 11,168, similar to neighboring countries like Colombia and Peru.

Schooling in Ecuador 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. An unusual feature of the education system in Ecuador is that the school year runs from May to February in the coastal area of the country (like most countries in the southern hemisphere), but from September through June in the highlands and *Oriente* regions of the country (like most countries in the northern hemisphere).[4]

There are 3.6 million children in the education system in Ecuador, 80 percent of who attend public schools; the remaining 20 percent attend private schools. There are more than 150,000 public

[3] Our study is conceptually similar to the MET project, but compares favorably in terms of design, compliance, and attrition. Contamination of the experiment was a serious issue in MET: Across the six different sites, compliance with the random assignment ranged from 66 percent (in Dallas) to 27 percent (in Memphis). Also, the sample of teachers in MET is drawn from those who *volunteered* to be in the study, which may limit external validity. Finally, there was substantial attrition: about 40 percent of the 4th through 8th grade sample in MET, and more than half of the high school sample, were lost to follow-up in a single year.

[4] According to the 2010 Census, 53 percent of the population lived on the coast, 42 percent lived in the highlands, and 5 percent lived in the sparsely populated *Oriente*.

sector teachers. The teacher salary scale is overwhelmingly determined by seniority. Roughly two-thirds of teachers in Ecuador are tenured, while the other third works on a contract basis.

Ecuador has made considerable progress expanding the coverage of the education system. In kindergarten, the focus of our paper, enrollment rates in 2012 were 93 percent, compared to 70 percent in 2000.[5] However, many children, especially the poor, appear to learn little in school. On a recent international test of 3rd grade children, 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. Sample and data

The bulk of our analysis focuses on a cohort of children entering kindergarten in the 2012 school year. In addition, we collected more limited data on another cohort of children entering kindergarten in the same schools, and assigned to the same teachers, in the 2013 school year.

We used administrative data from the Ministry of Education to construct our sample as follows. First, we limited the sample to full-time, public elementary schools on the coast (where the school year runs from May to February). Second, we limited the sample to schools that had at least two kindergarten classes. Finally, we took a random sample of 204 schools. We refer to this as our study sample of schools. Within these schools we attempted to collect data on all kindergarten children and their teachers.

The 204 schools in our study sample are a selected sample—they cover only schools with two or more kindergarten classrooms in the coastal region of the country. To get a sense of the external validity of our study, we also collected data on a nationally-representative sample of public, full-time elementary schools in Ecuador.[6]

[5] This is the fraction of children of five years of age who are enrolled in school, based on our calculations using the 2000 and 2012 *Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU)*, a nationally representative household survey in Ecuador.

[6] We drew a stratified, population-weighted sample of 40 schools from the coast, 40 from the highlands, and 20 from the *Oriente* region, with the weights given by the number of children enrolled in kindergarten in the 2012 school year. Within each school, we then collected data on all kindergarten teachers and a random sample of 10-11 kindergarten children.

(i) Baseline characteristics of children and families

We collected data on child gender and age. In addition, in the study sample (but not the national sample), we tested children at the beginning of the school year with the *Test de Vocabulario en Imágenes Peabody* (TVIP), the Spanish version of the widely-used Peabody Picture Vocabulary Test (PPVT) (Dunn et al. 1986). Performance on this test at early ages has been shown to predict important outcomes in a variety of settings, including in Ecuador.[7]

For the 2012 cohort (but not the 2013 cohort), we applied a household survey to the families of children in the study and national samples. This survey, which was fielded between November 2012 and January 2013 (close to the end of the school year), asked respondents (in most cases, the mother of the focal child) to rate the school and (separately) teacher on a 1-5 Likert scale. In addition, it included questions about basic household assets, living conditions, household composition, and the education level of all household members. Finally, parents were asked about learning inputs and activities.

Table I, upper panel, summarizes the baseline characteristics of children and families. Children are approximately five years of age on the first day of classes. Just under half of the children in the sample are girls, as expected. Mothers are on average in their early thirties, while fathers are in their mid-thirties. Education levels are similar for both parents—just under nine years of school (which corresponds to completed middle school). Sixty-one percent of children in the study sample (70 percent in the national sample) attended preschool. The average child in the study sample has a TVIP score that places her more than one standard deviation below the reference population that was used to norm the test, indicating that many children begin formal schooling with deep delays.[8] The children and households in the study and national samples are generally similar to each other, suggesting that the children we study in this paper are broadly representative of young children in kindergarten in Ecuador.

[7] Schady (2012) shows that children with low TVIP scores before they enter school are more likely to repeat grades and have lower scores on tests of math and reading in early elementary school in Ecuador; Schady et al. (2015) show that many children in Ecuador start school with substantial delays in receptive vocabulary, and that the difference in vocabulary between children of high and low socioeconomic status is constant throughout elementary school. Important references from the United States include Case and Paxson (2008), who show that low performance on the PPVT at early ages predicts wages in adulthood; and Cunha and Heckman (2007) who show that, by age 3 years, there is a difference of approximately 1.2 standard deviations in PPVT scores between children in the top and bottom quartiles of the distribution of permanent income, and that this difference is largely unchanged until at least 14 years of age.

[8] The TVIP was standardized on a sample of Mexican and Puerto Rican children. The test developers publish norms that set the mean at 100 and the standard deviation at 15 at each age (Dunn et al. 1986).

(ii) Data on child learning outcomes at the end of kindergarten

Data for the 2012 kindergarten cohort: We applied twelve separate tests at the end of the school year (between the 1st of November and the 21st of January). These tests were applied to children individually (as opposed to a class as a whole), and in total took between 30 and 40 minutes per child. Most children were tested in school, in a room that had been set up for this purpose.[9]

Ninety-six percent of children who attended kindergarten in our sample of study schools in the 2012 school year (including those who dropped out at some point during the school year) completed all 12 tests. The non-response rate, 4.4 percent, is very low relative to that found in other longitudinal studies of learning outcomes in developing countries.[10]

We applied four tests of language and early literacy, which covered child vocabulary, oral comprehension, and sound, letter, and word recognition. All of these are foundational skills, and strongly predict reading acquisition in early elementary school and beyond (Powell and Diamond 2012; Wasik and Newman 2009). We also applied four tests of math, which covered number recognition, sequencing, applied math problems, and identification of basic geometric figures. Mathematical knowledge at early ages has been shown to strongly predict later math achievement in a number of longitudinal studies (Duncan et al. 2007; Siegler 2009).

Finally, we applied four tests of executive function (EF). EF includes a set of basic self-regulatory skills which involve various parts of the brain, but in particular the prefrontal cortex.[11] It is an important determinant of how well young children adapt to and learn in school. Basic EF skills are needed to pay attention to a teacher; wait to take a turn or raise one's hand to ask a question; and remember steps in, and shift from one approach to another, when solving a math problem, among many other tasks that children are expected to learn and carry out in the classroom. Children with high EF levels are able to concentrate, stay on task, focus, be goal-directed, and make good use of learning opportunities. Low levels of EF are associated with low levels of self-control and "externalizing" behavior, including disruptive behavior, aggression, and inability to sit still and pay

[9] We attempted to test children who could not be located in school in their homes—a total of 390 children were tested at home.

[10] For example, in their analysis of school incentives in India, Duflo et al. (2012) report an attrition rate of 11 percent in the treatment group and 22 percent in the control group at mid-test; for the post-test, comparable values are 24 percent and 21 percent, respectively; in their analysis of student tracking in Kenya, Duflo et al. (2011) report a test non-response rate of 18 percent; and in their analysis of the effect of school vouchers in Colombia, Angrist et al. (2002) report a response rate for the tests of 60 percent, implying attrition of 40 percent.

[11] Volumetric measures of prefrontal cortex size predict executive function skills; children and adults experiencing traumatic damage to the prefrontal cortex sustain immediate (and frequently irreversible) deficits in EF (Nelson and Sheridan 2011, cited in Obradovic et al. 2012).

attention, which affects a child's own ability to learn, as well as that of her classmates (Séguin and Zelazo 2005).

Low levels of executive function in childhood carry over to adulthood. A longitudinal study that followed a birth cohort in New Zealand to age 32 years found that low levels of self-control in early childhood are associated with lower school achievement, worse health, lower incomes, and a higher likelihood of being involved in criminal activity in adulthood, even after controlling for IQ and socioeconomic status in childhood (Moffitt et al. 2011). A growing literature in economics stresses the importance of various behaviors (attentiveness, showing up on time) and abilities (flexibility, learning on the job, capacity to work with others), all of which have an EF component, in determining labor market outcomes (for summaries, see Heckman 2013; Heckman and Kautz 2012 and the references therein).

Although there have been competing definitions of executive function and how to measure it, there is a growing consensus that it includes three broad domains: inhibitory control, working memory, and cognitive flexibility. Sometimes, attention is added as separate domain. *Inhibitory control* refers to the ability to suppress impulsive behaviors and resist temptations; *working memory* refers to the ability to hold, update, and manipulate verbal or non-verbal information in the mind for short periods of time; *cognitive flexibility* refers to the ability to shift attention between competing tasks or rules.[12] *Attention* is the ability to focus and disregard external stimuli, which is why it is often grouped with working memory. We applied age-appropriate tests of all four EF domains.

We normalize each of the twelve end-of-year tests by subtracting the mean and dividing by the standard deviation of the national sample. We then create three test aggregates for language, math, and executive function, respectively. Each of the four tests within an aggregate receives the same weight. Like the underlying tests, the aggregates are normalized to have zero mean and unit standard deviation.

Data on the 2013 kindergarten cohort: We also collected data on a new cohort of kindergarten children who were taught by the same teachers in the study sample of schools in 2013. We applied four of the twelve tests to this new cohort—two tests of language (letter and word recognition, and the TVIP) and two tests of math (number recognition, and applied problems). Further details on the various tests, including on the distribution of scores, are given in Online Appendix A.

[12] This categorization and the definitions very closely follow Obradovic et al. (2012, pp. 325-27).

(iii) Data on teachers

Data on teacher characteristics: Table I, lower panel, provides sample averages for those teacher characteristics that are generally available in administrative data. There is no variation in degree—all teachers have a teaching degree and no further education. Virtually all teachers are women. The average teacher is in her early 40s. Sixty-four percent of teachers in the study sample are tenured, and 6 percent are "rookies" (have three years of experience or less).[13] The average class size in the study sample is 34 children.[14] The study and national samples generally appear to be quite similar to each other, with the exception of the proportion tenured, which is substantially (22 percentage points) higher in the national sample.

A growing body of evidence suggests that both intelligence and personality are important determinants of success in the labor market and elsewhere (Almlund et al. 2011). To see how these characteristics correlate with teacher effectiveness, we collected IQ, the Big Five personality test, a test of attention and inhibitory control, and data on early circumstances (for example, the education level of both parents) for teachers in the study (but not the national) sample.

To measure teacher IQ, we used the Spanish-speaking version of the Wechsler Adult Intelligence Scale (WAIS-III; the Spanish-speaking acronym is EIWA-III; see Wechsler 1939, and subsequent updates). The WAIS is a widely-used test of IQ. To measure teacher personality, we used the Big Five. The Big Five is a widely accepted taxonomy of personality traits (Almlund et al. 2011; Costa and McCrae 1992; Goldsmith et al. 1987). It focuses on five traits: neuroticism, extraversion, openness, agreeableness, and conscientiousness. The test consists of 240 questions, 48 for each trait, which the respondent answers on a 5-point, Likert-like scale (where 1 is "totally disagree" and 5 is "totally agree").[15] To measure attention and inhibitory control, we applied a test in

[13] We ran regressions in which we defined "rookies" as teachers with various levels of experience (0 years, 0-1 years, 0-2 years… 0-N years) and chose the value for years of experience which maximized the R-squared of a regression of learning outcomes on experience. Much as in the United States (Rivkin et al. 2005; Staiger and Rockoff 2010), the returns to experience in Ecuador rise sharply in the first three years, and then flatten out.

[14] Only 6.4 percent of teachers in the sample report having had a teacher's aide.

[15] Measurement error is a concern for both the IQ and Big Five tests (see Borghans et al. 2008 for a discussion). In the extreme, both tests could be mainly noise. Under these circumstances, the correlations across tests, or across dimensions within a test, would generally be insignificant. This is not the case in our data. In fact, the correlations we observe are similar to others reported in the literature. The Big Five trait that has most consistently been linked with intelligence is openness (Ackerman and Heggestad 1997; Moutafi et al. 2004). Austin et al. (2002) report that the "typical correlation magnitude" between IQ and openness is 0.3 (in our data it is 0.27), between IQ and neuroticism is -0.1 (in our data it is -0.07) and between IQ and extraversion is 0.1 (in our data it is 0.07). Three of the Big Five traits (neuroticism, openness, and agreeableness) are also significantly correlated with the CLASS. Van der Linden et al. (2010) report the results of a meta-analysis of the correlations across different traits in the Big Five in 212 samples. They show that neuroticism is negatively associated with the other four traits, with correlations that on average range from -0.12 (with openness) to -0.32 (with conscientiousness). In our data, neuroticism is negatively correlated with openness, conscientiousness, and

which subjects are quickly shown sets of incongruent stimuli (for example, the word "red" printed in blue ink) and are asked to inhibit a habitual or automated response (in this case, they would be asked to name the color of the ink rather than read the word) (Jensen and Rowher 1966; MacLeod 1991; Stroop 1935). Finally, we asked teachers about their parents' education and about a number of characteristics of their homes when they were children.

Data on teacher behaviors: The main measure of teacher behaviors (or interactions) we use in this paper is the CLASS (Pianta et al. 2007). A number of papers using US data have found that children exposed to teachers with better CLASS scores have higher learning gains, better self-regulation, and fewer behavioral problems (references on one or more outcomes include Howes et al. 2008 for pre-k; Grossman et al. 2010 for middle school; and Kane and Staiger 2012 for the MET study).

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" (scores of 3-5), or "high" (scores of 6-7). In practice, in our application of the CLASS (as well as in others), scores across different dimensions are highly correlated with each other. For this reason, we focus on a teacher's *total* CLASS score (given by the simple average of her scores on the 10 dimensions). We take this score to be a measure of Responsive Teaching (as in Hamre et al. 2014). As with the child tests, we normalize IQ, the Big Five, inhibition and attention, and the CLASS to have mean zero and unit standard deviation.[17]

To apply the CLASS in Ecuador, we filmed all kindergarten teachers in both our study and national samples of schools in the 2012 school year. In addition, in the study sample of schools, we filmed all kindergarten teachers in the previous (2011) school year. Teachers were filmed for a full day (from approximately eight in the morning until one in the afternoon); they did not know on

---

agreeableness (with correlations that range from -0.04 to -0.30), but not with extraversion. All the other correlations reported by Linden et al. are positive, with correlations that on average range from 0.14 to 0.31. In our data, these correlations are all positive as well, and range from 0.26 to 0.39.

[16] 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.

[17] We use the study sample of teachers for norming because IQ, the Big Five, the test of attention and inhibitory control and parental education were not collected for the national sample. Our results are very similar if we norm the CLASS with the national rather than the study sample.

what day they would be filmed until the day itself. Further details on the process of CLASS filming and coding are given in Online Appendix B.

Figure I graphs univariate densities of the distribution of total CLASS scores in the study and national samples. The average score is 3.7 in both samples. A few teachers have CLASS scores in the "low" range (scores of 1 or 2) but the vast majority, more than 80 percent, have scores between 3 and 4. In the figure we also graph the distribution of CLASS scores in a nationally-representative sample of 773 kindergarten classrooms in the United States (Clifford et al. 2003). The average CLASS score in this sample is 4.5. The difference in scores between the US and Ecuador samples is substantial, equivalent to 1.6 standard deviations of the US sample and 2.6 standard deviations of either of the two Ecuador samples. Fifteen percent of the teachers in the US sample, but none of the teachers in the Ecuador samples, have scores of 5 or higher.

(iv) As-good-as-random assignment

In order to identify teacher effects on learning we rely on as-good-as-random assignment of students to teachers, within schools. Because kindergarten is the first year of the formal education cycle in Ecuador, and because we needed an assignment rule that school headmasters could follow, and we could verify, this was no simple task.

For both the 2012 and 2013 school years, the assignment of children to teachers was carried out as follows. Working in close coordination with staff from the Ministry of Education, we ordered all children who had signed up for kindergarten in a given school by their last name and, if there was more than one child with the same last name, by their first name. We then went down the list and assigned children to kindergarten classrooms in alternating order.[18] Compliance with our assignment rule was almost perfect: Only 1.7 percent of children were found to be sitting in classrooms other

[18] For example, in a school with two kindergarten classes, and the following order of children (María Andrade, Jose Azuero, Miguel Azuero, Rosa Bernal … Emiliano Zapata), María Andrade and Miguel Azuero would be assigned to kindergarten classroom A, and Jose Azuero and Rosa Bernal would be assigned to kindergarten classroom B. If there were three kindergarten classes, María Andrade and Rosa Bernal would be assigned to classroom A, Jose Azuero to classroom B, and Miguel Azuero to classroom C. In the 2012 cohort of students, we gave no explicit instructions to headmasters about how teachers should be assigned to the classes formed with the as-good-as-random assignment rule. In principle, this could have allowed headmasters to purposefully assign teachers to one or another classroom, perhaps taking account of any misbalance of students that occurred in spite of random assignment. We do not believe that this was the case for a number of reasons. First, as we show in Online Appendix Table C1, observable teacher and student characteristics are orthogonal with each other in our sample. Second, when assigning the 2013 cohort of students to teachers we not only created the lists with the as-good-as-random rule described above but, in addition, randomly assigned classes to teachers. Compliance with this second stage of random assignment was 100 percent. The classroom effects estimated with the second cohort of students are indistinguishable from those that use the first cohort, as we show below.

than those they had been assigned to in one or both unannounced visits to schools to verify compliance. To avoid biases, we include these children in the classrooms they were assigned to, rather than those they were sitting in during the school visits. In this sense, our estimates correspond to Intent-to-Treat parameters (with a very high level of compliance with treatment). Further details and tests on the assignment protocol are provided in Online Appendix C.

3. Estimation strategy and results

Our analysis has two parts. We first estimate within-school differences in end-of-year test scores between children assigned to different kindergarten classrooms (controlling for baseline TVIP scores and other characteristics). These estimates allow us to recover the within-school variance of classroom quality. We then relate these differences in student test scores to differences in teacher characteristics and behaviors.

A. Classroom and teacher effects

To estimate classroom effects, we first regress end-of-year test score Y on test aggregate k of child i in classroom c in school s on a set of classroom indicators, child age, gender, whether she attended preschool before kindergarten, baseline child TVIP, mother's education, and the variables of housing conditions and assets we collected in the household survey.[19] We also include a dummy variable for children who were tested at home, rather than in school. These baseline controls, in particular the TVIP, are important to improve the balance of student characteristics across classrooms within the same school. The regression equation is:

$$(1) \qquad Y_{ics}^{k} = \delta_{cs}^{k} + \boldsymbol{X_{ics}}\beta_{1}^{k} + \varepsilon_{ics}^{k} \qquad k = 1,2,3$$

where δ_{cs}^{k} are classroom indicators, $\boldsymbol{X_{ics}}$ is a vector of child and household characteristics, and ε_{ics}^{k} is an i.i.d. error term. The classroom indicators δ_{cs}^{k} are the basis for our calculation of teacher and classroom effects.

[19] Mother's education is discretized to have five mutually exclusive categories: incomplete elementary schooling or less; complete elementary; incomplete secondary school; complete secondary; and some tertiary education. We are missing data on one or more characteristics for 829 children (6.0 percent) who attended kindergarten at some point in the 2012 school year. For these children, we replace the missing value with the mean for the sample (or the mode, when the variable is coded as multiple dummies, as in the case of mother's education), include a dummy variable that takes on the value of one if that variable is missing, and interact the dummy variable with the variables for the other characteristics (thus allowing the slope of the association between outcomes and a given variable X_1 to be different for children who are, or are not, missing a different variable X_2).

In practice, a number of estimation challenges arise. The first is that it is not possible to separate classroom and school effects. Random assignment of children to teachers took place *within* schools, so sorting (of teachers or children) may be an issue across, but not within, schools. Therefore, following (among others) Chetty et al. (2011) and Jacob and Lefgren (2008), we redefine each classroom effect relative to the school: $\gamma_{cs}^{k} = \delta_{cs}^{k} - \frac{\sum_{c=1}^{C_s} N_{cs}\delta_{cs}^{k}}{\sum_{c=1}^{C_s} N_{cs}}$, where γ_{cs}^{k} is the demeaned classroom effect, C_s is the number of kindergarten classrooms in school s and N_{cs} is the number of students in classroom c in school s. Of course, there are likely to be differences in classroom quality across schools. Because we ignore these cross-school differences, our estimates provide a lower bound on the total variation in classroom quality in our sample.[20]

As in much of the literature, we focus on estimating $V(\gamma_{cs}^{k})$, ($V(.)$ indicates variance). A complication arises since $V(\gamma_{cs}^{k})$ overestimates the true variance of classroom effects, because of sampling error. To purge $V(\gamma_{cs}^{k})$ of sampling error we estimate the variance of the sampling error (roughly following the procedure in Appendix B of Chetty et al. 2011), and then subtract it from the variance of the *estimated* classroom effects to obtain the variance of the *true* classroom effects—see Appendix D for details.[21]

As is well known in the literature, it is important not to confuse classroom and teacher effects. Specifically, the δ_{cs}^{k} (and γ_{cs}^{k}) parameters include both differences in teacher quality across classrooms, and random classroom shocks. These shocks could include the presence of a particularly difficult student who disrupts learning; the way in which children in a classroom relate to each other; or disruptions on the day in which the end-of-year tests were taken (for example, if children in a classroom have the flu, or are distracted by construction outside the classroom). With only one year of data, differences in teacher quality cannot be separated from these shocks, so the γ_{cs}^{k} provide estimates of classroom (not teacher) effects, and $V(\gamma_{cs}^{k})$ is the within-school variance in classroom quality (not teacher quality).

With two (or more) years of data, on the other hand, it is possible to separate classroom and teacher effects. Recall that, for a subset of math and language tests, we have estimates of learning

[20] Just over half of the variance in the distribution of CLASS scores in the study schools is across, rather than within, schools. If the distribution of classroom quality is similar to the distribution of the CLASS, we would be under-estimating the variance in classroom quality roughly by a factor of two (or would be under-estimating the standard deviation of classroom quality by a factor of 1.4).

[21] This is analogous to the empirical Bayes approach in Kane and Staiger (2002) and many subsequent papers, but our calculation of the variance of the sampling error explicitly accounts for the fact that the classroom effects are demeaned within each school, and that the within-school mean may also be estimated with error.

outcomes for two cohorts of children taught by the same teacher. Given random assignment of students to teachers, classroom shocks should be uncorrelated across the two cohorts. We can calculate $COV(\gamma_{cs}^{t}, \gamma_{cs}^{t+1})$. As discussed in Hanushek and Rivkin (2012) and McCaffrey et al. (2009), the square root of this covariance is an estimate of the standard deviation of the teacher effects, purged of classroom shocks.[22]

Table II reports the standard deviation of classroom effects (first four columns) and teacher effects (last column) for various test aggregates and breakdowns of the data. The first column refers to the largest possible estimation sample: all children with valid test data enrolled in kindergarten in the 204 study sample of schools. The next column limits the sample to children who were taught by the *same* teacher for the entire 2012 school year. The remaining columns refer to classrooms taught by teachers who taught kindergarten in the study sample of schools in both the 2012 and 2013 school years, and consider only the smaller set of tests applied to both cohorts.

There are a number of important results in the table. First, a one-standard deviation increase in classroom quality, corrected for sampling error, results in 0.11 standard deviations higher test scores in both language and math. These results are close to those reported in the US literature.[23]

Second, we estimate classroom effects of 0.07 standard deviations on executive function. We are aware of only a handful of earlier papers that use non-experimental methods to explore the effects of some education intervention (like access to preschool or a reformed curriculum) on measures of child EF (Bierman et al. 2008; Gormley et al. 2011; Weiland and Yoshikawa 2013). They do not explicitly estimate classroom effects on EF. Early measures of EF are potentially very important for future outcomes, and providing credible estimates of classroom effects on EF is a major contribution of this paper.[24]

---

[22] The focus on this literature is generally on value added measures of teacher quality. Since we have twelve end-of-year tests and only one beginning-of-year test, the TVIP, we can only compute a standard value added measure for the TVIP. For the other tests the most we can do is to control for beginning of year TVIP scores (and other covariates), to construct an approximate value added measure. Nevertheless, given that (1) we randomize students to teachers within each school, and (2) we define teacher quality relative to a school mean, the variance of teacher effects assessed using a standard value added measure and our approximate value added measure should be the same. This is because randomization ensures that any potential beginning-of-year test results on the twelve tests we apply at the end of the year would be the same on average for students assigned to each teacher within a school, and therefore should not affect differences in value added relative to a school mean.

[23] We note, however, that in the US literature classroom effects on math tend to be larger than on language (Jackson et al. 2014). This does not appear to be the case in Ecuador.

[24] Our results on classroom effects on executive function relate to a small but growing literature on teacher effects on school outcomes other than test scores. Jackson (2014) uses a large sample of schools in North Carolina and estimates teacher effects for high-school students on a composite measure of absences, suspensions, grades, and on-time grade progression (which he refers to as a "non-cognitive factor"). Using multiple cohorts of students with the same teacher, he finds teacher effects on this composite measure that are larger than teacher effects in either English or math, although

Third, we estimate teacher (rather than classroom) effects of 0.09 on language and math. Moreover, by comparing the classroom and teacher effects we can approximate the magnitude of the classroom shocks. If we assume that 0.11 is the true value of the teacher effect on language and math, this implies a classroom shock of roughly two-thirds the magnitude of the teacher effect (with standard deviation = 0.063, since $\sqrt{(0.063^2 + 0.09^2)} = 0.11$).[25] These estimates show that, much as has been found in the United States, classroom shocks can have sizeable effects on learning. If we assume that classroom shocks are roughly two-thirds of teacher effects for executive function, then teacher effects on EF would be about 0.05 standard deviations.

Fourth, although this is not reported in the table, we examine whether the classroom effects on one subject are correlated with those for other subjects. If we focus on the 2012 cohort only, the classroom effects on math and language have a correlation of 0.71. The correlation between the classroom effects on the "academic" subjects and executive function is somewhat lower, but still sizeable (0.50 for math, and 0.54 for language). These results suggest that there was more learning in some classrooms than in others in *all* of the subjects we measure.

Finally, we estimate the correlation across years for the subset of teachers and tests where data are available for two cohorts of children. The cross-year correlation in classroom effects on language is 0.42, and the cross-year correlation in math is 0.32; the cross-year correlations between classroom effects on language in one year and math in the other year are 0.28-0.38. The fact that the cross-year correlations are substantially lower than those that use data from a single year confirm the importance of classroom shocks, or non-permanent components of teacher effectiveness. However, the general picture that emerges from these correlations and the other results reported above is that some kindergarten teachers are more effective than others, in multiple subjects, year after year.[26]

---

the teacher effects in English (0.03 standard deviations) and math (0.07 standard deviations) are smaller than most of those reported in the literature. In a similar vein, Jennings and DiPrete (2010) report that kindergarten teacher effects on 1st grade behavioral problems (as reported by 1st grade teachers) are larger than the corresponding teacher effects in reading or math.

[25] We thank an anonymous referee for this suggestion.

[26] The content matter for different subjects in kindergarten is straightforward (for example, letters, numbers, patterns), and would be simple for a teacher to master. As a result, the *way* in which kindergarten teachers teach that subject matter may be the most important determinant of their effectiveness. In secondary school, on the other hand, subject matter is harder to master, and teachers will likely have to specialize (see Cook and Mansfield 2014). An effective teacher will need to know the subject matter *and* how to teach it well. As a result, teacher quality is more likely to be portable across subjects in kindergarten and the early years of elementary school than in secondary school.

B. Teacher characteristics, behaviors, and child learning

The second part of the analysis relates differences across classrooms in learning outcomes to the characteristics and behaviors of teachers. For this purpose, we run regressions of the following form:

$$(2) \qquad Y_{ics}^{k} = a_s^k + \boldsymbol{X}_{ics}\beta_1^k + \ +\bar{\boldsymbol{X}}_{\boldsymbol{cs}}\beta_2^k + \boldsymbol{C}_{\boldsymbol{cs}}\Phi_1^k + \boldsymbol{B}_{\boldsymbol{cs}}\Phi_2^k + \varepsilon_{ics}^k \qquad k = 1,2,3$$

where a_s^k is a school indicator, so all coefficients are estimated using only within-school variation. Comparing equations (1) and (2), we replace the classroom indicator, δ_{cs}^k, with the school indicator a_s^k, classroom averages of the child characteristics $\bar{\boldsymbol{X}}_{\boldsymbol{cs}}$, and observable teacher characteristics, $\boldsymbol{C}_{\boldsymbol{cs}}$, and behaviors, $\boldsymbol{B}_{\boldsymbol{cs}}$. Standard errors throughout are clustered at the school level.

One point that requires elaboration is that as-good-as-random assignment was used to assign *teachers* to students, not teacher *characteristics* or *behaviors*. It is therefore important to be cautious in interpreting the parameters Φ_1^k and Φ_2^k because the *measured* characteristics and behaviors of teachers may also be correlated with other *unmeasured* teacher attributes which could themselves affect student learning.

Another concern is that some teacher characteristics and attributes could suffer from reverse causality. The measures of teacher IQ, personality, attention and inhibitory control were collected after the end of the 2012 school year, and it is conceivable that responses to the questions on the Big Five, or performance on the IQ and attention and inhibitory control tests were affected by the students that a teacher had taught recently. In Online Appendix E we present an intermediate set of results where we separate teacher attributes into two groups: those measured before the start of the school year, and those measured after. The results are very similar to the ones we report in the paper.

Reverse causality may also bias the coefficients on a teacher's contemporaneous CLASS score. For example, teachers who by chance were assigned a group of particularly attentive and well-behaved students may be given higher CLASS scores, but these students could also be better learners, even after accounting for their observed characteristics. For this reason, we use a teacher's lagged, rather than current, CLASS score in most of our estimates. In practice, this means that we

restrict the sample to teachers who taught kindergarten in the study sample of schools for the entire 2011 and 2012 school years, and their students.[27]

Results from the regressions of child learning outcomes on teacher characteristics and behaviors are reported in Table III. Columns with odd numbers refer to regressions that only include one teacher characteristic at a time, while those with even numbers include all at once. The table shows that a teacher's lagged CLASS score is associated with better learning outcomes—children assigned to teachers with a one-standard deviation higher CLASS score have between 0.05 and 0.07 standard deviations higher end-of-year test scores. Children with inexperienced teachers have test scores that are 0.17 standard deviations lower. None of the other teacher characteristics, including her tenure status, IQ, the five dimensions of the Big Five, inhibitory control and attention, and the education of her parents are consistently associated with student learning.[28] These results make clear how difficult it is to predict teacher effectiveness, even with a much richer set of teacher characteristics than is generally available.

Many of the teacher characteristics and behaviors are measured with error. This is perhaps particularly apparent for the CLASS, as we are approximating the quality of the interactions between teachers and children with the interactions observed on a single day. To explore how measurement error in the CLASS may affect our estimates, and the conclusions one might draw for policy design, we run a number of supplemental regressions. (These regressions do not include teacher attributes other than the CLASS.)

In calculating a teacher's CLASS score, and following established CLASS protocols, coders scored four 20-minute segments of a teacher in her class on a single day of school. (The CLASS score for that teacher is the average across the four segments.) A less expensive way of collecting

[27] When a teacher leaves a school (in the middle of either school year, or between school years) we lose all students in her class. Moreover, because all of our regressions include school fixed effects, when one teacher leaves a school with only two kindergarten classes, we lose all students in that school.

[28] In the United States, certification status is only weakly correlated with teacher effectiveness (Dobbie and Fryer 2013; Kane et al. 2008); in India, there is no difference in the effectiveness of contract teachers and civil service teachers (Muralidharan and Sundararaman 2011); in Kenya, children randomly assigned to civil service teachers generally have *lower* end-of-year test scores than those assigned to contract teachers (Duflo et al. 2011). Teachers in Ecuador are selected for tenure on a school-by-school basis (as *vacantes*, or slots, are given to schools by the Ministry of Education). All candidates for tenure have to take the *Concurso de Méritos y Oposiciones*, a test which has a written component and a demonstration class. Teaching experience also factors into a teacher's score. The applicant for a slot in a given school who has the highest score on the *Concurso* is awarded the tenured position. There are two possible reasons why, in spite of this rigorous selection process, tenure is not associated with better learning outcomes in our sample. First, the *Concurso* is relatively recent, and was first applied in 2008. Before that, selecting teachers for tenure was an ad hoc process. Only 22 of the 269 teachers that are the basis for most of the calculation in this paper received tenure in 2008 or later. Second, it is not clear that the written questions on the test actually distinguish better from worse teachers. School administrators, peer teachers, and parents receive little concrete guidance on how to score the demonstration class.

data on teachers might have considered only one, rather than four, segments. A comparison of the coefficient in column (1) (in which the CLASS is calculated on the basis of four segments) and column (2) (in which the CLASS is calculated on the basis of only the first segment) in Table IV shows that the coefficient on the CLASS is smaller (0.06, rather than 0.08) when only one segment is used to assess teaching practices. This is consistent with a reduction in measurement error as the amount of time that a teacher is observed increases. It also leaves open the possibility that if a teacher were observed on multiple days, perhaps over the course of a school year, measurement error would be reduced further.

To limit the potential for reverse causality, we generally use a teacher's lagged (rather than contemporaneous) CLASS score. However, using the lagged score may introduce other complications if there is drift in teacher quality (as suggested by Chetty et al. 2014a; Goldhaber and Hansen 2013). A comparison of the coefficient in column (1), 0.08 (using the lagged CLASS), and column (3), 0.06 (using the contemporaneous CLASS), suggests that reverse causality and drift in a teacher's CLASS are not major sources of concern in our sample, since we obtain similar estimates regardless of whether we use the current or lagged teacher score. It is also possible to average a teacher's CLASS score over the two years. Consistent with a reduction in measurement error, estimates that use this average are somewhat larger (a point estimate of 0.09, with a standard error of 0.02) than those that use the data from one year only.

Finally, we report estimates in which the lagged CLASS score is used as an instrument for the current CLASS score. The correlation between the two CLASS measurements is positive and highly significant in our data, so there is a strong first stage.[29] In addition, it is hard to imagine why the CLASS score in $t-1$ would be correlated with the regression error term in the equation that uses data from year t.[30] The IV regression in column (5) of the table suggests that, once the CLASS is purged of measurement error, a one-standard deviation increase in the CLASS is associated with a 0.18 standard deviation increase in learning outcomes.

[29] The first-stage coefficient is 0.42, with a standard error of 0.09.

[30] First, teacher videos are allocated randomly to coders, so it is very unlikely that the same teacher would be coded by the same pair of coders in both years. Therefore, coding error should be independent over time. Second, temporary shocks to teachers that affect their teaching could be correlated over time over a very short horizon (for example, if a teacher has the flu one day she may also have it for the subsequent two or three days), but not over a year. Therefore, measurement error coming from the observation of a teacher on a single day should be uncorrelated across years. Third, because of as-good-as-random assignment in year t, the characteristics of students taught by a teacher in year t should be uncorrelated with the characteristics of students taught by that same teacher in year $t-1$, at least after we account for school fixed effects.

It is worth considering which of these results is most useful for policy purposes. One policy objective could be to use the CLASS to identify effective teachers (for example, for promotion). In this case, the OLS results are most relevant. Moreover, there is no obvious disadvantage to using contemporaneous, rather than lagged, CLASS scores, and observing a teacher multiple times is likely to produce more accurate measures of her effectiveness than observing her once only.

Another policy objective might be to estimate how much learning outcomes could increase if teacher behaviors, as measured by the CLASS, were to improve (say, through an in-service training program targeting those behaviors), or to understand the sources of differences in teacher effectiveness. In that case, the IV results might be more informative (subject to the caveat that the coefficients in both the OLS and IV regressions may not have a causal interpretation).

C. Transmission mechanisms

The coefficients in regressions of test scores on teacher quality are not structural primitives (as stressed by Chetty et al. 2014a; Todd and Wolpin 2003). They include not only the direct effect of teachers on student learning but also any behavioral responses, in particular by parents, to offset or augment the effect of being assigned to a better or worse teacher.

To understand how better teachers produce more learning, we begin by analyzing whether parents can distinguish better from worse teachers. To this effect, we regress the score that parents gave to teachers (on the 1-5 scale) on different measures of teacher quality: estimated value-added, lagged CLASS, and an indicator for inexperienced teachers.[31] (Each regression includes only one teacher characteristic at a time.) We note that the scores parents give to teachers are generally high. Only 0.1 percent of teachers were scored as "very bad" (score = 1), 0.3 percent as "bad", 2 percent as "average", 37.5 percent as "good", and 58 percent as "very good" (score =5). Given this distribution of the scores, we report the results from various parametrizations, including simple OLS, ordered probit, and an OLS specification in which the parent perception variable is coded as a dummy that takes on the value of one for scores of five, and zero for scores between one and four.[32]

[31] We recalculate value-added for each observation used in this table, removing from the calculation the score of the corresponding student. Therefore, for each parent, we compute value added based only on the test scores of their child's peers, which we call "End-of-grade peer test score", to account for the possibility that a parent's valuation of the teacher may be greatly influenced by their own child's learning.

[32] Mothers with more schooling generally give higher scores to teachers. In an OLS regression of the score given to a teacher on school fixed effects and a dummy variable for mothers who have completed at least secondary school, the coefficient on high-education mothers is 0.06 (with a standard error of 0.02). However, we find no evidence that mothers with more schooling are better able to tell good from bad teachers. For example, in a regression of the score given to a teacher on the dummy for high-education mothers, the CLASS, and the interaction between the two, the

The results, presented in Table V, indicate that parents generally give higher scores to better teachers. Looking at the second column of the table, parents are 15 percentage points more likely to classify a teacher who produces one-standard deviation higher test scores as "very good" rather than "good" or less. They also give significantly higher scores to teachers with better CLASS scores and more experienced teachers.

We then directly consider parenting behaviors. In the household survey, we asked about a large number of inputs into child development and learning (including the availability of books, pencils, and toys of various kinds) and behaviors (including whether parents read to, sang to, or played with, their kindergarten children). We begin by analyzing whether these investments are correlated with child learning, controlling for baseline child TVIP. Most, but not all, of the inputs and behaviors predict student learning—see Online Appendix F. We keep all those that are positively associated with child learning outcomes, and create two indices—one for the availability of inputs (like books for children) another for parent behaviors (like parents reading to children).[33]

Finally, we run regressions of the index of inputs (or behaviors) on measures of classroom or teacher quality, one at a time. None of the coefficients in Table VI is significant. For example, a one standard deviation increase in learning outcomes in a classroom is associated with a very small decline in the inputs index (a coefficient of -0.003, from a sample average of 0.50), and a similarly small decline in the behaviors index (a coefficient of -0.001, from a sample average of 0.49). We conclude that, although we cannot rule out subtle adjustments on other margins we do not measure, the evidence is most consistent with parents making at most modest changes in response to random differences in teacher quality.

D. Out-of-sample predictions

The fact that we have data on learning outcomes for two cohorts of children assigned to the same teachers in the same schools allows us to carry out some simple, policy-relevant, out-of-sample

---

coefficient on high-education mothers is 0.06 (with a standard error of 0.02), that on the CLASS is 0.05 (with a standard error of 0.02), and the coefficient on the interaction term is 0.01 (with a standard error of 0.01).

[33] These indices are the simple mean of all of the inputs (or behaviors). They range from zero (for a household that has none of the inputs, or a parent who carries out none of the activities) to one (for household where all inputs are available, or a parent who carries out all of the activities). Both indices predict student learning, by construction: A one-unit change in the inputs index is associated with 0.25 standard deviations more learning (with a standard deviation of 0.03), while a one-unit change in the behavior index is associated with 0.09 standard deviations more learning (with a standard deviation of 0.03). There is no exogenous source of variation in inputs or behaviors, so these coefficients do not have a clean causal interpretation. Nevertheless, they show that the association between parental investments and child learning outcomes is substantial.

predictions.[34] By using the four language and math tests that were applied to the two cohorts of children, we can test how well data in year *t* allow us to predict teacher performance in year *t+1*. In particular, we focus on whether teacher characteristics and behaviors predict current value added once lagged value added has been controlled for.

These results are reported in Table VII. The first column in each panel reports the results of regressions of the average residualized test score in 2013 on the corresponding average in 2012. For the four tests as a whole, the correlation between value added in the two years is 0.36 —far from unity, but higher than what has generally been found in the US literature: Koedel et al. (2015) report that the correlation for studies that include school fixed effects ranges from 0.18 to 0.33. We next add the lagged CLASS to this regression. These results, in the second column in each panel, show that the CLASS is not a significant predictor of student learning gains, conditional on lagged value added. Next, we add the full set of teacher characteristics and behaviors. These results, in the third column in each panel, show that these teacher attributes are not significant, individually or jointly.[35]

In sum, Table VII shows that, once teacher value added is known, other teacher characteristics and behaviors are not significant predictors of out-of-sample learning outcomes in language and math among kindergarten students in Ecuador. Our results are broadly consistent with those from the MET study: Mihaly et al. (2013) calculate the optimal weights that should be given to different measures of teacher quality in predicting the stable component of value added, and conclude that between 65 percent (for elementary school language) and 85 percent (for elementary school math) of the weight would be given to value added. On the other hand, Rockoff and Speroni (2010) find that subjective evaluations of teachers by mentors predict learning gains in 3rd through 8th grades in math and English in New York City, even after they condition on previous year value added.[36]

---

[34] We thank Raj Chetty for suggesting this exercise.

[35] We also tested whether the score that parents give to teachers on the 1-5 scale predicts value added in year *t+1*, conditional on value added in year *t*. The coefficient on the parental score is very small, is as likely to be positive as negative, and is never significant at conventional levels.

[36] Kane et al. (2011) show that a measure of teaching behaviors and practices used to evaluate teachers in the Cincinnati public school system predicts learning gains. Other studies, including Rockoff et al. (2011) and Dobbie (2011), show that composite indices of teacher characteristics can predict value added. However, these studies do not control for lagged value added. In other words, they focus on whether teacher characteristics can predict value added, rather than on whether these characteristics predict out-of-sample learning outcomes, conditional on value added. Like others (Rockoff et al. 2011; Dobbie 2013), we used principal components to aggregate all of the teacher characteristics and behaviors. We retained all components that have an eigenvalue larger than 1 (five components) and regressed learning outcomes in *t+1* on value added in *t* and the five components. The second component is a significant predictor of student learning outcomes, but the five components are not jointly significant.

4. Conclusions

In this paper we have shown that there are substantial differences in the amount of learning that takes place in language, math, and executive function across kindergarten classrooms in Ecuador, a middle-income country. These differences are associated with differences in teacher behaviors and practices, as measured by a well-known classroom observation tool, the Classroom Assessment Scoring System (CLASS). We also show that parents can generally tell better from worse teachers, but do not meaningfully alter their investments in children in response to random shocks to teacher quality.

We conclude by considering some possible policy implications. First, our results show that value added is a useful summary measure of teacher quality in Ecuador. However, to date, no country in Latin America regularly calculates the value added of teachers.[37] Indeed, in virtually all countries in the region, decisions about tenure, in-service training, promotion, pay, and early retirement are taken with no regard for (and in most cases no knowledge about) a teacher's effectiveness. Value added is no silver bullet, but knowing which teachers produce more or less learning among equivalent students would be an important step to designing policies to improve learning outcomes.

Nevertheless, using value added as the *only* measure of teaching effectiveness has limitations. Collecting high-quality test score data may not always be feasible, especially for very young children.[38] It is also possible that measures of teacher quality other than value added, perhaps including teaching practices, could predict child outcomes other than test scores, test scores in subsequent years, or outcomes in adulthood (as discussed in Algan et al. 2013; Chetty et al. 2011; Jackson et al. 2014). Finally, value added is silent about *what* makes some teachers more effective than others.

Our results on the CLASS suggest that teachers vary a great deal in how they interact with children in the classroom. Teachers with higher CLASS scores produce more learning. These results may not have a causal interpretation because teacher behaviors may be correlated with other

[37] In Chile (Mizala and Urquiola 2013), and in the state of Pernambuco, Brazil (Ferraz and Bruns 2011) schools are ranked by a composite measure which includes test scores, and those with high scores on this composite measure receive additional resources. See also Behrman et al. (2015) for a pilot in Mexico.

[38] The data we collected for this study are likely to be higher quality than those that could realistically be collected regularly for the universe of kindergarten children in Ecuador (or elsewhere). The fact that we used twelve different tests, and that children were tested individually (rather than having paper-and-pencil tests handed out to a class as a whole) probably reduced measurement error substantially because young children are more likely to be distracted, or fail to understand instructions, in group testing situations. If measurement error in value added were substantial, it is possible that other measures of teacher quality would significantly predict out-of-sample child learning, even after accounting for a teacher's value added with a different group of children.

(unmeasured) teacher attributes. The evidence from some, but by no means all, pilots of innovative mentoring and in-service training programs in the United States suggests that the CLASS is malleable, and that programs that improve teacher behaviors in the classroom can, under some circumstances, improve child learning (Bierman et al. 2008, and Downer et al. 2013). There would be high returns to policy experimentation and careful evaluation of this kind in Latin America, and in other middle-income countries.

References

Ackerman, P., and E. Heggestad. "Intelligence, Personality, and Interests: Evidence for Overlapping Traits." *Psychological Bulletin* 121(2): 219-45.

Algan, Y., P. Cahuc, and A. Shleifer. 2013. "Teaching Practices and Social Capital." *American Economic Journal: Applied Economics* 5(3): 189-210.

Almlund, M., A.L. Duckworth, J. Heckman, and T. Kautz. 2011. "Personality Psychology and Economics." In E, Hanushek, S. Machin, and L. Woessman, Eds., *Handbook of the Economics of Education*, Vol. 4 (pp. 1-182). Oxford: Elsevier.

Anderson, P.J. 2002. "Assessment and Development of Executive Functioning (EF) in Childhood." *Child Neuropsychology* 8(2): 71-82.

Angrist, J., E. Bettinger, E. Bloom, E. King, and M. Kremer. 2002. "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment." *American Economic Review* 92(5): 1535-58.

Austin, E., I. Deary, M. Whiteman, F.G.R. Fowkes, N. Pedersen, P. Rabbitt, N. Bent, and L. McInnes. 2002. "Relationships between Ability and Personality: Does Intelligence Contribute Positively to Personal and Social Adjustment?" *Personality and Individual Differences* 32: 1391-1411.

Behrman, J., S. Parker, P. Todd, and K. Wolpin. 2015. Aligning Learning Incentives of Students and Teachers: Results from a Social Experiment in Mexican High Schools. *Journal of Political Economy* 123(2): 325-64.

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

Bierman, K., C. Domitrovich. R. Nix, S. Gest, J. Welsh, M. Greenberg, C. Blair, K. Nelson, and S. Gill. 2008. "Promoting Academic and Social-Emotional School Readiness: The Head Start REDI Program." *Child Development* 79(6): 1802-17.

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

Case, A., and C. Paxson. 2008. "Stature and Status: Height, Ability, and Labor Market Outcomes." *Journal of Political Economy* 116(3): 499-532.

Chetty, R., J. Friedman, N. Hilger, E. Saez, D. Schanzenbach, and D. Yagan. 2011. "How Does your Kindergarten Classroom Affect your Earnings? Evidence from Project STAR." *Quarterly Journal of Economics* 126(4): 1593-1660.

Chetty, R., J. Friedman, and J. Rockoff. 2014a. "Measuring the Impact of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review* 104(9): 2593-2632.

Chetty, R., J. Friedman, and J. Rockoff. 2014b. "Measuring the Impact of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review* 104(9): 2633-2679.

Clifford, D., D. Bryant, M. Burchinal, O. Barbarin, D. Early, C. Howes, R. Pianta, and P. Winton. "National Center for Early Development and Learning Multistate Study of Pre-Kindergarten, 2001-2003." Data available at http://doi.org/10.3886/ICPSR04283.v3

Cook, Jason, and R. Mansfield. 2014. "Task-Specific Experience and Task-Specific Talent: Decomposing the Productivity of High School Teachers." Working Paper, Cornell University ILR School.

Costa, P., and R. McCrae. 1992. Revised NEO Personality Inventory (NEO-PI-R) and NEO Five Factor Inventory (NEO-FFI). Professional Manual. Odessa, FL: *Psychological Assessment Resources*.

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

Dobbie, W. 2011. Teacher Characteristics and Student Achievement: Evidence from Teach for America." Unpublished manuscript, Harvard University.

Dobbie, W., and R. Fryer. 2013. "Getting Beneath the Veil of Effective Schools: Evidence From New York City." *American Economic Journal: Applied Economics* 5(4): 28-60.

Downer, J., R. Pianta, M. Burchinal, S. Field, B. Hamre, J. LoSale-Crouch, C. Howes, K. LaParo and C. Scott-Little. 2013. "Coaching and Coursework Focused on Teacher-Child Interactions during Language-Literacy Instruction: Effects on Teacher Outcomes and Children's Classroom Engagement." Unpublished manuscript, University of Virginia.

Duflo, E., P. Dupas, and M. Kremer. 2011. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." *American Economic Review* 101(5): 1739-74.

Duflo, E., R. Hanna, and S. Ryan. 2012. "Incentives Work: Getting Teachers to Come to School." *American Economic Review* 102(4): 1241-78.

Duncan, G., C. Dowsett, A. Claessens, K. Magnuson, A. Huston, P. Klebanov et al. 2007. "School Readiness and Later Achievement." *Developmental Psychology* 43: 1428-46.

Dunn, L., D. Lugo, E. Padilla, and L. Dunn. 1986. *Test de Vocabulario en Imágenes Peabody*. Circle Pines, MN: American Guidance Service.

Espy, K.A. 2004. "Using Developmental, Cognitive, and Neuroscience Approaches to Understand Executive Functions in Preschool Children." *Developmental Neuropsychology* 26(1): 379-84.

Ferraz, C., and B. Bruns. 2011. "Paying Teachers to Perform: The Impact of Bonus Pay in Pernambuco, Brazil." Unpublished, available at http://files.eric.ed.gov/fulltext/ED530173.pdf.

Goldhaber, D., and M. Hansen. 2013. "Is it Just a Bad Class? Assessing the Long-Term Stability of Teacher Performance." *Economica* 80(319): 589-612.

Goldsmith, H.H., A.H. Buss, R. Plomin, M.K. Rothbart., A. Thomas, S. Chess, R.A. Hinde, R.B. McCall. 1987. "Roundtable: What is Temperament? Four Approaches." *Child Development* 58(2): 505-29.

Gormley, W., D. Phillips, K. Newmark, K. Welti., and S. Adelstein. 2011. "Social-Emotional Effects of Early Childhood Education Programs in Tulsa." *Child Development* 82(6): 2095-2109.

Grossman, P., S. Loeb, J. Cohen, K. Hammerness, J. Wyckoff, D. Boyd, and H. Lankford. 2010. "Measure for Measure: The Relationship between Measures of Instructional Practice in Middle School English Language Arts and Teachers' Value Added Scores." NBER Working Paper 16015.

Heckman, J. 2013. *Giving Kids a Fair Chance (A Strategy that Works)*. Cambridge, MA: MIT Press.

Heckman, J., and T. Kautz. 2012. "Hard Evidence on Soft Skills." *Journal of Labor Economics* 19(4): 451-64.

Hamre, B., B. Hatfield, R. Pianta and F. Jamil. 2014. "Evidence for General and Domain-Specific Elements of Teacher-Child Interactions: Associations with Preschool Children's Development." *Child Development* 85(3): 1257-1274.

Hanushek, E., and S. Rivkin. 2012. "The Distribution of Teacher Quality and Implications for Policy." *Annual Review of Economics* 4: 131-57.

Howes, C., M. Burchinal, R. Pianta, D. Bryant, D. Early, R. Clifford and O. Barbarin. 2008. "Ready to Learn? Children's Pre-Academic Achievement in Pre-Kindergarten Programs." *Early Childhood Research Quarterly* 23: 27-50.

Jacob, B., and L. Legfren. 2008. "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education." *Journal of Labor Economics* 26(1): 101-36.

Jackson, K., J. Rockoff, and D. Staiger. 2014. "Teacher Effects and Teacher-Related Policies." *Annual Review of Economics* 6: 801-25.

Jacob, B., L. Lefgren, and D. Sims. 2010. "The Persistence of Teacher-Induced Learning Gains." *Journal of Human Resources* 45(4): 915-43.

Jennings, J., and T. DiPrete. 2010. "Teacher Effects on Social/Behavioral Skills in Early Elementary School." *Sociology of Education* 83(2): 135-59.

Jensen, A.R., and W.D. Rowher. 1966. "The Stroop Color-Word Test: A Review." *Acta Psychologica* 25(1): 36-93.

Kane, T., and D. Staiger. 2002. "Volatility in School Test Scores: Implications for Test-Based Accountability Systems." *Brookings Papers on Education Policy*: 235-283.

Kane, T., and D. Staiger, 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." NBER Working Paper 14607.

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.

Koedel, G., K. Mihaly, J. Rockoff. 2015. "Value-Added Modeling: A Review." Forthcoming, *Economics of Education Review*.

Kremer, M., C. Brannen, and R. Glennerster. 2013. "The Challenge of Education and Learning in the Developing World." *Science* 340: 297-300.

MacLeod, C.M. "Half a Century of Research on the Stroop Effect: An Integrative Review." *Psychological Bulletin* 109(2): 163-203.

McCaffrey, D., T. Sass, J. Lockwood, and K. Mihaly. 2009. "The Intertemporal Variability of Teacher Effect Estimates." *Education Finance and Policy* 4: 572-606.

Mihaly, K., D. McCaffrey, D. Staiger, and J.R. Lockwood. 2013. "A Composite Measure of Effective Teaching." Unpublished manuscript, available at http://metproject.org/downloads/MET_Composite_Estimator_of_Effective_Teaching_Research_Paper.pdf.

Mizala, A., and M. Urquiola. 2013. "School Markets: The Impact of Information Approximating Schools' Effectiveness." *Journal of Development Economics* 103: 313-35.

Moffitt, T., L. Arseneault, D. Belsky, N. Dickson, R. Hancox, H. Harrington, R. Houts, R. Poulton, B. Roberts, S. Ross, M. Sears, M, Thomson, and A. Caspi. 2011. "A Gradient of Childhood Self-Control Predicts Health, Wealth, and Public Safety." *Proceedings of the National Academy of Sciences* 108(7): 2693-98.

Moutafi, J., A. Furnham, and L. Paltiel. 2004. "Why is Conscientiousness Negatively Correlated with Intelligence?" *Personality and Individual Differences* 37: 1013-22.

Muñoz-Sandoval, A., R. Woodcock, K. McGrew, and N. Mather. 2005. *Batería III Woodcock-Muñoz.* Itasca, IL: Riverside Publishing.

Muralidharan, K., and V. Sundararaman. 2011. "Teacher Performance Pay: Experimental Evidence from India." *Journal of Political Economy* 119(1): 39-77.

Murnane, R., and A. Gaminian. 2014. "Improving Educational Outcomes in Developing Countries: Lessons from Rigorous Evaluations." NBER Working Paper 20284.

Nelson, C, and M. Sheridan. 2011. "Lessons from Neuroscience Research for Understanding Causal Links between Family and Neighborhood Characteristics and Educational Outcomes." In G. Duncan and R. Murnane, Eds., *Whither Opportunity: Rising Inequality, Schools, and Children Life Chances* (pp. 27-46). New York: Russell Sage Foundation.

Obradovic, J., X. Portillo, and T. Boyce. 2012. "Executive Functioning and Developmental Neuroscience." In R. Pianta, Ed., *Handbook of Early Childhood Education* (pp. 324-51). New York and London: The Guilford Press.

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

Pop-Eleches, C., and M. Urquiola. 2013. "Going to a Better School: Effects and Behavioral Responses." *American Economic Review* 103(4): 1289-1324.

Powell, D., and K. Diamond. 2012. "Promoting Early Literacy and Language Development." In R. Pianta, Ed., *Handbook of Early Childhood Education* (pp. 194-216). New York and London: The Guilford Press.

Reubens, A.2009. Early Grade Mathematics Assessment (EGMA): A Conceptual Framework Based on Mathematics Skills Development in Children. Available at https://www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&id=193

Rivkin, S., E. Hanushek, and J. Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica* 73(2): 417-58.

Rockoff, J., B. Jacob, T. Kane, and D. Staiger. 2011. "Can You Recognize an Effective Teacher when You Recruit One?" *Education Finance and Policy* 6(1): 43-74.

Rockoff, J., and C. Speroni. 2010. "Subjective and Objective Evaluations of Teacher Effectiveness." *American Economic Review, Papers and Proceedings* 100(2): 261-266

Rothstein, J. 2010. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics* 125(1): 175-214.

RTI International. 2009. Early Grade Reading Assessment Toolkit, 2009. Available at https://www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&id=149

Schady, N. 2012. "El Desarrollo Infantil Temprano en América Latina y el Caribe: Acceso, Resultados y Evidencia Longitudinal de Ecuador." In *Educación para la Transformación*:, eds. Marcelo Cabrol and Miguel Székely. Washington, DC: Inter-American Development Bank.

Schady, N., J. Behrman, C. Araujo, R. Azuero, R. Bernal, D. Bravo, F. Lopez-Boo, K. Macours, D. Marshall, C. Paxson, and R. Vakis. 2015. "Wealth Gradients in Early Childhood Cognitive Development in Five Latin American Countries." *Journal of Human Resources* 50(2): 446-63.

Séguin, J., and P. Zelazo. 2005. "Executive Function in Early Physical Aggression." In R. Tremblay, V. Hartup, and J. Archer, Eds., *Developmental Origins of Aggression* (pp. 307-29). New York: The Guilford Press.

Senn, T.E., K.A. Espy, and P.M. Kaufmann. 2004. "Using Path Analysis to Understand Executive Function Organization in Preschool Children." *Developmental Neuropsychology* 26(1): 445-64.

Shonkoff, J. and D. Phillips. 2000. *From Neurons to Neighborhoods. The Science of Early Childhood Development*. Washington, D.C.: National Academy Press.

Siegler, R. 2009. "Improving Preschoolers' Number Sense Using Information-Processing Theory." In O. Barbarin and B. Wasik, Eds., *Handbook of Childhood Development and Education* (pp. 429-54). New York and London: The Guilford Press.

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

Stroop, R. 1935. "Studies of Inference in Serial Verbal Reactions." J*ournal of Experimental Psychology* 18(6): 643-662.

Todd, P., and K. Wolpin. 2003. "On the Specification and Estimation pf the Production Function for Cognitive Achievement." *The Economic Journal* 113(February): F3-F33.

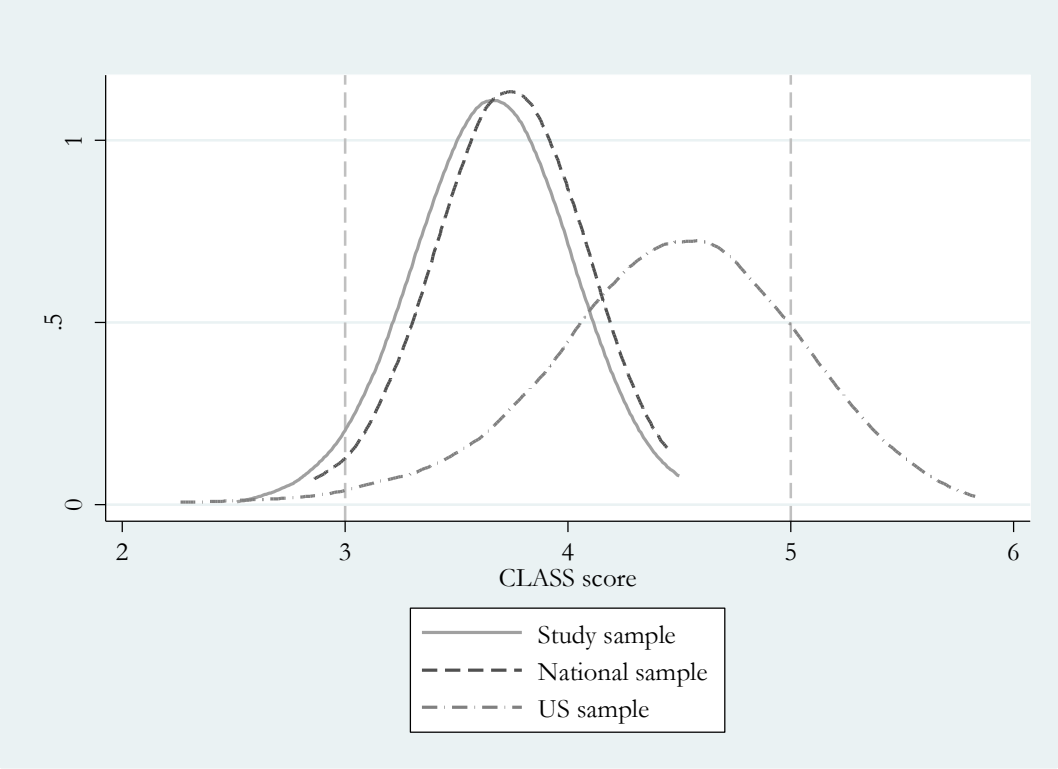
Van der Linden, D., J. te Nijenhuis, and A. Bakker. 2010. "The General Factor of Personality: A Meta-Analysis of Big Five Intercorrelations and a Criterion-Related Validity Study." *Journal of Research in Personality* 44: 315-27.

Wasik, B., and B. Newman. 2009. "Teaching and Learning to Read." In O. Barbarin and B. Wasik, Eds., *Handbook of Childhood Development and Education* (pp. 303-27). New York and London: The Guilford Press.

Wechsler, D. 1939. *The Measurement of Adult Intelligence*. Baltimore(MD): Williams and Witkins.

Weiland, C., and H. Yoshikawa. 2013. "Impacts of a Prekindergarten Program on Children's Mathematics, Language, Literacy, Executive Function, and Emotional Skills." *Child Development* 84(6): 2112-30.

Figure I: Distribution of CLASS scores, study, national and US samples



Notes: The figure graphs univariate densities of the CLASS score of kindergarten teachers in 2012 in the study and national samples in Ecuador, and in a nationally-representative sample of kindergarten classrooms in the United States (Clifford et al. 2003). The study sample is the sample used in most calculations in our paper, while the national sample is a nationally representative sample used for comparison. The CLASS is scored on a 1-7 scale; scores of 1-2 indicate poor quality, scores of 3-5 indicate intermediate levels of quality, and scores of 6-7 indicate high quality. Calculations are based on an Epanechnikov kernel with optimal bandwidth.

Table I: Summary Statistics

	Study sample			National sample		
	Mean	S.D.	Obs.	Mean	S.D.	Obs.
Children						
Age (months)	59.35	5.24	15,302	56.66	5.66	1,032
Proportion female	0.49	0.50	15,434	0.49	0.50	1,034
TVIP	82.82	15.87	13,850	-	-	-
Mother's age (years)	30.23	6.57	13,662	30.64	6.63	983
Father's age (years)	34.56	7.91	10,644	34.24	7.75	782
Mother's years of schooling	8.78	3.81	13,652	8.36	3.80	983
Father's years of schooling	8.49	3.84	10,618	8.28	3.76	780
Proportion who attended preschool	0.61	0.49	14,395	0.70	0.46	1,019
Teachers						
Age (years)	42.23	9.58	448	43.14	10.31	218
Proportion female	0.99	0.10	450	0.98	0.13	218
Proportion with 3 years of experience or less	0.06	0.24	450	0.05	0.21	218
Proportion tenured	0.64	0.48	450	0.86	0.35	218
Number of students in classroom	34.22	8.00	451	31.73	7.83	218

Notes: This table reports means and standard deviations of the characteristics of children entering kindergarten in 2012 and of their teachers, measured at the beginning of the school year. The study sample is the sample used in the rest of the paper, while the national sample is a nationally representative sample used for comparison. The TVIP is the *Test de Vocabulario en Imágenes 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 II: Within-school standard deviations of classroom and teacher effects

	(1)	(2)	(3)	(4)	(5)
	Classroom effects				**Teacher effects**
	2012 cohort (12 tests)		2012 cohort (4 tests)	2013 cohort (4 tests)	
Sample Restriction	Whole sample	Classes with the same teacher throughout the year	Classes for which teachers are the same in both cohorts of children		2012 and 2013 cohorts (4 tests)
Language	0.11	0.11	0.10	0.10	0.09
Math	0.11	0.12	0.11	0.11	0.09
Executive function	0.07	0.07	--	--	--
Total	0.11	0.11	0.12	0.10	0.10
Students	13,565	9,962	5,904	6,023	11,927
Teachers	451	334	196	196	196
Schools	204	150	87	87	87

Notes: The table reports the within-school standard deviations of classroom effects (columns 1 through 4), and teacher effects (column 5), adjusted for sampling error. In columns 1 and 2 classroom effects are calculated based on all twelve tests administered in the study for the 2012 cohort, which are used to calculate three indices of student performance in language, math, and executive function (which are then standardized to have mean zero and variance equal to one). Columns 3, 4, and 5 use only the 4 tests administered to the 2013 cohort two construct two indices, for language and math. Column 1 uses the whole sample of students, column 2 takes only classrooms for which the teacher remained the same throughout the year, and columns 3, 4 and 5 use only classrooms in which a teacher was observed both in the 2012 and 2013 cohorts. In columns 1 and 2, the control variables used in the model include all baseline child and household characteristics. In the remaining columns, controls include only child age, gender, and the baseline TVIP score, since household data are not available for the 2013 cohort.

Table III: Teacher characteristics and behaviors, and child learning outcomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Language		**Math**		**Executive function**		**Total**	
	Bivariate	Multivariate	Bivariate	Multivariate	Bivariate	Multivariate	Bivariate	Multivariate
Lagged CLASS	0.06*	0.04	0.08**	0.05	0.06**	0.04	0.07**	0.05*
	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.03)	(0.02)
Teacher has 3 years of experience or less	-0.15*	-0.12	-0.16	-0.11	-0.12	-0.09	-0.17*	-0.13*
	(0.07)	(0.06)	(0.09)	(0.07)	(0.06)	(0.07)	(0.07)	(0.06)
Tenured teacher	0.05	-0.00	0.08	0.05	0.01	-0.02	0.06	0.01
	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
IQ	0.04*	0.04	0.04*	0.02	0.03	0.03	0.04*	0.03
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Neuroticism	0.00	0.01	0.00	0.02	0.02	0.03	0.01	0.02
	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Extraversion	0.03	0.02	0.03	0.02	0.03*	0.02	0.04*	0.02
	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)	(0.02)
Openness	0.01	0.01	0.02	0.03	0.01	0.02	0.02	0.02
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Agreeableness	-0.00	-0.01	-0.00	-0.01	-0.02	-0.03	-0.01	-0.02
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Conscientiousness	-0.02	-0.02	-0.03	-0.04	-0.02	-0.02	-0.03	-0.03
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Inhibitory control & attention	0.02	0.00	0.03	0.01	0.03*	0.02	0.03	0.01
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Parents' education	0.01	0.00	0.01	0.01	0.00	0.00	0.01	0.00
(Average years)	(0.01)	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)
Students	7,978	7,978	7,978	7,978	7,978	7,978	7,978	7,978
Classrooms	269	269	269	269	269	269	269	269
Schools	125	125	125	125	125	125	125	125
R-squared		0.47		0.36		0.30		0.46
F-test (p-value)		0.11		0.00		0.03		0.01

Notes: The table reports estimates from regressions of test scores on teacher characteristics and behaviors. All regressions are limited to children in schools in which at least two teachers taught kindergarten in both the 2011 and 2012 school years. All regressions include baseline student and household characteristics, their classroom averages, and school fixed effects. Specifications in columns with odd numbers (bivariate) correspond to regressions in which each teacher characteristics or behaviors are included one at a time, while those with even numbers (multivariate) correspond to regressions in which all are included simultaneously. Standard errors (in parentheses) clustered at the school level. * significant at 5%, ** at 1%.

Table IV: Teacher behaviors and child learning outcomes, OLS and IV estimates

	(1)	(2)	(3)	(4)	(5)
	OLS				**IV**
	Lagged CLASS		**Current CLASS**	**Average CLASS**	**Current CLASS**
	4 video segment average	Single video segment	4 video segment average	4 video segment average	4 video segment average
Language	0.06*	0.05*	0.05**	0.07**	0.14*
	(0.03)	(0.02)	(0.02)	(0.02)	(0.06)
Math	0.08**	0.06**	0.07**	0.09**	0.18**
	(0.03)	(0.02)	(0.02)	(0.03)	(0.07)
Executive function	0.06**	0.03	0.03	0.05**	0.13*
	(0.02)	(0.02)	(0.02)	(0.02)	(0.05)
Total	0.08**	0.06**	0.06**	0.09**	0.18**
	(0.03)	(0.02)	(0.02)	(0.02)	(0.06)
Students	7,978	7,978	7,978	7,978	7,978
Classrooms	269	269	269	269	269
Schools	125	125	125	125	125

Notes: The table reports estimates from regressions of test scores on CLASS scores. All regressions are limited to children in schools in which at least two teachers taught kindergarten in both the 2011 and 2012 school years. All regressions include baseline student and household characteristics, their classroom averages, and school fixed effects. Lagged CLASS was measured in the 2011 school year, while current CLASS was measured in the 2012 school year. In column 2 CLASS is constructed from one video segment, while in the remaining columns it is based on four video segments from one day of teaching. In the instrumental variables regression, current CLASS is instrumented with lagged CLASS. The coefficient on the instrument in the first-stage regression is 0.42 (0.093). Standard errors (in parentheses) clustered at the school level. * significant at 5%, ** at 1%.

Table V: Parent perceptions and observable measures of teacher quality

	(1) **OLS**	(2) **OLS dummy**	(3) **Ordered probit**
End-of-grade peer test score	0.18**	0.15**	0.41**
	(0.07)	(0.05)	(0.15)
Lagged CLASS	0.05**	0.04**	0.12**
	(0.02)	(0.01)	(0.03)
Teacher has 3 years of experience or less	-0.16**	-0.13**	-0.32**
	(0.04)	(0.04)	(0.1)
Students	7,873	7,873	7,873
Classrooms	269	269	269
Schools	125	125	125

Notes: The table reports estimates of regressions of parent perception of teacher quality (1-5 scale, 1 = very bad, 5 = very good) on measures of teacher quality. All regressions limited to children in schools in which at least two teachers taught kindergarten in both the 2011 and 2012 school years. Each cell in the table corresponds to a separate regression. Column 1 corresponds to OLS regression estimates. In column 2 the dependent variable is a dummy variable which takes the value of one if a parent's perception of the teacher is very good, and zero otherwise. Column 3 presents estimates from an ordered probit. Mean parents' perception of teacher quality is 4.5, the median is 5.0; 0.1 percent of teachers are classified as very bad, 0.3 percent as bad, 4.2 percent as average, 37.5 percent as good, and 57.9 percent as very good. End-of-grade peer test score is a measure of teacher value added recalculated for each observation used in this table, removing from the calculation the score of the corresponding student. All regressions include school fixed effects and baseline child and household characteristics and their averages. Standard errors (in brackets) clustered at school level. * significant at 5%, ** at 1%.

Table VI: Parental responses to differences in teacher quality

	(1) **Home stimulation index**	(2) **Home inputs index**
End-of-grade peer test score	-0.00	-0.00
	(0.03)	(0.02)
Lagged CLASS	0.01	0.00
	(0.01)	(0.01)
Teacher has 3 years of experience or less	-0.06	-0.01
	(0.03)	(0.02)
Students	7,933	7,952
Classrooms	269	269
Schools	125	125

Notes: The table reports estimates of regressions of parental responses to differences in teacher quality. The dependent variables in each column are indices of home stimulation and home inputs, standardized to have mean zero and unit standard deviation. All regressions are limited to children in schools in which at least two teachers taught kindergarten in both the 2011 and 2012 school years. Each cell in the table corresponds to a separate regression. All regressions include school fixed effects and baseline child and household characteristics and their averages. Standard errors (in brackets) clustered at school level. * significant at 5%, ** at 1%.

Table VII: Teacher value added in 2013, and teacher characteristics and behaviors in 2012

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
		Language			**Math**			**Total**	
Value added 2012	0.38*	0.37*	0.31	0.29*	0.27*	0.20	0.36**	0.34*	0.27
	(0.15)	(0.16)	(0.17)	(0.12)	(0.12)	(0.13)	(0.12)	(0.13)	(0.15)
CLASS 2012		0.01	0.01		0.02	0.02		0.01	0.02
		(0.02)	(0.02)		(0.03)	(0.03)		(0.02)	(0.03)
Teacher has 3 years of experience or less			0.00			0.00			0.00
			(0.00)			(0.00)			(0.00)
Tenured teacher			0.03			0.07			0.05
			(0.06)			(0.06)			(0.06)
IQ			0.02			-0.02			0.00
			(0.03)			(0.03)			(0.03)
Neuroticism			0.04			0.04			0.04
			(0.04)			(0.04)			(0.04)
Extraversion			-0.01			-0.01			-0.01
			(0.02)			(0.02)			(0.02)
Openness			0.03			0.01			0.03
			(0.03)			(0.03)			(0.03)
Agreeableness			0.01			0.02			0.02
			(0.03)			(0.03)			(0.03)
Conscientiousness			-0.02			-0.08*			-0.06*
			(0.03)			(0.03)			(0.03)
Inhibitory control & attention			-0.01			-0.00			-0.00
			(0.04)			(0.03)			(0.04)
Parents' education (Average years)			-0.00			0.00			-0.00
			(0.01)			(0.01)			(0.01)
Students	6,023	6,023	6,023	6,023	6,023	6,023	6,023	6,023	6,023
Classrooms	196	196	196	196	196	196	196	196	196
Schools	87	87	87	87	87	87	87	87	87
R-squared	0.14	0.14	0.20	0.09	0.10	0.27	0.14	0.15	0.25
F-test (p-value)		0.66	0.91		0.60	0.17		0.60	0.63

Notes: The table reports multivariate regressions of value added for each teacher in the 2013 school year, on value added in 2012, and teacher characteristics. Columns 1-3 consider valued added in language, columns 4-6 consider value added in math, and columns 7-9 consider value added on an index of math and language. Standard errors (in parentheses) clustered at the school level. * significant at 5%, ** at 1%.

Online Appendix

Teacher Quality and Learning Outcomes in Kindergarten

M. Caridad Araujo
Pedro Carneiro
Yyannú Cruz-Aguayo
Norbert Schady

Online Appendix A: End-of-year tests

This appendix provides additional information on the end-of-year tests we applied to the study and national samples in the 2012 cohort.

We applied four tests of language and early literacy: (1) the Test de Vocabulario en Imágenes Peabody (TVIP), the Spanish version of the much-used Peabody Picture Vocabulary Test (PPVT) (Dunn et al. 1986). In the TVIP, children are shown slides, each of which has four pictures, and are asked to identify the picture that corresponds to the object (for example, "boat") or action (for example, "to measure") named by the test administrator. The test is a measure of receptive vocabulary because children do not have to name the objects themselves and because children need not be able to read or write; (2) oral comprehension, in which the enumerator reads the child a very short passage and asks her simple questions about its content; (3) letter and word recognition; and (4) identification of the first sound of a letter or word. All of these tests other than the TVIP were taken from the Spanish-speaking version of the Woodcock-Johnson battery of tests of child development and achievement (Muñoz-Sandoval et al. 2005), and from an adapted version of the Early Grade Reading Assessment (RTI International 2009). Vocabulary and oral comprehension are part of children's early oral language abilities. Letter and word recognition is an early measure of alphabet awareness, and identification of the first sound of a letter or word is an early measure of phonological awareness. Both are measures of children's early code-related skills.

We applied four tests of math: (1) number recognition; (2) a number series test, in which children are asked to name the number that is missing from a string of numbers (for example, 2, 4, -, 8); (3) block rotation, in which children are asked to identify geometric figures that are the same, even though they have been rotated (some items are two-dimensional, others three-dimensional); and (4) applied math problems, in which the enumerator reads out a simple problem to a child (for example, "Imagine you have four cars and are given three more. How many cars would you have in total?"). All of the math tests were taken from the Spanish-speaking version of the Woodcock-Johnson battery (Muñoz-Sandoval et al. 2005), and from an adapted version of the Early Grade Math Assessment (Reubens 2009).

Finally, we applied four test of executive function. The first test is the Stroop Day-Night test (Stroop 1935). On this test, children are shown slides with either (1) a moon and stars in a dark sky, or (2) a sun in a bright sky, in quick succession. They are then asked to say "day" whenever the moon is shown, and "night" when the sun is shown. The test measures *response inhibition* because the child is asked to say the opposite of what would be her natural response (which has to be inhibited).

In the second test the child is read a string of numbers and is asked to repeat them in order or backwards. Repeating numbers backwards is a more difficult task because it involves remembering the numbers, manipulating them, and then repeating the altered sequence. The test also varies in difficulty in that earlier questions have short number strings (for example, 3-7-9), while later items have longer strings (for example, 6-4-0-5-6). This is a widely-used test of *working memory*.

The third test is the dimensional change card sort (Zelazo et al. 2013). In this test, children are shown two wooden boxes. The first box has a card with a blue truck on the inside cover, and the second box has a card with a red apple. In the first part of the test (the "shapes" part), children are asked to put cards in the box that corresponds to the right shape (trucks with trucks, apples with apples), no matter their color. In the second part (the "color" part of the test), children are asked to sort cards according to color (red cards with red cards, and blue cards with blue cards), no matter what picture is on the card. Empirically, at about five years of age (but not earlier), most children can switch between two incompatible sets of rules such as these (Zelazo et al. 1996). This is a test of *cognitive flexibility* because it requires children to ignore an earlier rule and replace it with a new one.

In the final test we used, children are shown a picture and asked to point at a series of objects in the correct order (for example, "point at the pig, then at the umbrella, and then at the chicken that is closest to the barn"). This test measures *attention*, although it clearly has a working memory element as well.

All of the tests were piloted in Ecuador before they were applied to the study and national samples of kindergarten children we use in our paper. We made minor changes to the tests, as needed (for example, clarifying the instructions, taking out questions that were clearly too difficult for the children in our study). During data collection, the order in which tests were carried out was assigned randomly—each enumerator had a different order of tests, and, within this order, the starting point varied from student to student.[1] Tests were applied to children individually (as opposed to a class as a whole), and in total took between 30 and 40 minutes per child.[2]

Appendix Figure A1 provides histograms of test scores in the study sample for each individual test. The figure shows there is a clear heaping of mass at various points of the distribution of the executive function tests. Most children had little difficulty with the Stroop Day-Night test of response inhibition—more than half (55 percent) answered all questions correctly, suggesting that this test may have been too easy for children in our study. On the test of cognitive flexibility, 87 percent of children answered all questions in the first part of the test correctly; of these children, 56 percent then answered all questions on the second part correctly (indicating that they understood that the sorting rule had changed), while 25 percent did not answer any of these questions on the second part correctly (indicating that they were unable to switch from one sorting rule to the other,

[1] For example, for tests A through L, enumerator 1 might have the order C, A, F, E, L, B, I, D, H, G, K, J, and would apply the tests in the following order to child 1 (F, E, … A), and in the following order to child 2 (G, K… H), while enumerator 2 might have the order L, A, K, G, B, J, D, C, F, E, I, H, and would apply the tests in the following order to child 1 (K, G… A), and in the following order to child 2 (C, F… D), with the order of tests and the starting test for each enumerator determined by a random number generator.

[2] There are a number of important advantages to applying tests to children of this age individually. Any test that would have required the child to read and follow any instructions would have been impossible to apply, as children at this age (in Ecuador and elsewhere) cannot read. None of the EF tests can be applied in a group setting. Furthermore, young children are much more likely to be distracted in a group testing situation, and this could have introduced substantial measurement error to the tests. Of course, applying tests individually, rather than to a class as a whole, is also considerably more expensive.

despite repeated prompting by the enumerator). Children also had difficulty with the longer strings in the working memory test. Very few children, only 7 percent, could correctly repeat any of the five-number strings, and even fewer, less than 2 percent, could repeat any of the strings backwards (even the easiest, three-number strings). Appendix Figure A1 also shows a heaping of mass at the top or bottom of the distribution of some of the individual math and language tests.

Appendix Figure A2 shows univariate density estimates of the distribution of the average language score, the average math score, the average executive function score, and the total test score for the study and national samples. The figure shows that taking the averages across tests makes the distributions appear much closer to normal.

We also note that test performance generally appears to be very low. Most of the tests do not have scales that indicate how children at different ages "should" be performing. The TVIP is an exception. The TVIP has been standardized on samples of Mexican and Puerto Rican children to have an average score of 100 and a standard deviation of 15 at all ages (Dunn et al. 1986). We can use the tables provided by the test developers to convert the raw scores (number of correct responses) into externally-normed, age-specific scores. The mean TVIP score of children at baseline is 83, more than one standard deviation below the average in the reference population; children at the 10^{th} percentile have a TVIP score of 61, two-and-a-half standard deviations below. Consistent with what has been found in other data from Ecuador (Paxson and Schady 2007; Schady 2011; Schady et al. 2015), many of the children we study in this paper have deep deficits in receptive language development. For the other tests, we do not have data to make international comparisons. Nevertheless, performance on these tests also seems to be very low. The median child correctly answered five of the 14 questions on the oral comprehension test, and four of the 17 questions on the applied math problems test. Outcomes in other tests seem to be even lower—on average, at the end of the year, children recognized only two letters and four numbers, and most children could not identify the first sound of any letter or word (even though this is something that, according to the kindergarten curriculum, teachers in Ecuador should be covering).

Finally, we note there are substantial socioeconomic gradients in test scores. Appendix Figure A3 graphs the average score (separately for math, language, executive function, and for the average across the twelve tests) of children in the study sample by month of age (horizontal axis), separately for children of mothers of three education categories: elementary school dropouts (11 percent of the sample); elementary school graduates (58 percent of the sample, half of these women began, but did not graduate from, high school); and high school graduates (32 percent of the sample, about a third of these women have some education beyond high school). All lines slope upwards from left to right, indicating that older children have higher test scores than younger children: On average, each month of age is associated with 0.03 standard deviations higher test scores. Children of mothers with more education have higher scores. On average, across all twelve tests, children of mothers who are high school graduates have scores that are 0.43 standard deviations higher than children of mothers who only completed elementary schooling; children of elementary school graduates, in turn, have scores that are 0.29 standard deviations higher than children of elementary school dropouts.

The education gradients are steepest in language, and least steep (but still considerable) in executive function.

References

Dunn, L., D. Lugo, E. Padilla, and L. Dunn. 1986. *Test de Vocabulario en Imágenes Peabody*. Circle Pines, MN: American Guidance Service.

Muñoz-Sandoval, A., R. Woodcock, K. McGrew, and N. Mather. 2005. *Batería III Woodcock-Muñoz*. Itasca, IL: Riverside Publishing.

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

Reubens, A.2009. Early Grade Mathematics Assessment (EGMA): A Conceptual Framework Based on Mathematics Skills Development in Children. Available at https://www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&id=193

RTI International. 2009. Early Grade Reading Assessment Toolkit, 2009. Available at https://www.Zelaeddataglobal.org/documents/index.cfm?fuseaction=pubDetail&id=149

Schady, N. 2011. "Parental Education, Vocabulary, and Cognitive Development in Early Childhood: Longitudinal Evidence from Ecuador." *American Journal of Public Health* 101(12): 2299-307.

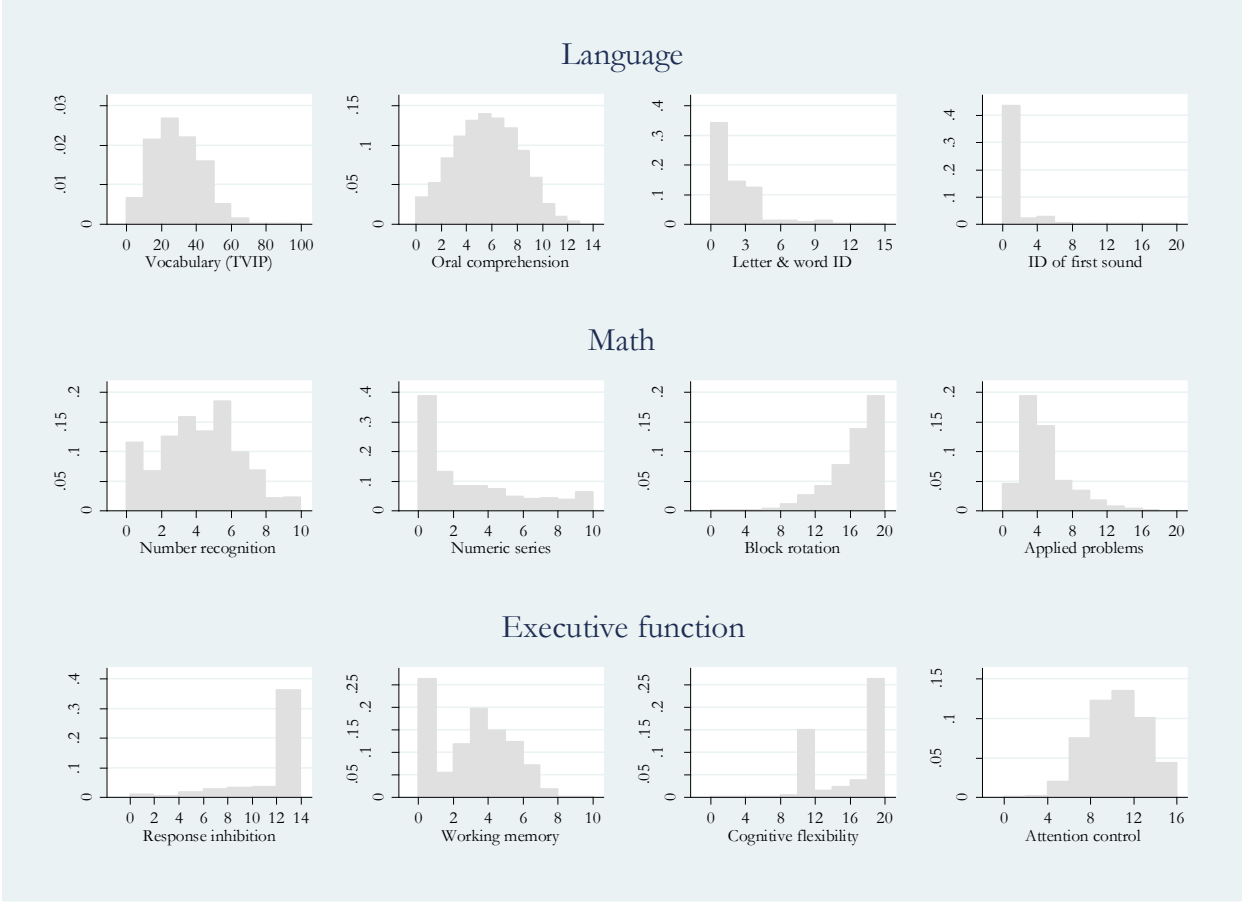
Schady, N., J. Behrman, C. Araujo, R. Azuero, R. Bernal, D. Bravo, F. Lopez-Boo, K. Macours, D. Marshall, C. Paxson, and R. Vakis. 2015. "Wealth Gradients in Early Childhood Cognitive Development in Five Latin American Countries." *Journal of Human Resources* 50(2): 446-63.

Stroop, R. 1935. "Studies of Inference in Serial Verbal Reactions." J*ournal of Experimental Psychology* 18(6): 643-662.

Zelazo, P., J. Anderson, J. Richler, K. Wallner-Allen, J. Beaumont, and S. Weintraub. 2013. "NIH Toolbox Cognition Battery: Measuring Executive Function and Attention." *Monographs of the Society for Research in Child Development* 78(4): 16-33.

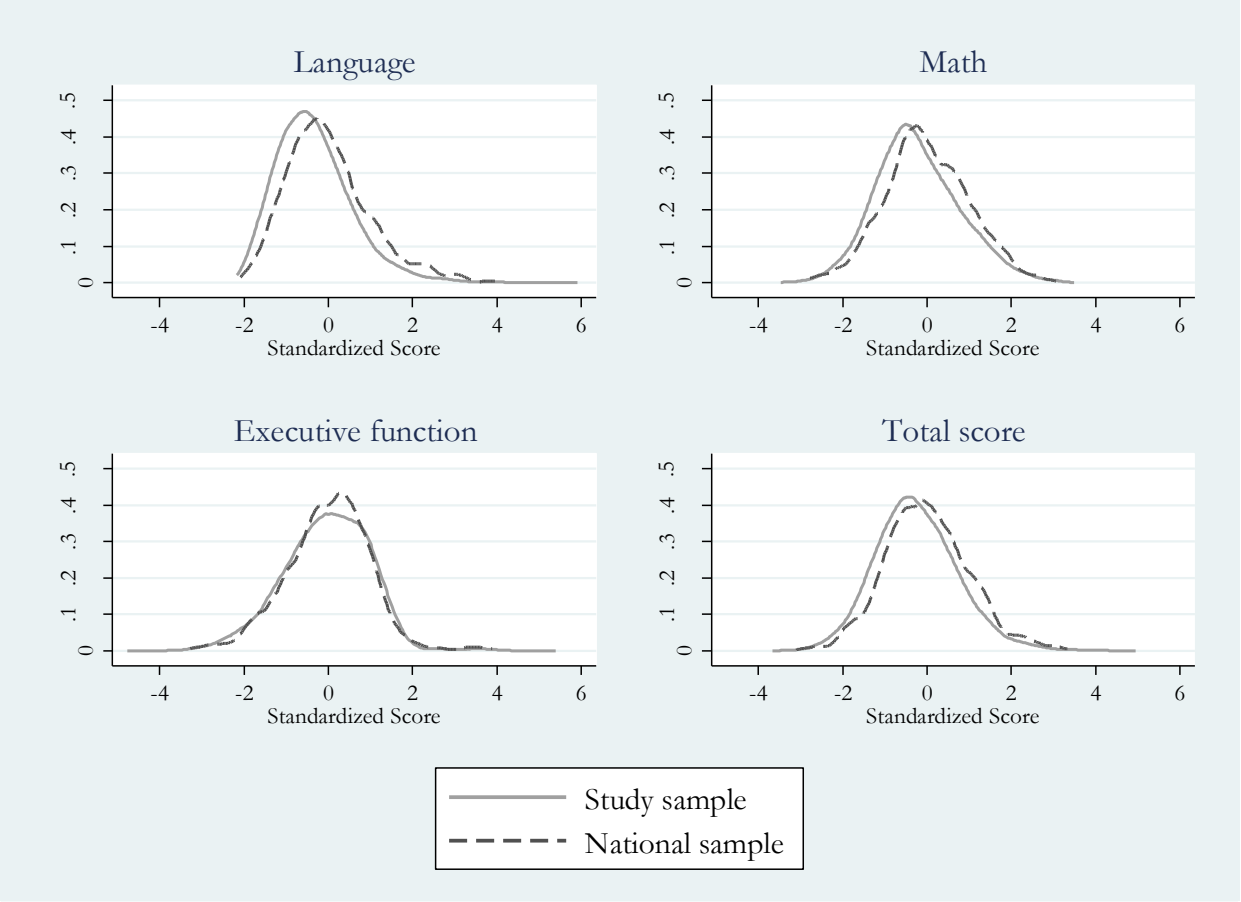
Zelazo, P., D. Frye, and T. Rapus. 1996. "An Age-Related Disassociation between Knowing Rules and Using Them." *Child Development* 11(1): 37-63.

Appendix Figure A1: Test score histograms



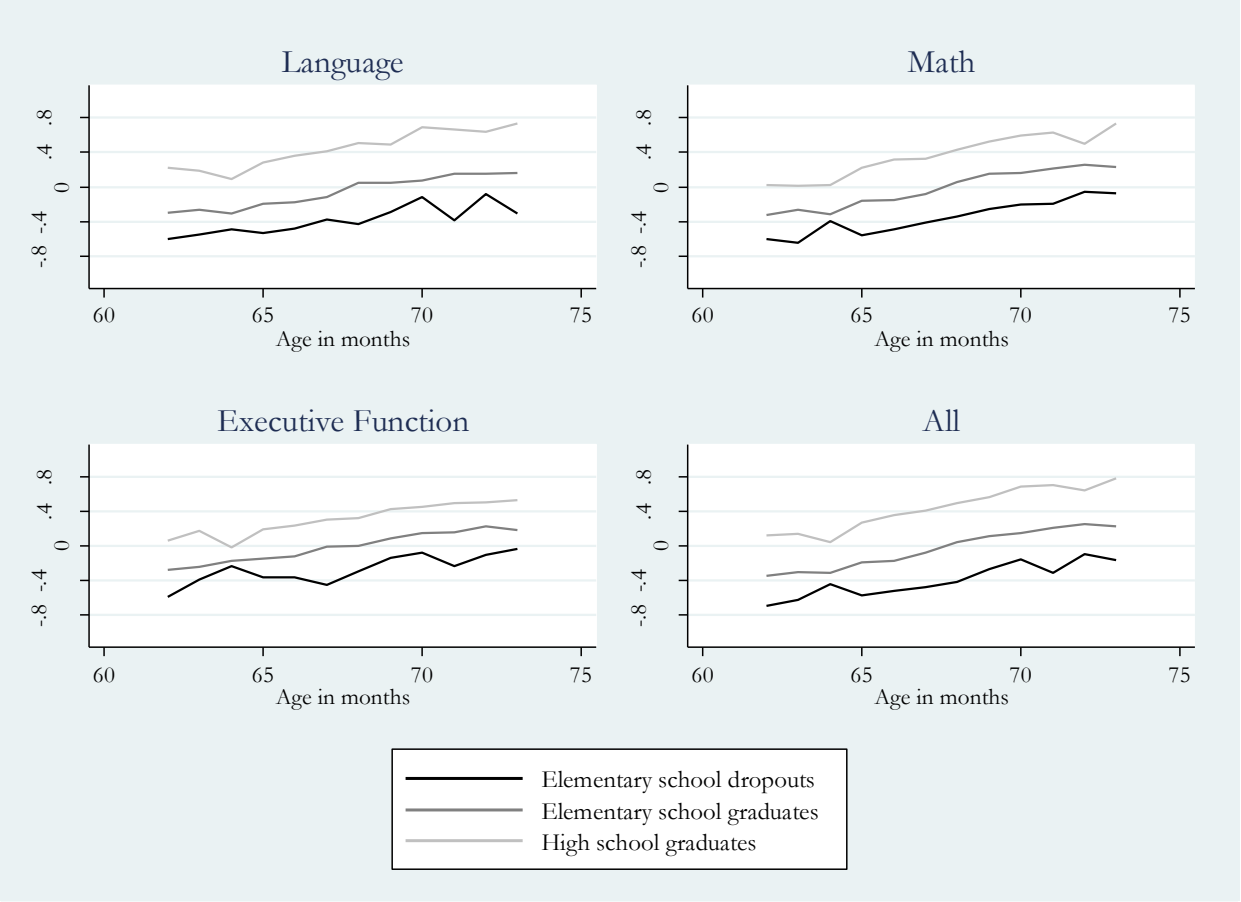
Notes: Figure presents histograms of the raw test scores (number of correct responses) by test.

Appendix Figure A2. Univariate densities of test score aggregates



Notes: Figure presents univariate densities of test score aggregates in standard deviation units (mean zero, unit standard deviation).

Appendix Figure A3: Gradients in standardized test scores by age and mother's education



Notes: Figure presents average test scores, in standard deviation units (mean zero, unit standard deviation) of test scores, by age in months of the child and years of education of the child's mother.

Appendix B: Application of the CLASS in Ecuador

The main measure of teacher behaviors (or interactions) we use in this paper is 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).

To give a better sense of the behaviors that are measured by the CLASS, we cite at length from Berlinski and Schady (pp. 136-37, 2015), which draws heavily on Cruz-Aguayo et al. (2015):

> *"Emotional support*: In classrooms with high levels of emotional support, teachers and students have positive relationships and enjoy spending time together. Teachers are aware of, and responsive to, children's needs, and prioritize interactions that place an emphasis on students' interests, motivations, and points of view. In classrooms with low levels of emotional support, teachers and students appear emotionally distant from one another, and there are instances of frustration in interactions. Teachers seldom attend to children's need for additional support and, overall, the classroom follows a teacher's agenda with few opportunities for student input. Many studies from the United States have found associations between the teachers' provision of emotionally supportive interactions in the classroom and students' social-emotional development.[3]
>
> *Classroom organization*. In highly organized classrooms, teachers are proactive in managing behavior by setting clear expectations; classroom routines allow for students to get the most out of their time engaged in meaningful activities; and teachers actively promote students'

[3] Perry et al. (2007) found that across 14 first-grade classrooms, higher emotional support at the beginning of the year was associated with more positive peer behavior and less problem behaviors as the year progressed. Similarly, in an examination of 36 first grade classrooms serving 178 6- and 7-year-old students, emotionally supportive classrooms demonstrated decreased peer aggression over the course of the year (Merritt et al. 2012). Emotional climate appears to influence academic outcomes, as well. In a sample of 1,364 third grade students, the classroom's emotional support was related to a child's reading and mathematics scores at the end of the year (Rudasill et al. 2010).

engagement in those activities. In less organized classrooms, teachers might spend much of their time reacting to behavior problems; classroom routines are not evident; students spend time wandering or not engaged in activities; and teachers do little to change this. When teachers manage behavior and attention proactively, students spend more time on-task and are better able to regulate their attention (Rimm-Kaufman et al. 2009). Students in better organized and managed classrooms also show larger increases in cognitive and academic development (Downer et al. 2010).[4]

Instructional support. In classrooms with high levels of instructional support, a teacher promotes higher order thinking and provides quality feedback to extend students' learning. At the low end, rote and fact-based activities might be common, and students receive little to no feedback about their work beyond whether or not it is correct. In these classrooms, teachers do most of the talking or the room is quiet. The quality of instructional support provided in a classroom is most consistently linked with higher gains in academic outcomes, such as test scores.[5]"

In practice, in our application of the CLASS, scores across different dimensions are highly correlated with each other, as can be seen in Appendix Table B2. 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 kindergarten teachers in the study and national samples of schools in the 2012 school year. In addition, in the study sample of schools, we filmed all kindergarten teachers in the previous (2011) school year. Teachers were filmed for a full school day

[4] For example, data from 172 first graders across 36 classrooms in a rural area of the United States demonstrated that classroom organization was significantly predictive of literacy gains (Ponitz et al. 2009).

[5] References include Burchinal et al. (2008, 2010); Hamre and Pianta (2005); and Mashburn et al. (2008). For example, examining 1,129 low-income students enrolled in 671 pre-kindergarten classrooms in the United States, Burchinal et al. (2010) found a significant association between instructional support and academic skills; classrooms demonstrating higher instructional support had students who scored higher on measures of language, reading, and math than those enrolled in classrooms with low-quality instructional support. Similarly, Mashburn et al. (2008) used data from the United States and found that the instructional support of a classroom was related to all five academic outcomes measured (receptive language, expressive language, letter naming, rhyming, and applied math problems).

(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 4,900 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 compared average CLASS scores for the first and fourth segments. We found 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 in the *current* or *previous* year. 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, the inter-coder reliability ratio was high, 0.92, 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, we carried out some additional calculations, summarized in Appendix Table B3. First, we calculated the reliability ratio of the scores across segments within a day for a given teacher. The cross-segment reliability ratio between the 1st and 4th segment is 0.77. Second, we make use of the fact that a subsample of teachers was filmed for two or three days in 2011. (On average, 2 days elapsed between the first and second day of filming, and 4 days between the first and third day of filming.) For these teachers, we can therefore calculate the cross-day reliability ratio, comparing the scores they received in days 1 and 2 (for 105 teachers), and between days 1 and 3 (for 45 teachers). The cross-day reliability ratio is 0.83 for days 1 and 2, and 0.86 for days 1 and 3. We note that this pattern—large increases in measured relative to "true" variability with more segments per day and more days of filming, 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

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

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.

Burchinal, M., C. Howes, R. Pianta, D. Bryant, D. Early, R. Clifford, and O. Barbarin. 2008. "Predicting Child Outcomes at the End of Kindergarten from the Quality of Pre-Kindergarten Teacher-Child Interactions and Instruction." *Applied Developmental Science* 12(3): 140-53.

Burchinal, M., N. Vandergrift, R. Pianta, and A. Mashburn. 2010. "Threshold Analysis of Association between Child Care Quality and Child Outcomes for Low-Income Children in Pre-Kindergarten Programs." *Early Childhood Research Quarterly* 25(2): 166-76.

Cruz-Aguayo, Y., J. LoCasale-Crouch, S. Schodt, T. Guanziroli, M. Kraft-Sayre, C. Melo, S. Hasbrouck, B. Hamre, and R. Pianta. 2015. "Early Classroom Schooling Experiences in Latin America: Focusing on What Matters for Children's Learning and Development." Unpublished manuscript, Inter-American Development Bank.

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.

Hamre, B., and R. Pianta. 2005. "Can Instructional and Emotional Support in the First-Grade Classroom Make a Difference for Children at Risk of School Failure?" *Child Development* 76(5): 949-67.

Hamre, B., B. Hatfield, R. Pianta and F. Jamil. 2014. "Evidence for General and Domain-Specific Elements of Teacher-Child Interactions: Associations with Preschool Children's Development." *Child Development* 85(3): 1257-1274.

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., R. Pianta, B. Hamre, J. Downer, O. Barbarin, D. Bryant, M. Burchinal, D. Early, and C. Howes. 2008. "Measures of Classroom Quality in Prekindergarten and Children's Development of Academic, Language, and Social Skills." *Child Development* 79(3): 732-49.

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.

Merritt, E.G., S.B. Wanless, S.E. Rimm-Kaufman, C. Cameron, and J.L. Peugh. 2012. "The Contribution of Teachers' Emotional Support to Children's Social Behaviors and Self-Regulatory Skills in First Grade." *School Psychology Review* 41(2): 141-59.

Perry, K.E., K.M. Donohue, and R.S. Weinstein. 2007. "Teaching Practices and the Promotion of Achievement and Adjustment in First Grade." *Journal of School Psychology* 45(3): 269-92.

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.

Ponitz, C.C., S.E. Rimm-Kaufman, L.L. Brock, and L. Nathanson. 2009. "Early Adjustment, Gender Differences, and Classroom Organizational Climate in First Grade." *Elementary School Journal* 110(2): 142-62.

Rimm-Kaufman, S., R. Pianta, and M. Cox. 2000. "Teachers' Judgments of Problems in the Transition to Kindergarten." *Early Childhood Research Quarterly* 15(2) 147-66.

Rudasil, K., K. Gallagher, and J. White. 2010. "Temperamental Attention and Activity, Classroom Emotional Support, and Academic Achievement in Third Grade." *Journal of School Psychology* 48(2): 113-34.

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).

Appendix Table B2: Pairwise correlation of CLASS dimensions

		Emotional Support					Classroom Organization				Instructional Support				
		Positive Climate	Negative Climate	Teacher Sensitivity	Regard for Students Perspectives	Emotional Support Total	Behavior Management	Productivity	Instructional Learning Formats	Classroom Organization Total	Concept Development	Quality of Feedback	Language Modeling	Instructional Support Total	Total CLASS score
Emotional Support	Positive Climate	1													
	Negative Climate	0.45	1												
	Teacher Sensitivity	0.89	0.44	1											
	Regard for Students Perspectives	0.54	0.36	0.51	1										
	Emotional Support Total	0.95	0.62	0.94	0.65	1									
Classroom Organization	Behavior Management	0.56	0.56	0.55	0.27	0.61	1								
	Productivity	0.52	0.28	0.54	0.23	0.53	0.68	1							
	Instructional Learning Formats	0.75	0.36	0.73	0.36	0.74	0.70	0.74	1						
	Classroom Organization Total	0.68	0.45	0.68	0.32	0.70	0.89	0.90	0.91	1					
Instructional Support	Concept Development	0.40	0.12	0.40	0.30	0.40	0.27	0.37	0.44	0.40	1				
	Quality of Feedback	0.53	0.12	0.54	0.35	0.52	0.32	0.41	0.53	0.47	0.63	1			
	Language Modeling	0.39	0.10	0.40	0.24	0.38	0.22	0.34	0.43	0.37	0.77	0.67	1		
	Instructional Support Total	0.50	0.13	0.50	0.33	0.48	0.30	0.42	0.53	0.46	0.91	0.86	0.91	1	
	Total CLASS score	0.88	0.54	0.87	0.52	0.90	0.79	0.78	0.90	0.91	0.56	0.64	0.53	0.65	1

Notes: The table reports the pairwise correlation coefficient of CLASS dimensions in 2012 for 451 teachers. All the correlations in the table are significant at the 99 percent confidence level, except for three correlations that are significant at the 90 percent confidence level.

Appendix Table B3: Sources of measurement error in the CLASS

	N	Correlation	Reliability Ratio
Inter-coder (2012)	451	0.86	0.92
Inter-segment (1st and 4th segments) (2012)	451	0.44	0.77
First and second day (2011)	105	0.72	0.83
First and third day (2011)	45	0.76	0.86

Notes: The table reports the correlation and reliability ratio of the CLASS, for different coders (first row), different segments within a day for the same teacher (second row), and different days for the same teacher (third and fourth rows).

Appendix C: Checks on assignment of children and teachers and other robustness tests

We argue that identification in our paper is based on as-good-as-random assignment of children to teachers. We would therefore expect that any differences in student composition across classrooms would be orthogonal with teacher quality. Appendix Table C1 shows that this is indeed the case. The table reports the results from regressions of a given child or household characteristic on three key teacher characteristics (the lagged CLASS score, a dummy variable for inexperienced teachers, and whether the teacher had a tenured position). There are 18 coefficients in the table, each from a different regression. All of the coefficients are small in magnitude, and none is significant at conventional levels.

We also ran regressions of the within-school alphabetical order of children on child characteristics (age, gender, the baseline TVIP score, mother's education, the wealth aggregate, and whether she attended preschool), and school fixed effects (seven different regressions). For six of these seven variables, the coefficient is not significant. The only exception is in the regression of order on child age, which implies that the child who is ordered last in a school by her name is, on average, 0.34 months younger than the child who is ordered first (with a standard error of 0.013).[6]

No shows, attritors, and late enrollments: There were 951 children (6 percent of the total) whose parents signed them up for kindergarten in our study sample of schools, but who never showed up in practice, or dropped out in the first days of the school year; we refer to these children as "no shows". There were 580 children for whom we have the baseline TVIP but do not have end-of-year test scores (4 percent of all children for whom we have the baseline TVIP); we refer to these children as "attritors". Finally, there were 661 children for whom we have end-of-year test scores, but no baseline TVIP (5 percent of children with end-of-year test score data). We assume that these children enrolled in school at some point during the course of the school year (rather than at the beginning); we refer to them as "late enrollments".

To see how no shows, attritors, and late enrollments could affect our results, we carry out a number of calculations. First, we run regressions of the dummy variable for each of these three groups on observable teacher characteristics, including these characteristics one by one (with the exception of the Big Five characteristics, which we include in a single regression). There are 33 regression coefficients, and only two are significant at the 5 percent level or higher: tenured teachers in the no show regression (a point estimate of 0.012, with a standard error of 0.006) and the extraversion Big Five trait in the late enrollments regression (a point estimate of -0.007, with a standard error of 0.003). Moreover, when we include all of the teacher characteristics together, and carry out an F-test for joint significance, the p-value on the F-test is 0.34 for no shows, 0.25 for attritors, and 0.26 for late enrollments. We conclude from these regressions that there is no evidence that the decision to be a no show, attritor, or late enrollment is affected by the observable characteristics of teachers.

[6] By our as-good-as-random assignment rule, twins would always be assigned to different classrooms; this need not occur with true random assignment.

We next regress the three dummies (for no shows, attritors, and late enrollments, respectively) on school and teacher fixed effects (leaving out one teacher per school), and test the joint significance of the teacher fixed effects. The p-values of the F-tests of joint significance are 0.11 for no shows, 0.27 for attritors, and <0.01 for new enrollments. We conclude that there is no evidence that parents' decisions not to enroll a child in school in the 2012 school year (no shows), or to withdraw their children from school during that year (attritors) were affected by the observed or unobserved characteristics of teachers. There is a stronger prima facie case that parents' decision to enroll their children in school after the beginning of the school year (late enrollment) could be affected by teacher quality.

Because late enrollments were not signed up for school and were not in class in the first weeks of the school year, we could not assign them to classes using the as-good-as-random assignment rule. Moreover, we do not have the baseline TVIP for these children. To be conservative, we do not include new enrollments in the main results in the paper. In practice, however, including these children in the calculations (and giving them the average baseline TVIP value in their school) has a negligible effect on the estimates we report.

A final estimation challenge arises because teachers move schools (for example, if they apply for, and are awarded, a tenured position in a different school), or temporarily stop teaching (for example, during maternity leave). In practice, 53 teachers left the schools in our study sample *between* the 2011 and 2012 school years, and 65 moved *within* the 2012 year. We note that these values are lower than those found in studies with US data. For example, 32 percent of teachers in Domitrovich et al. (2008) attrited from the sample between one year and the next; in Downer et al. (2013) 22 percent of teachers left the sample between random assignment and the end of the first year of the study, and a further 23 percent attrited between the first and second years.

To test whether the teachers who moved schools are different from those who did not we generated dummy variables for teachers who moved between the 2011 and 2012 years, and within the 2012 year, respectively. We then ran regressions of these dummies on our measures of the 2011 CLASS, years of experience, and the dummy variable for tenured teachers, with and without school fixed effects (12 separate regressions). These regressions, reported in Appendix Table C2, show (unsurprisingly) that teachers who moved schools generally have less experience and are much less likely to be tenured than those who did not move.

A concern that is potentially important is whether teachers are more likely to move within the school year if, by chance, they were assigned a particularly difficult group of students. To test for this, we ran regressions of a dummy variable for teachers who left the school at some point during the 2012 school year on the variables for baseline TVIP, child age, gender, education of the mother, and household wealth, with and without school fixed effects (10 separate regressions). Appendix Table C3 shows that the coefficients on these variables are all very small and, in the regressions that include school fixed effects, none are anywhere near conventional levels of significance. In other words, we find no evidence that the observable characteristics of students affected a teacher's decision to move schools.

References

Domitrovich, C., S. Gest, S. Gill, K. Bierman, J. Welsh, and D. Jones. 2008. "Fostering High-Quality Teaching with an Enriched Curriculum and Professional Development Support: The Head Start REDI Program." *American Educational Research Journal* 46(2): 567-97.

Downer, J., R. Pianta, M. Burchinal, S. Field, B. Hamre, J. LoSale-Crouch, C. Howes, K. LaParo and C. Scott-Little. 2013. "Coaching and Coursework Focused on Teacher-Child Interactions during Language-Literacy Instruction: Effects on Teacher Outcomes and Children's Classroom Engagement." Unpublished manuscript, University of Virginia.

Appendix Table C1: Verifying as-good-as random assignment

		(1) Age (months)	(2) Female	(3) TVIP	(4) Wealth index	(5) Mother's years of schooling	(6) Attended preschool
	Mean	60.32	0.49	0.01	0.00	8.79	0.60
	Standard deviation	4.94	0.50	1.00	0.99	3.79	0.49
	N:	14,052	14,066	13,739	13,691	12,971	14,066
CLASS 2011		-0.04	0.00	-0.01	-0.02	0.10	0.00
		(0.08)	(0.01)	(0.02)	(0.02)	(0.06)	(0.01)
		11,237	11,250	10,965	10,950	10,376	10,924
Teacher has 3 years of experience or less		-0.02	0.00	-0.03	-0.03	-0.32	0.01
		(0.23)	(0.02)	(0.05)	(0.04)	(0.17)	(0.03)
		11,237	11,250	10,965	10,950	10,376	10,924
Tenured teacher		0.05	-0.01	0.02	0.04	0.12	-0.01
		(0.15)	(0.01)	(0.03)	(0.02)	(0.11)	(0.02)
		11,237	11,250	10,965	10,950	10,376	10,924

Note: All regressions include school fixed effects. Standard errors clustered at the school level. The wealth aggregate is the first principal component of all the housing characteristics and asset ownership variables collected in the household survey. It includes whether the household has piped water and (separately) sewerage in the home; three separate variables for the main material of the floor, walls, and roof of the house; and whether the household owns a television, computer, fridge or washing machine (four separate variables)* significant at 5%, ** at 1%.

Appendix Table C2: Characteristics of teachers who left sample

		(1)	(2)	(3)	(4)
		Teacher left school between the 2011 and 2012 school years		Teacher left school within the 2012 school year	
	Mean				
CLASS 2011	0.00	-0.3 (0.14)*	-0.12 (0.22)	-0.37 (0.15)*	-0.20 (0.22)
Years of experience	14.05	-0.37 (1.47)	1.56 (2.20)	-6.16 (1.04)**	-3.73 (2.09)
Tenured	0.53	-0.26 (0.07)**	-0.21 (0.11)*	-0.44 (0.06)**	-0.34 (0.12)**
School fixed effects		No	Yes	No	Yes

Notes: Dependent variable is a dummy for teachers who left the sample between years (columns 1 and 2) or within year (columns 3 and 4); explanatory variable is given in first column of table. Sample for regressions in columns (1) and (2) is all teachers in sample schools who completed the 2011 school year and for whom explanatory variables are available (449 teachers). Sample for regressions in columns (3) and (4) is all teachers who began the 2012 school year and for whom 2011 CLASS is available (369 teachers). Standard errors clustered at school level. * significant at 5% level, ** at 1% level.

Appendix Table C3: Characteristics of students in classes where teacher left within year

	(1)	(2)	(3)	(4)
	Observations	Mean		
TVIP	11,486	0.00	-0.09*	-0.03
			(0.04)	(0.03)
Age (months)	11,472	60.30	-0.13	-0.05
			(0.17)	(0.13)
Female	11,486	0.49	0.00	0.01
			(0.02)	(0.01)
Wealth index	11,156	0.02	-0.04	0.00
			(0.07)	(0.02)
Mother's years of schooling	10,560	8.83	-0.07	-0.02
			(0.19)	(0.09)
School fixed effects			No	Yes

Notes: Dependent variable in first column of the table; explanatory variable is a dummy for teachers who left the sample during the 2012 school year. Sample is all children who began 2012 school year. Standard errors clustered at school level. *significant at 5% level, ** at 1% level.

Appendix D: Dealing with sampling error in the estimates of $V(\gamma_{cs}^k)$ and $\rho_{kl} = \frac{COV(\gamma_{cs}^k, \gamma_{cs}^l)}{\sqrt{V(\gamma_{cs}^k)V(\gamma_{cs}^l)}}$

To calculate classroom effects purged from measurement error, we start from our equation (1):

$$Y_{ics}^k = \delta_{cs}^k + \boldsymbol{X_{ics}}\beta^k + \varepsilon_{ics}^k \qquad k = 1,2,\dots K$$

where Y_{ics}^k are child test scores, δ_{cs}^k are classroom indicators, $\boldsymbol{X_{ics}}$ is a vector of child and household characteristics, and ε_{ics}^k is an i.i.d. error term. Given the large sample size we use, we can estimate β^k quite precisely, so in what follows we ignore sampling variation in these estimates.

Therefore, conditioning on $\boldsymbol{X_{ics}}$, $\hat{\delta}_{cs}^k = \delta_{cs}^k + \frac{\sum_{i=1}^{N_{cs}} \varepsilon_{ics}^k}{N_{cs}}$. Let α_{cs}^k be the true classroom effects, net of school effects denoted by θ_s^k. Then $\delta_{cs}^k = \theta_s^k + \alpha_{cs}^k$. Finally, define the demeaned classroom effect as $\gamma_{cs}^k = \delta_{cs}^k - \frac{\sum_{d=1}^{C_s} N_{ds}\delta_{ds}^k}{\sum_{d=1}^{C_s} N_{ds}} = \alpha_{cs}^k - \frac{\sum_{d=1}^{C_s} N_{ds}\alpha_{ds}^k}{\sum_{d=1}^{C_s} N_{ds}}$. We then have that:

$$\hat{\gamma}_{cs}^k = \hat{\delta}_{cs}^k - \frac{\sum_{d=1}^{C_s} N_{ds}\hat{\delta}_{ds}^k}{\sum_{d=1}^{C_s} N_{ds}} = \left(\delta_{cs}^k - \frac{\sum_{d=1}^{C_s} N_{ds}\delta_{ds}^k}{\sum_{d=1}^{C_s} N_{ds}}\right) + \left(\frac{\sum_{i=1}^{N_{cs}} \varepsilon_{ics}^k}{N_{cs}} - \frac{\sum_{d=1}^{C_s}\sum_{i=1}^{N_{ds}} \varepsilon_{ids}^k}{\sum_{d=1}^{C_s} N_{ds}}\right)$$

(see also equation (7) in Chetty et al (2011), and their Appendix B). We can rewrite this as:

$$\hat{\gamma}_{cs}^k = \gamma_{cs}^k + \left[\frac{\left(\sum_{d=1}^{C_s} N_{ds}\right) - N_{cs}}{N_{cs}\left(\sum_{d=1}^{C_s} N_{ds}\right)}\sum_{i=1}^{N_{cs}} \varepsilon_{ics}^k - \frac{1}{\sum_{d=1}^{C_s} N_{ds}}\sum_{d=1,d\neq c}^{C_s}\sum_{i=1}^{N_{ds}} \varepsilon_{ids}^k\right]$$

Finally:

$$V(\hat{\gamma}_{cs}^k) = V\left\{\gamma_{cs}^k + \left[\frac{\left(\sum_{d=1}^{C_s} N_{ds}\right) - N_{cs}}{N_{cs}\left(\sum_{d=1}^{C_s} N_{ds}\right)}\sum_{i=1}^{N_{cs}} \varepsilon_{ics}^k - \frac{1}{\sum_{d=1}^{C_s} N_{ds}}\sum_{d=1,d\neq c}^{C_s}\sum_{i=1}^{N_{ds}} \varepsilon_{ids}^k\right]\right\}$$

Assuming that the ε_{ics}^k are homoscedastic with variance σ^2 and independent of γ_{cs}^k (because of random assignment), after some algebra we get that:

$$V(\hat{\gamma}_{cs}^k) = V(\gamma_{cs}^k) + E\left\{\frac{\left[\left(\sum_{d=1}^{C_s} N_{ds}\right) - N_{cs}\right]}{N_{cs}\left(\sum_{d=1}^{C_s} N_{ds}\right)}\sigma^2\right\}$$

This means that:

$$V(\gamma_{cs}^k) = V(\hat{\gamma}_{cs}^k) - E\left\{\frac{\left[\left(\sum_{d=1}^{C_s} N_{ds}\right) - N_{cs}\right]}{N_{cs}\left(\sum_{d=1}^{C_s} N_{ds}\right)}\sigma^2\right\}$$

In order to estimate $\rho_{kl} = \frac{COV(\gamma_{cs}^k, \gamma_{cs}^l)}{\sqrt{V(\gamma_{cs}^k)V(\gamma_{cs}^l)}}$ we still need to develop an estimate of the numerator of this expression free of sampling error. Let σ^{kl} be the covariance between ε_{ids}^k and ε_{ids}^l. If we estimate equation (1) jointly for both tests, we can recover this covariance from the variance covariance matrix of the residuals. Then, proceeding as above, we can show that:

$$COV(\gamma_{cs}^k, \gamma_{cs}^l) = V(\hat{\gamma}_{cs}^k, \hat{\gamma}_{cs}^l) - E\left\{\frac{\left[\left(\sum_{d=1}^{C_s} N_{ds}\right) - N_{cs}\right]}{N_{cs}\left(\sum_{d=1}^{C_s} N_{ds}\right)}\sigma^{kl}\right\}$$

References

Chetty, R., J. Friedman, N. Hilger, E. Saez, D. Schanzenbach, and D. Yagan. 2011. "How Does your Kindergarten Classroom Affect your Earnings? Evidence from Project STAR." *Quarterly Journal of Economics* 126(4): 1593-1660.

Appendix E: Teacher characteristics, behaviors, and learning outcomes

This Appendix reports estimates of regressions of learning outcomes on teacher characteristics and behaviors, separating variables which are observed before child learning outcomes are measured, and those that are contemporaneous with learning outcomes.

Appendix Table E1: Teacher characteristics and behaviors, and child learning outcomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Language		**Math**		**Executive function**		**Total**	
Teacher has 3 years of experience or less	-0.12		-0.09		-0.09		-0.12*	
	(0.06)		(0.07)		(0.06)		(0.06)	
Tenured teacher	0.02		0.06		-0.01		0.03	
	(0.04)		(0.04)		(0.04)		(0.04)	
Parents' education	0.01		0.01		0.00		0.01	
(Average years)	(0.01)		(0.00)		(0.01)		(0.01)	
Lagged CLASS	0.05*		0.07*		0.06*		0.07**	
	(0.03)		(0.03)		(0.02)		(0.03)	
IQ		0.04*		0.04		0.03		0.04*
		(0.02)		(0.02)		(0.02)		(0.02)
Neuroticism		0.00		0.01		0.02		0.01
		(0.02)		(0.02)		(0.02)		(0.02)
Extraversion		0.03		0.03		0.03		0.04
		(0.02)		(0.02)		(0.02)		(0.02)
Openness		0.01		0.03		0.02		0.02
		(0.02)		(0.02)		(0.02)		(0.02)
Agreeableness		-0.02		-0.02		-0.03		-0.03
		(0.02)		(0.02)		(0.02)		(0.02)
Conscientiousness		-0.02		-0.05*		-0.03		-0.04*
		(0.02)		(0.02)		(0.02)		(0.02)
Inhibitory control & attention		0.01		0.02		0.03		0.02
		(0.02)		(0.02)		(0.02)		(0.02)
R-squared	0.47	0.47	0.35	0.35	0.30	0.30	0.46	0.46
F-test (p-value)	0.06	0.21	0.00	0.07	0.05	0.02	0.01	0.04
Students	7,978	7,978	7,978	7,978	7,978	7,978	7,978	7,978
Classrooms	269	269	269	269	269	269	269	269
Schools	125	125	125	125	125	125	125	125

Notes: The table reports estimates from regressions of test scores on teacher characteristics and behaviors. All regressions are limited to children in schools in which at least two teachers taught kindergarten in both the 2011 and 2012 school years. All regressions include baseline student and household characteristics, their classroom averages, and school fixed effects. Standard errors (in parentheses) clustered at the school level. * significant at 5%, ** at 1%.

Appendix F: Parental inputs, behaviors, and learning outcomes

This Appendix provides additional information on the association between end-of-year child learning outcomes and individual parental inputs and behaviors (controlling for baseline child TVIP). Appendix Table F1 shows that three behaviors that parents carry out with children significantly predict end-of-year test scores: reading books, telling stories, and singing to the child. These behaviors are aggregated into a parental behaviors index. Appendix Table F2 shows that the availability of most toys and learning materials are significantly associated with higher end-of-year test scores. All of the inputs other than the first two in the table are aggregated into a parental inputs index. Within each index, each included behavior or input receives equal weight.

Appendix Table F1: Parental behaviors and child test scores

	Obs.	Mean	Coefficient (standard error)
Read books, watch pictures or drawings in a book with child	7,945	0.41	0.08** (0.02)
Tell stories to child	7,946	0.42	0.08** (0.02)
Sing to, or sing with, child	7,946	0.64	0.08** (0.02)
Go for a walk with child	7,952	0.78	0.04 (0.02)
Play with child with his toys	7,943	0.51	0.02 (0.02)
Draw or paint with child	7,949	0.66	0.01 (0.02)
Play with child to name or count objects or colors	7,948	0.77	0.03 (0.03)

Notes: Dependent variable is test score aggregate for twelve tests (mean zero, unit standard deviation), explanatory variable given in first row of table. Each row corresponds to a separate regression. All specifications include TVIP as a control. Standard errors clustered at the school level. * significant at 5%, ** at 1%.

Appendix Table F2. Parental inputs and child test scores

	Obs.	Mean	Coefficient (standard error)
Toys made at home	7,952	0.12	0.02
			(0.03)
Toys bought in a store	7,952	0.93	0.06
			(0.04)
Musical toys	7,952	0.42	0.13**
			(0.02)
Blocks for construction	7,952	0.47	0.14**
			(0.02)
Writing, drawing, or painting material	7,952	0.60	0.07*
			(0.03)
Toys that require physical movement	7,952	0.80	0.06*
			(0.03)
Dolls or toys for pretend play	7,952	0.76	0.14**
			(0.03)
Children's coloring books	7,952	0.38	0.12**
			(0.02)
Children's story books	7,952	0.33	0.13**
			(0.02)
Toys to learn shapes, figures, or colors	7,952	0.24	0.09**
			(0.03)

Notes: Dependent variable is test score aggregate for twelve tests (mean zero, unit standard deviation), explanatory variable given in first row of table. Each row corresponds to a separate regression. All specifications include TVIP as a control. Standard errors clustered at the school level. * significant at 5%, ** at 1%.