IDB WORKING PAPER SERIES Nº IDB-WP-00939

Mothers, Teachers, Peers, and the Gender Gap in Early Math Achievement

Pedro Carneiro
Yyannú Cruz-Aguayo
Norbert Schady

Inter-American Development Bank
Social Sector - SCL/SCL

April, 2018

Mothers, Teachers, Peers, and the Gender Gap in Early Math Achievement

Pedro Carneiro*
Yyannú Cruz-Aguayo**
Norbert Schady**

*University College London
**Inter-American Development Bank

April, 2018

Cataloging-in-Publication data provided by the
Inter-American Development Bank
Felipe Herrera Library
Carneiro, Pedro.
Mothers, teachers, peers, and the gender gap in early math achievement / Pedro Carneiro, Yyannú Cruz Aguayo, Norbert Schady.
p. cm. — (IDB Working Paper Series ; 939)
Includes bibliographic references.
1. Mathematics-Study and teaching (Early childhood)-Ecuador. 2. Academic achievement-Sex differences-Ecuador. 3. Mathematical ability in children-Testing-Ecuador. 4. Sex differences in education-Ecuador. I. Cruz Aguayo, Yyannú. II. Schady, Norbert Rüdiger, 1967- III. Inter-American Development Bank. Social Sector. IV. Title. V. Series.
IDB-WP-939

http://www.iadb.org

Mothers, Teachers, Peers, and the Gender Gap in Early Math Achievement

Pedro Carneiro
Yyannú Cruz-Aguayo
Norbert Schady

April 1, 2018[1]

[1] Carneiro: University College London, Institute for Fiscal Studies, and Centre for Microdata Methods and Practice. Cruz-Aguayo and Schady: Inter-American Development Bank. Carneiro gratefully acknowledges the support of the ESRC for CEMMAP (ES/P008909/1) and the ERC through grant ERC-2015-CoG-682349. We thank Jere Behrman, Gregory Elacqua, Costas Meghir, and Andrew Morrison for their comments, Alejandra Campos, Nicola Dehnen, Nicolás Fuertes, Rafael Hernández and Matías Martínez for outstanding research assistance, and the Government of Ecuador for collaboration at every step in this research project.

Abstract

We study the determinants of math achievement among children in early elementary school using data from a unique experiment in which children were randomly assigned to classrooms within schools for four consecutive grades. As a result, each child in our sample was exposed to four separate, orthogonal shocks to the quality of teachers and peers. We first show that there are steep socioeconomic gradients and a substantial boy-girl gap in math test scores. However, among children of mothers with university education, there is no difference in the math achievement of girls and boys. We use the experimental design to test for differences in how boys and girls respond to different measures of the quality of their classrooms, teachers, and peers. We find no evidence of differential responses by gender.

JEL classification: I21

Keywords: test scores, teachers, gender

1. Introduction

Parents, teachers, and peers are the most important determinants of learning outcomes in childhood. In this paper, we use data from a unique, multi-year policy experiment in Ecuador, a middle-income country in South America, to analyze how these inputs affect the relative math achievement of boys and girls in early elementary school.

Our focus on math is motivated by two important facts. First, early math test scores strongly predict later educational attainment and achievement. In the United States, children who have one-standard deviation higher math scores at the beginning of kindergarten have math scores that are 0.41 standard deviations higher in 1st through 8th grades, on average (Duncan et al. 2007). Relative to the rest of the population, children in the lowest quartile of the distribution of math test scores at ages 6, 8, and 10 are 13 percentage points less likely to graduate from high school, and 34 percentage points less likely to attend college (Duncan and Magnuson 2011).

The second reason we focus on math is that girls lag behind boys in math achievement in elementary school and, especially, in high school. In the United States, Fryer and Levitt (2010) use the ECLS-K to show that, by 3rd grade, boys have math test scores that are 0.2 standard deviations higher than girls. The average score of girls on the math component of the Scholastic Aptitude Test (SAT) is 30-40 points (0.3-0.4 standard deviations) below that of boys. Since early math scores predict later math scores, understanding the determinants of early math achievement may provide some clues about later career choices, including about the relative absence of women in science, technology, engineering, and math (STEM) jobs.

We study early math achievement using data from a cohort of approximately 13,500 children who were assigned to kindergarten classrooms within 204 schools with a rule that is as-good-as-random. These children were re-assigned to different classrooms in 1st, 2nd, and 3rd grades. Compliance with the as-good-as random assignment was very high—98.8 percent on average. At the end of each grade, age-appropriate math tests were applied to children. We have rich data on teachers, including on the quality of the interactions between teachers and children, as measured by a rubric known as the Classroom Assessment Scoring System (CLASS, Pianta et al. 2007).

We make several contributions to the literature on human capital formation in childhood. Like others, we find that children of low socioeconomic status (SES) and girls in our sample have substantially lower math achievement: A child whose mother has university education has math scores that are 0.52 standard deviations higher than a comparable child (in age and gender) whose mother has only an elementary school education. Girls have math scores that are on average 0.14 standard deviations lower than boys, with particularly large gender gaps in the right tail of the distribution. The variance of test scores

of boys is larger than that of girls, although the difference is modest—a ratio of the boy-girl variance of 1.09

Next, we carefully explore boy-girl differences in achievement for different components of the math tests. This is motivated by earlier findings in the literature that suggest that, because boys and girls excel at different math tasks, the magnitude of the boy-girl gap in achievement can be very sensitive to test content.[2] We show that in our sample boys outperform girls in all sections of the tests we use, although there are differences in the magnitude of the gender achievement gap by task.

Having established these facts, we turn to possible explanations for the gender gap we observe. We first focus on the quality of home inputs, proxied by maternal education. We show that, on average, girls with mothers who have university education do *not* have lower math test scores than boys of similarly-educated mothers. The "protective" effect of maternal university education on the math achievement of girls, relative to boys, appears to be most pronounced among single mothers. We cannot cleanly establish causality between maternal education and child achievement, but our results suggest that mothers may play an important role in determining the early math achievement of girls.

Schools could in principle affect the gender gap in math achievement in various ways. For example, all teachers in Ecuador could call on boys in class more than on girls, and this could affect math achievement. Our data are not well-suited to analyze such behaviors. However, we can analyze how girls and boys respond to as-good-as-random variation in the quality of classrooms, teachers, and peers. This is of interest in its own right and, as we discuss in the next section, is a question that has not been resolved in the literature.

To test for differential responses to classroom quality by boys and girls, we build on our earlier work on learning outcomes in kindergarten (Araujo et al. 2016), as well as on the large literature on teachers in the United States.[3] We begin by estimating "classroom effects" separately for boys and girls. These classroom effects measure the extent to which there is more (or less) learning in a given classroom, relative to others within the same grade and school (the level at which the as-good-as-random assignment was carried out). We show that classroom effects for boys and girls are very close in magnitude and are not statistically different from each other.

---

[2] Casey et al. (1995) show that, when questions which rely on visuospatial abilities are statistically removed from the math SAT, the sex difference in scores disappears; Gallagher and DeLisi (1994) and Gallagher et al. (2002) find a similar pattern for the Quantitative Reasoning section of the Graduate Record Examination (GRE). Some of the differences in the magnitude of the gender gap that are reported in the literature, even for children of the same age, in the same country, at the same point in time, could also reflect differences in test content. For example, the boy-girl difference in math achievement of U.S. high school students on the 2015 PISA math test, applied to 15-year-olds, was 0.09 standard deviations (OECD 2016), while the boy-girl difference in the SAT in 2015 was 0.31 standard deviations (Perry 2016).

[3] See Chetty et al. (2014a, 2014b); Jackson et al. (2014); Jacob et al. (2010); Rivkin et al. (2005); and Staiger and Rockoff (2010), among many important contributions.

Furthermore, in kindergarten, we collected data on the learning outcomes of two cohorts of children assigned to the same teachers in subsequent grades, so we can calculate "teacher effects" that purge the classroom effects of any classroom shocks. Here, too, we cannot reject the null hypothesis that the effects for boys and girls are the same, although these results are imprecise.

Next, we turn to data on the quality of teacher-child interactions, as measured by the CLASS. In our earlier work, we showed that kindergarten children who were randomly assigned to teachers with higher CLASS scores learned more. In this paper, we first show that the CLASS is more strongly associated with learning outcomes in kindergarten and 1st grade than in 2nd and 3rd grades. Averaging across grades, however, we cannot reject the null that the association between teacher CLASS scores and math achievement is the same for boys and girls.

We also test whether girls benefit more from having a female teacher than boys, as has been found by Dee (2005) in the United States, and Muralidharan and Sheth (2016) in India. Teacher gender does not affect the math achievement of boys or girls in our sample.

In the last part of our analysis, we turn to the possible role of peers. Despite as-good-as-random assignment, there is some naturally occurring variation in the baseline test scores of children in one or another class in the same school. We exploit this variation to test whether having peers with higher baseline test scores increases own math achievement, and whether any effect is different for boys and girls. We show that peer quality measured in this way does not predict overall math achievement, or the gender gap in achievement.

Similarly, there is some variation in the proportion of classmates who are girls in our sample. This allows us to test whether having more girls in the classroom affects math achievement, as suggested by Hoxby (2000). We show that having a higher proportion of female classmates improves the math test scores of girls, although the magnitude of the effect is modest and the coefficient is only borderline significant.

In sum, by and large there do not appear to be substantive or significant differences in how boys and girls respond to variation in the quality of classrooms, teachers, or peers. On the other hand, the associations between maternal university education and the relative test scores of girls and boys, especially in single-mother households, suggest that there are aspects of the home environment that are important in explaining early gender gaps in math achievement.

The rest of the paper proceeds as follows. In section 2, we briefly discuss the earlier literature on gender, SES, and early math achievement. Section 3 describes the setting and our data. Section 4 discusses identification and presents our main results. We conclude in Section 5.

2. Gender gaps in math achievement[4]

A number of papers in the economics and psychology literature, primarily from the United States, have described, and attempted to explain, gender gaps in math achievement. Three striking facts emerge from this literature.

First, gender gaps in math achievement are more apparent among older than younger children. There is no evidence of a gender gap in math abilities among infants, toddlers, and children of preschool age (Spelke 2005). In elementary school, Fryer and Levitt (2010) use the ECLS-K to show that, by 3rd grade, boys in the United States have math test scores that are 0.2 standard deviations higher than girls, and Bharadwaj et al. (2012) show that 4th grade boys in Chile have math test scores that are 0.09 standard deviations higher than girls. The strongest evidence of gender gaps in achievement comes from standardized tests taken in high school. The average score of girls on the math component of the Scholastic Aptitude Test (SAT) is 30-40 points (0.3-0.4 standard deviations) below that of boys, and there has been no substantive change in this gender gap for almost 50 years (Perry 2016). Data from the Programme for International Student Assessment (PISA), a test applied to 15-year-olds, show that boys outperform girls in most countries.

Second, there are differences in the *kinds* of math tasks that males and females excel at. Females tend to outperform males in basic computational tasks, while males do better than females on tests that cover material that is not taught in the school curriculum (Geary 1996; Halpern 2000). The largest male advantage is found in math tasks that involve visuospatial abilities, with gender differences of up to one standard deviation (Halpern et al. 2007).

Third, the boy-girl gap in math achievement is largest in the right tail of the distribution. Early work by Benbow and Stanley (1980; 1983) found ratios close to 13 to 1 in the extreme right tail (top 0.01 percent) of the distribution of the math component of the 7th grade SAT, although this ratio appears to have decreased and stabilized around 4 to 1 (Wai et al. 2010). Ellison and Swanson (2010) show that the male-female ratio among those scoring 800 on the SAT is 2.1 to 1. Using data from the American Mathematics Competition, which focuses on very high math achievers, they find male-female ratios larger than 10 to 1 in the extreme right tail of the distribution.

Although the basic facts about gender differences in math achievement are widely accepted, there is vigorous scholarly (and popular) debate about why boys outperform girls in math, in particular on the role of biological and environmental influences. Some have argued for the importance of hormonal differences, or differences in brain structure, between men and women. When women are administered male hormones in preparation for female-to-male sex-change surgery, their performance on some math

[4] In this section, our discussion of the evidence from psychology draws heavily on Halpern et al. (2007).

tasks (in particular, 3-D spatial cognition) improves (Van Goozen et al. 1994; 1995; Voyer et al. 1995). Evolutionary psychologists and anthropologists have pointed out that males in traditional societies carry out tasks that require visuospatial abilities, including tribal warfare and large-game hunting. This, in turn, could have resulted in an evolutionary process that favors brain development in these abilities among males (see the discussion in Halpern et al. 2007).

Others have focused on environmental influences. Crowley et al. (2001) show that, when parents and children use interactive science exhibits at a museum, parents are three times more likely to explain science concepts to boys than girls. In a science game that involved playing with magnets, mothers were more likely to ask boys than girls questions about the scientific process, like their hypotheses and how to test them (Tenenbaum et al. 2005).

Nosek et al. (2009) argue that national differences in gender–science stereotypes predict national sex differences in science and math achievement. Using data from PISA, Else-Quest et al. (2010) and Guiso et al. (2008) argue that the gender gap in math achievement is larger in countries where there is less gender equality in other outcomes, although the empirical work underlying these estimates has been brought into question (Stoet and Geary 2015). Nollenberger et al. (2016) also use the PISA data to show that, among second-generation immigrants, the more unequal are various outcomes by gender in parents' country of origin, the larger are the boy-girl difference in math achievement of their children—even though children of immigrants from different countries are all exposed to the same institutional context, given by their country of birth.

"Stereotype threat" may also play a role. In a well-known study with college students, participants were divided in two groups before taking a math test. In one group, participants were told before taking the test that men generally outperform women, while in the other group participants were not told anything about gender differences in performance. In the first group, men had higher test scores than women, but in the second group they did not (Spencer et al. 1999). A different explanation is found in Niederle and Vesterlund (2010), who argue that the gender difference in math achievement is largely driven by the fact that girls perform less well than boys in competitive test-taking environments.

Environmental influences also include teachers and schools. Teachers could have stereotypes about the math ability of boys and girls that affect how they treat students (as in Lavy and Sand 2015). Teachers may call more on boys than girls in class.[5] Even in the absence of differences in how teachers perceive or treat boys and girls, there could be differences in how boys and girls respond to inputs like

[5] For the United States, see Sadker and Sadker (1995), and Halpern et al. (2007, p. 22, and the references therein); see also Bassi et al. (2016) for a sample of 4th grade classrooms in Chile. Our data are not well-suited to analyze behaviors that are common to all teachers.

teacher quality or better peers. This question has received attention in both psychology and economics. In the economics literature, Autor et al. (2016) use data from Florida and argue that there are larger benefits from school quality for boys than girls. On the other hand, Hastings et al. (2006) and Deming et al. (2014) use data from lottery assignment in the Charlotte-Mecklenburg school district and argue that attending a first-choice school benefits girls more than boys.

3. Setting and data

We study SES, gender, and math achievement in Ecuador, a middle-income country in South America. As is the case in most other Latin American countries, math achievement of young children in Ecuador is low (Berlinski and Schady 2015; Naslund-Hadley and Bando 2015).

The data we use come from an experiment in which an incoming cohort of children was assigned to kindergarten teachers in 204 schools in the 2012 school year with an assignment rule that is as-good-as-random.[6] These children were reassigned to 1st grade teachers in 2013, to 2nd grade teachers in 2014, and to 3rd grade teachers in 2015. Compliance with the assignment rules was very high—98.8 percent on average.[7] As a result, for the main cohort of children we follow, we have four exogenous, orthogonal shocks to classroom quality. In addition, a second cohort of kindergarten children was assigned to teachers with the same assignment rule in 2013. Thus, for the majority of kindergarten teachers in our sample (excepting those who moved schools, or taught a grade other than kindergarten in either year), we have data on the learning outcomes of children in their classrooms for two consecutive years.

As-good-as-random assignment means that we can deal effectively with concerns about any purposeful matching of students with teachers that can arise in a non-experimental setting (Chetty et al. 2014a; Rothstein 2010). Throughout, we work with a balanced panel of 9,013 children for whom we have data on maternal education; their receptive vocabulary at the beginning of kindergarten, as measured by the *Test de Vocabulario en Imágenes Peabody* (TVIP), the Spanish version of the widely used Peabody Picture Vocabulary Test (PPVT) (Dunn et al. 1986);[8] and valid test score data in all four grades. We provide further

[6] These schools are a random sample of all public schools that had at least two kindergarten classrooms in the coastal region of the country. See Araujo et al. (2016) for details.

[7] Random assignment experiments of students to teachers in the United States have had much lower compliance rates. For example, contamination was a serious issue in the influential Measuring Effective Teaching (MET) project (Kane and Staiger 2012). Across the six different sites, compliance with the random assignment ranged from 66 percent (in Dallas) to 27 percent (in Memphis).

[8] Performance on this test at early ages has been shown to predict important outcomes in a variety of settings, including in Ecuador. 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.

details on the assignment rules and compliance in Appendix A, and on changes in the sample that arise as children transfer into or out of our sample of schools in Appendix B.

At the end of each grade, we applied age-appropriate math tests to children. Test scores are normalized by subtracting the grade-specific mean and dividing by the grade-specific standard deviation. Details on test scoring and aggregation are provided below and in Appendix C.

Table 1, Panel A, summarizes the characteristics of children and their families. Children were approximately 5 years old on the first day of kindergarten. Half of them are girls, as expected. At the time children enrolled in kindergarten, mothers were on average in their early thirties, and fathers were in their mid-thirties. Education levels are similar for both parents—just under nine years of school (which corresponds to completed middle school). The average child in our sample has a TVIP score that places her more than 1 standard deviation below the reference population that was used to norm the test, indicating that many children begin formal schooling with significant delays.[9] The baseline TVIP score of girls in our sample is lower than that of boys, a difference of about 0.08 standard deviations.

Table 1, Panel B, summarizes the characteristics of teachers in our sample. Teachers are in their mid-40s, on average, and 89 percent of them are women. The average teacher has 18 years of experience, and three-quarters are tenured (rather than working on a contract basis). Average class size is 37 students.

We measure the quality of teacher-student interactions using the CLASS (Pianta et al. 2007). The CLASS measures teacher behaviors in three domains: emotional support, classroom organization, and instructional support. In practice, in our application of the CLASS (as well as in others), scores across domains are highly correlated with each other. For this reason, we focus on a teacher's total CLASS score. A number of papers using data from the United States have found that children exposed to teachers with better CLASS scores have larger learning gains, better self-regulation, and fewer behavioral problems (references include Howes et al. 2008; Grossman et al. 2013; Kane and Staiger 2012). In our earlier work (Araujo et al. 2016) we found that kindergarten children who were randomly assigned to a teacher with a one-standard deviation higher CLASS score had 0.08-0.09 standard deviation higher math test scores. Further details on the CLASS rubric and the process of filming teachers and coding video footage are given in Appendix D.

Figure 1 graphs univariate densities of the distribution of total CLASS scores for teachers in our study, separately by grade. The CLASS is scored on a 1-7 scale, with scores of 1 or 2 being "low", between 3 and 5 being "medium", and 6 or 7 being "high" quality. The average score of teachers in our sample is 3.3. A few teachers have scores in the "low" range, but the vast majority, more than 80 percent, have scores

[9] 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).

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 United States and Ecuador samples is substantial, equivalent to 2.4 standard deviations of the U.S. sample and 4.7 standard deviations of the Ecuador sample. Fifteen percent of the teachers in the U.S. sample but none of the teachers in the Ecuador sample have scores of 5 or higher.

4. Estimation strategy and results

A. Descriptive results

To motivate our results, Figure 2 presents math achievement by gender, averaged across the four grades. Panel A shows that girls have 0.14 standard deviations lower math achievement than boys. Panel B focuses on distributional differences. The figure shows that girls are slightly over-represented in the bottom part of the distribution of achievement: 53 percent of those below the 10th percentile are girls. There are larger gender differences in the right tail of the distribution: The percentages of girls above the 75th, 90th, 95th, and 99th percentiles of the distribution are 43 percent, 40 percent, 39 percent, and 36 percent, respectively. Put differently, in the very right tail of the distribution of math achievement, there are almost twice as many boys as girls.

Figure 3 focuses on maternal education gradients in math achievement. For this purpose, we classify children into those whose mothers have only elementary school education (39 percent of the sample), those with at least some secondary school (51 percent of the sample), and those with at least some post-secondary education (10 percent of the sample; university, for short—we cannot distinguish among different kinds of post-secondary education in our data). Panel A shows that children of mothers with university education have achievement levels that are 0.52 standard deviations higher than children of mothers with only elementary school education.

The other panels in the figure focus on distributional differences in math achievement by maternal education. Panel B shows that children of low-education mothers are over-represented at the bottom of the distribution, and under-represented at the top. Forty-seven percent of children in the lowest decile of math achievement, but only 29 percent in the highest decile, have mothers with elementary school education. Conversely, Panel D shows that children of high-education mothers are under-represented at the bottom of the distribution, and over-represented at the top. Six percent of children in the lowest decile of math achievement, but 15 percent in the highest decile, have mothers with university education.

B. Maternal education, gender, and math achievement

To further explore the associations between SES, gender, and math achievement, we stack the data for the four grades, and run regressions of the following form:

$$(1)\quad Y_{ihg} = \alpha_g + \beta_1 Age_{ihg} + \beta_2 E2_{ihg} + \beta_3 E3_{ihg} + \beta_4 Female_{ihg} + \beta_5 (E3*Female)_{ihg} + \varepsilon_{ihg},$$

where the subscripts i, h, and g stand for individuals, households, and grades, respectively; Y_{ihg} is the math test score; $E2_{ihg}$ and $E3_{ihg}$ are indicator variables for children with mothers (or fathers, in some specifications) with secondary school and university education, respectively; $Female_{ihs}$ is an indicator variable for girls; $(E3*Female)_{ihg}$ refers to girls of mothers (fathers) with university education; Age_{ihg} is child age and age squared at the beginning of kindergarten; α_g is a set of grade-specific intercepts; and ε_{ihg} is the error term. Standard errors are clustered at the student level.[10]

Our first set of regression results are in Table 2. In each grade, our test has three sections: *number recognition and basic arithmetic* (section 1), *number sense* (section 2), and *word problems* (section 3). Panel A reports the results from regressions of math achievement on a single variable—the indicator for girls. These results are therefore estimates of the boy-girl gap in achievement for different sections of the test, or different ways of scoring or aggregating questions and sections. Columns (1) through (3) show there are large differences in the magnitude of the estimated gender gap in achievement across test sections. The gender gap is 0.115 standard deviations for the *number recognition and basic arithmetic* section, 0.162 for the *number sense* section, and 0.066 for the *word problems* section.[11]

The other columns in Table 2 compare different ways of scoring responses and aggregating test sections. In our preferred specification in column (4) of Table 2, questions within each section are aggregated by Item Response Theory (IRT), and each section is given one-third of the weight in the total score.[12] With this aggregation, the overall boy-girl gap is 0.143 standard deviations. An alternative, in column (5), is to simply count the number of correct responses in each section (rather than using IRT),

[10] We have also considered models including an interaction between E2 and Female. However, the estimated coefficient is always small and statistically insignificant.

[11] Conceivably, this pattern could arise if there were differences in the quality of the test across sections—in the extreme, if a test were completely inappropriate for children of a given age, we would expect there to be no differences in achievement by gender. However, this does not appear to be the case in our data: The achievement gaps between boys whose mothers have more or less education are very similar for all three sections. Rather, it appears that the gender differences in achievement by section reflect true underlying differences by gender in skills, and that these gender differences are larger for some skills—*number sense*—than for others—*word problems*. We also note that the pattern of results in Table 2 is consistent with what has been reported in the psychology literature: Halpern et al. (2007, p. 29) argue that there is "converging evidence" that the largest gender gaps in math achievement favoring boys are found in tasks such as place value and representation on a number line—precisely the math skills that are in the *number sense* section—while the smallest gaps are found in math tasks that are represented in a verbal format—as is the case in the *word problems* section of our tests.

[12] In columns (1) through (3) in the table, individual questions within each section are also aggregated by IRT.

but still give one-third of the weight to each section. With this aggregation, the gender gap is 0.150 standard deviations. Finally, in columns (6) and (7) we test the robustness of the results to different ways of choosing the weights that are given to each section. In both columns, the questions in each section are aggregated by IRT, as in column (4). In column (6), however, we use factor analysis to aggregate the three sections, which results in an estimate of the boy-girl gap of 0.157 standard deviations. In column (7) we use the approach proposed by Anderson (2008), which results in an estimated gap of 0.136 standard deviations.[13]

In sum, although there are differences in the boy-girl gap in achievement for different sections of our math test, our results are robust to alternative ways of calculating a total math score. On the other hand, any conclusion one might draw on the *evolution* of the gender gap as children age is very sensitive to choices of test scoring and aggregation.[14] These concerns have also been highlighted in recent research on the evolution of the Black-White test score gap in the United States (Boyd and Lang 2012). For this reason, we do not analyze changes in the gender gap as children age in this paper.

Panel B refers to results that also include the controls for age, maternal education, and the interaction between maternal university education and girls. The most interesting result in Panel B of Table 2 is the coefficient on the interaction term between maternal university education and gender. This coefficient is generally of about the same magnitude, but of opposite sign, as the coefficient on girls, and is highly significant. F-tests reported at the bottom of the table show that we can never reject the null that the sum of the two coefficients is equal to zero. In other words, among children whose mothers have university education, the math achievement of boys and girls is essentially the same.

To further explore the association between maternal education, gender, and math achievement, we first control for household wealth.[15] These results are in Table 3. Column (1) in the table reproduces our preferred specification from Panel B of Table 2. In column (2) we add a fourth-order polynomial in wealth. Unsurprisingly, controlling for wealth substantially reduces the coefficients on the maternal

[13] The steps for constructing the Anderson aggregate are as follows (1) Suppose we have N tests in an aggregate: test1, test2, test3, ..., testN. Calculate the variance-covariance matrix for these N tests. It should be a symmetric NN matrix. Call that COV; (2) calculate the inverse of COV, call that INVCOV. That should also be a symmetric NN matrix; (3) compute weights for each test, which are needed to construct the weighted average that forms the index. The denominator of each weight is the sum of all the elements in INVCOV. The numerator in each weight is the sum of all the elements in each line of INVCOV (line 1 for test1, line 2 for test2, ..., line N for testN). Weights add up to 1, since the sum of all these line sums will be the sum of all the elements in INVCOV; (4) compute the aggregate using these weights to form a weighted average of all tests.

[14] Depending on how tests are aggregated, one can show that the gender gap increases between kindergarten and 1st grade, but then falls between 1st and 3rd grade; increases from kindergarten to 2nd grade but then falls between 2nd and 3rd grade; or increases in every grade between kindergarten and 3rd grade.

[15] To construct our measure of wealth, we aggregate the following variables by principal components: whether the household has piped water inside the home, whether it is connected to the public sewerage system, the main materials of the roof, walls, and floors (three separate variables), and whether the household has a fridge, TV, computer, and washing machine (four separate variables). The wealth index is given by the first principal component.

education indicators. However, the coefficient on the interaction term between maternal university education and gender changes very little: It is 0.137 (0.056) in column (1), and 0.132 (0.056) in column (2).

Next, we create indicator variables for households in the lowest 40 percent of the distribution of wealth, between the 40th and 90th percentiles, and above the 90th percentile of the distribution. We partition the wealth data in this way to closely mimic the distribution of education. In column (3) of Table 3 we include the indicator variables for medium and high wealth, as well as an interaction term between high wealth and gender. The coefficient on the interaction term is small and is not statistically significant. Finally, in column (4), we include the indicator variables for maternal education and wealth, as well as both interaction terms. The interaction term between high wealth and gender is -0.013 (0.059), while the interaction term between maternal education and gender is 0.143 (0.060). In sum, Table 3 shows that wealth per se does not predict the test scores of girls, relative to boys, and controlling for wealth does not change the association between maternal university education and the gender gap in math achievement.

We also partition the data into children of single mothers, and children who live with both biological parents.[16] Table 4 focuses on children who live with both biological parents. The table shows that the coefficients on the interaction between university education and gender are somewhat larger for mothers than fathers, although neither is significant:[17] 0.105 (0.066) for mothers in column (1), and 0.063 (0.074) for fathers in column (3). The coefficient on the interaction between paternal university education and girls is reduced by half when the regression also includes maternal university education and wealth, as can be seen in column (6). On the other hand, the coefficient on the interaction between maternal university education and girls is unaffected by the inclusion of these additional controls.

Table 5 refers to single mothers, and has the same structure as Table 3.[18] We first note that the coefficient on girls in this sample is somewhat smaller, -0.114 (0.045), than in the sample of married or

[16] We do this for both practical and substantive reasons. On the practical side, the household survey only collected information on the education of all household members who were co-residing with the focal child. Thus, in households with both parents, we have information on the education of fathers and mothers, while in single-mother households there is no information on the education of fathers. There are 9,457 children for whom we have data on child age, gender, the baseline TVIP score, and the end-of-grade math test scores for kindergarten through 3rd grade. Of these children, 7,131 (75.4 percent) lived with both biological parents at the time of the survey, and 1,891 (20 percent) lived only with their mother. We have data on the education of mothers for 9,013 of these 9,457 children, and data on the education of fathers for 7,119 of children who lived with both biological parents, as well as 137 of the 138 children who lived only with their fathers. There are also 270 children (2.9 percent of the total) who lived with neither biological parent, and we do not have data on the education of either parent for these children.

[17] Of course, the education of both parents in two-parent households is correlated: The correlation between years of completed schooling of father and mothers is 0.48. However, there are 438 children who have university-educated mothers but fathers without university education, and 314 children who have university-educated fathers but mothers without university education. These are the two groups that are important for the estimates in Table 4.

[18] We refer to women as "single mothers" if they do not co-reside with the biological father of the focal child. 332 of the 1891 women who do not live with the father of the focal child live with a new partner. To sharpen our comparisons, we do not include these women in the sample of single mothers, or in the sample of children who live with both parents.

cohabitating women, -0.147 (0.021). Moreover, in this sample the association between maternal university education and the math achievement of girls, relative to boys, is very large: the coefficient on the interaction term is 0.239 (0.119). The next three columns in the table show that this conclusion does not change if we control for wealth.

Finally, we use quantile regressions to analyze differences in math achievement between boys and girls at different points in the distribution. To keep the number of results manageable, we focus on maternal education. The regressions for all quantiles are estimated simultaneously, and standard errors are generated using a block bootstrap procedure (each child is a block). We report results in Table 6, for the full sample of children (Panel A) and for the sample of children of single mothers (Panel B).

In both panels, boys outperform girls throughout the distribution of math achievement. The girl disadvantage is larger at the top than at the bottom of the distribution—significantly so, in the case of the full sample of women. There are differences in the two panels, however, in the "protective" effect of university education on the math achievement of girls, relative to boys. In the full sample, boys outperform girls at the 75th percentile of the distribution or above even among children of university-educated mothers. In the sample of single mothers, we can never reject that the math achievement of sons and daughters of university-educated women is the same.

In sum, our data suggest that maternal education, living arrangements, and child gender interact in important ways. That said, we do not know what highly-educated mothers *do* to boost the math achievement of girls, relative to boys, and why this effect is particularly large among single mothers. It is possible that highly-educated women who select into single parenthood have unmeasured characteristics that are correlated with the math achievement of their daughters. It is also possible that mothers have preferences for the math achievement of their daughters, and are better able to act on these preferences when they are highly-educated and single.[19] Our data do not allow us to test these or other hypotheses.

C. Classroom effects

The most novel aspect of our data is the as-good-as-random assignment of children to classrooms, with very high levels of compliance, in four consecutive grades. We use this to carefully analyze possible differences by gender in the returns to the quality of classrooms, teachers, and peers.

However, our results are very similar if we reclassify these women as single mothers. These results are available from the authors upon request.

[19] A number of papers have shown that when women control resources, investments in daughters increase more than investments in sons (for example, Duflo 2003; Rangel 2006). There is no exogenous change of resources in our data, but the finding that maternal university education boosts the math achievement of girls, relative to boys, more in single-mother- than in two-parent-households may be an indication that bargaining power is important: In two-parent households, mothers need to bargain with fathers, in single-mother households, they do not.

We calculate classroom effects, separately by grade, and also separately for boys and girls. For this purpose, as is standard in the literature, we (1) regress end-of-grade math scores on a fourth-order polynomial in lagged math scores, child age and its square, the indicator variables for mothers with secondary school and university education, and classroom fixed effects;[20] (2) construct residualized math scores by subtracting the estimated effects of lagged math scores, age, and maternal education; (3) calculate classroom and school means of these residualized scores; and (4) subtract the school means from the classroom means of residualized scores.

The distribution of the demeaned classroom averages we calculate has mean zero, by construction. In principle, the standard deviation of this distribution is an estimate of how much more learning there is in a classroom with one standard deviation higher quality. In practice, however, the estimated variance (or standard deviation) includes both the variance of the true classroom effects and the variance of sampling error, and is therefore an upward-biased estimate of the variance of classroom effects. As in our earlier work (Araujo et al. 2016), we use a standard Empirical Bayes procedure to estimate the variance of the sampling error, and subtract this from the variance of the observed classroom effects.[21] Furthermore, because we are interested in comparing the magnitude of the classroom effects for boys and girls, we use a block bootstrap procedure (each school is a block) to calculate standard errors of each of the estimated classroom effects.

Results are reported in Table 7. The table has five columns—the first column corresponds to the stacked data for all four grades, and the next columns refer to grade-specific estimates of the classroom effects. The table shows that classroom effects decline monotonically by grade, from 0.13 to 0.10. We cannot reject the null that estimated effects are the same for boys and girls in any grade. In the stacked data, the confidence interval for the boy-girl difference is reasonably tight, and we can rule out differences in classroom effects by gender of 0.014 standard deviations or larger.[22]

D. Teachers and teacher behaviors

[20] In the regressions that do not separate the sample into boys and girls, we also include the indicator variable for gender. We do not have data on lagged math scores for children in kindergarten. In this case, and following our earlier work (Araujo et al. 2016), we include a fourth-order polynomial in the TVIP.

[21] The Empirical Bayes correction we apply takes account of the fact that the school mean is unknown, and must therefore be estimated. See Araujo et al. (2016), Appendix D, and Chetty et al. (2011). In the most flexible formulation this correction factor can be grade-specific. However, estimating so many parameters leads to substantial imprecision, especially when we compare parameters by gender. Therefore, in the estimates we report, we fix the variance of the sampling error to be equal to the average of the grade-specific variances. Conclusions from the more flexible specification with grade-specific estimates of the sampling error are the same, but the estimates are less precise.

[22] The standard deviation of classroom effects for the whole sample is smaller than the standard deviation of classroom effects for boys and girls. This need not be surprising. To see this, consider a case with two random variables with the same variance. The variance of the mean of these variables will be lower than the variance of the individual variables, as long as they are not perfectly correlated (which we do not expect because of within-school random assignment).

As is well known from the literature on teachers, classroom effects include both teacher quality and idiosyncratic classroom shocks (for example, the presence of a particularly difficult child who disrupts learning). With two or more years of data on learning outcomes for children assigned to the same teachers, it is possible to calculate teacher (rather than classroom) effects by taking the square root of the cross-cohort covariance of the classroom effects for the same teacher (Hanushek and Rivkin 2012; McCaffrey et al. 2009). For kindergarten only, our experiment collected data on learning outcomes for two cohorts of children as-good-as-randomly assigned to the same teachers.[23] We use these data to calculate kindergarten teacher effects, once again separately for boys and girls.

Panel A of Table 8 shows that the kindergarten teacher effects we estimate in this way appear to be somewhat larger for girls than boys—0.11 compared to 0.07. However, as is the case with the classroom effects, we cannot reject the null that the teacher effects are the same, although the confidence interval for this difference is large, suggesting that the test has low power.

Although classroom and teacher effects for boys and girls are indistinguishable from one another in our data, it is possible that some teachers are especially effective at teaching boys, while others are especially effective at teaching girls. To explore this, we calculate the cross-cohort correlations between the estimated learning gains a given kindergarten teacher produces for all children, for boys only, and for girls only, and the bootstrapped standard errors of these correlations.[24] These estimates are reported in Panel B of Table 8. The top, left-hand value in the table, 0.29, is the cross-cohort correlation of the teacher effects for all children. The other two diagonal elements in the table indicate that the gender-specific correlations are 0.10 for boys, and 0.24 for girls.

Panel B of Table 8 provides no evidence that some kindergarten teachers are consistently more effective with boys while others are consistently more effective with girls. We cannot reject the null that knowing how much learning a teacher produced among boys in cohort 1 predicts her effectiveness with *girls* in cohort 2 as well as knowing how much learning this teacher produced among girls in cohort 1 (the correlations of 0.17 and 0.24 are statistically indistinguishable from each other); similarly, we cannot reject the null that knowing how much learning a teacher produced among girls in cohort 1 predicts her

[23] For the second kindergarten cohort, we only collected data on two of the three math sections—*number recognition and basic arithmetic*, and *word problems*, but not the third domain, *number sense*. Our estimated teacher effects therefore only refer to the average of these two domains.

[24] To calculate these correlations, we need the variance of the teacher effect for boys, the variance of the teacher effect for girls, and the covariance between the teacher effects for boys and girls. The two variances are given by the square of the gender-specific standard deviations of teacher quality reported in Table 8. The covariance can in principle be estimated by taking the covariance between the boy-specific classroom effects in cohort 1 and the girl-specific classroom effects in cohort 2 for each teacher, or the covariance between the girl-specific classroom effects in cohort 1 and the boy-specific classroom effects in cohort 2 for each teacher. In practice, we take the average of these two quantities.

effectiveness with *boys* in cohort 2 as well as knowing how much learning this teacher produced among boys in cohort 1 (the correlations of 0.13 and 0.10 are statistically indistinguishable from each other).

We next turn to an analysis of teacher characteristics and behaviors. For this purpose, we run regressions of math achievement of the following form:

$$(2)\quad Y_{ihjgs} = \alpha_{sg} + \beta_1 Y_{ihjg\text{-}1s} + \beta_2 X_{ihjgs} + \beta_3 T_{jgs} + \beta_6 (T_{jgs} * Female_{ihjgs}) + \varepsilon_{ihjgs},$$

where Y_{ihjgs} is the math test score of child i in household h with teacher j in grade g and school s; $Y_{ihjg\text{-}1s}$ is a vector containing terms of a fourth-order polynomial in lagged test scores for this child; α_{sg} is a vector of school-by-grade fixed effects; X_{ihjgs} includes child age and its square, an indicator variable for girls, and the indicator variables for mothers with secondary school and university education, respectively; T_{jgs} is a given teacher characteristic or behavior; and ε_{ihjgs} is the error term. Standard errors are clustered at the classroom level, as recommended in Abadie et al. (2017).[25]

We begin by analyzing how teacher gender affects the math achievement of girls and boys. Dee (2005) and Muralidharan and Sheth (2016) argue that female teachers particularly benefit girls using data from the United States and India, respectively. We can only carry out this analysis for 3rd grade because in earlier grades almost all teachers are women.[26] In 3rd grade, 22 percent of the teachers are male, and roughly one-third of the schools in the sample have at least one male and one female teacher. These results, in Panel C of Table 8, show that the coefficient on having a female teacher is very small for boys and girls, and we cannot reject the null hypothesis that they are the same.

Finally, we turn to teacher-child interactions, as measured by the CLASS.[27] Panel D of Table 8 shows that children who are taught by teachers with a one-standard deviation higher CLASS have 0.043 standard deviations higher math achievement, on average. We note that the CLASS is likely to have substantial measurement error, which would bias these coefficients towards zero. In our earlier work (Araujo et al. 2016) we used the CLASS score for kindergarten teachers teaching an earlier cohort of children to instrument the current CLASS with the lagged CLASS.[28] The coefficient on the CLASS in

[25] However, our results are very robust to alternative ways of clustering—for example, clustering at the school level—or to no clustering at all. These results are available from the authors upon request.

[26] The proportions of the teachers that are female in kindergarten, 1st grade, and 2nd grade are 99 percent, 93 percent, and 87 percent, respectively.

[27] The CLASS is missing for two teachers in the sample, who had 20 and 8 children, respectively, in their classes. As a result, the sample size for the stacked (all grades) regression in Panel D of Table 8 has 132 (28 times 4) fewer observations than the regressions in the other tables in the paper.

[28] The coefficient on the CLASS for kindergarten teachers in Table 8, 0.054 (0.013) is not the same as the comparable OLS coefficient in Table IV in Araujo et al. (2016), 0.07 (0.02). This difference is a result of differences in the estimation samples. In Table IV in Araujo et al. (2016), the sample was limited to kindergarten teachers for whom lagged CLASS scores were available, and this restriction does not apply in Table 8 above. On the other hand, the sample in Table IV in

these IV regressions was more than twice as large as those in the comparable OLS regressions. We do not have data on the lagged CLASS scores of teachers in grades other than kindergarten. It seems likely to us, however, that the coefficient on the CLASS for these other grades suffers from a similar degree of attenuation bias (unless measurement error in CLASS is substantially different across grades).

Turning to possible differences by gender, Panel D of Table 8 shows that the association of the CLASS with the math achievement of girls goes down monotonically between kindergarten (0.073) and 3rd grade (0.014).[29] No such pattern is apparent for boys. Most importantly for the discussion in our paper, however, when we stack all the data (first column) we cannot reject the null that the coefficient on the CLASS is the same for boys and girls. Like our earlier results on teacher effects and on the effect of female teachers, the results in Panel D of Table 8 suggest that boys and girls respond similarly to as-good-as-random variation in teacher quality.

E. Peers

Much as girls could in principle be more (or less) sensitive than boys to variations in teacher quality, there could also be gender differences in response to peer quality. In spite of random assignment, there is some naturally occurring variation in the composition of peers in different classrooms in the same grade and school, and we can exploit this variation for identification.

To analyze peer effects, we proceed in the same way as with teachers, but replace the teacher variable T_{jgs} in equation (2) with a variable reflecting the composition of peers in the classroom, P_{cgs}, where the subscript c stands for classrooms. We consider two measures of peer "quality". The first is the leave-*i*-out (jackknifed) mean of lagged test scores in the classroom. This builds on a substantial literature that has tested for linear-in-means effects of peer achievement on own achievement, with mixed results, including in elementary school.[30] The second measure builds on Hoxby (2000), and is the proportion of peers in the classroom who are girls.

Results are in Table 9, which has the same structure of Panel D of Table 8. To put things in context, we note that the median within-school, across-classroom difference in lagged peer test scores is

---

Araujo et al. (2016) included math test scores for all children in the kindergarten classrooms in the sample, while in the current paper we restrict the sample to children with test score data in all four grades, as well as data on maternal education and the baseline TVIP.

[29] The variance of the CLASS is larger in kindergarten (0.080), than in 1st grade (0.052), 2nd grade (0.057), or 3rd grade (0.056). If instead of standardizing the CLASS to have a grade-specific zero mean and unit standard deviation we work with the unstandardized scores, the coefficients on the CLASS in the "all children" row in Table 8 are 0.191 (0.044) for kindergarten, 0.263 (0.053) for 1st grade, 0.134 (0.055) for 2nd grade, and 0.096 (0.042) for 3rd grade.

[30] Important references for elementary school include Angrist and Lang (2004), Graham (2008), Hanushek et al. (2003), Hoxby (2000), Imberman et al. (2012), and Whitmore (2005), all with U.S. data.

0.16 standard deviations; at the 90th percentile, the difference is 0.42 standard deviations.[31] Similarly, the median within-school, across-classroom difference in the number of boys and girls is 2, relative to a mean class size of 37 children; at the 90th percentile, the difference is 8.[32]

Panel A shows there is no evidence that the lagged math test scores of peers affect own math achievement, in any grade, for boys or for girls. In the stacked data, we can rule out positive peer effects of 0.052 standard deviations or larger with 95 percent confidence. Panel B, on the other hand, finds suggestive evidence that having more girls as classmates raises own math achievement of girls. In the stacked data, a 10-percentage point increase in the proportion of peers who are girls leads to a 0.021 standard deviation increase in own test scores of girls, although this result is only borderline significant.[33] Moreover, we cannot reject the null that the effect is the same for boys and girls.

In sum, and subject to the caveat that in our data the cross-classroom variation in peer quality is modest in magnitude, which would keep us from credibly extrapolating to policies that entail large changes in peer composition—for example, ability tracking or single-sex education—we find that the effects of peers on own math achievement are at most small, and do not vary by gender.

5. Conclusion

In this paper, we study the math gender gap in early elementary school. Our analysis of how different inputs—mothers, teachers, peers—affect early math achievement, including differences in achievement between boys and girls, has implications for a broader understanding of the process of human capital formation at early ages. The data we use are unique, including four rounds of as-good-as-random assignment of children to classrooms, teachers, and peers.

Much as has been found by others, we show that boys have higher math achievement than girls on average. Underlying the overall boy advantage, however, there is interesting heterogeneity. Among children of mothers with university education, there is no difference in mean math achievement by gender. Further research to identify the effects of socioeconomic status generally, and maternal education specifically, on math learning at early ages seems to us an important priority.

We next turn to the possible role of classroom, teachers, and peers. By and large, we find little evidence that boys and girls respond differently to these inputs into the production of early math

[31] In their sample of non-tracked schools in Kenya, Duflo et al. (2011) report median differences in lagged peer test scores of 0.17 standard deviations.

[32] For these calculations, when there are more than two classrooms in a school, we select two at random. Peer scores and the proportion female are calculated for all children for whom we have lagged achievement data and for all children in the class, respectively, including those who entered the panel after the beginning of kindergarten or left at some point between kindergarten and the end of 3rd grade (and who are therefore not included in the main regressions of the paper).

[33] This implies that, in a class of 40 students, increasing the number of girls from 20 to 30 would increase the test score of girls by 0.052 standard deviations.

achievement. Thus, as-good-as-random variation in the quality of these dimensions of the school environment is unlikely to explain the gender differences in the means and distributions of test scores we observe in our data. A corollary of this conclusion is that across-the-board improvements in the quality of teachers or peers would probably do little to close the gender gap in math achievement. There would be high returns to policy experimentation, and careful evaluation, of more targeted school-based interventions that seek to ensure that girls do not fall behind in math achievement as they progress through the school system.

References

Anderson, Michael L. 2008. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecederian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association* 103(484): 1481-95.

Angrist, Joshua, and Kevin Lang. 2004. "Does School Integration Generate Peer Effects? Evidence from Boston's Metco Program." *American Economic Review* 94(5): 1613-34.

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

Autor, David H., David N. Figlio, Krzystof Karbownik, Jeffrey Roth, and Melanie Wasserman. 2016. "School Quality and the Gender Gap in Educational Achievement." *American Economic Review, Papers and Proceedings* 106(5): 289-95.

Bassi, Marina, Rae Lesser Blumberg, and Mercedes Mateo. 2016. "Under the 'Cloak of Invisibility': Gender Bias in Teaching Practices and Learning Outcomes." Inter-American Development Bank Working Paper 696.

Benbow, Camilla Persson, and Julian C. Stanley. 1980. "Sex Differences in Mathematical Ability: Fact or Artifact?" *Science* 210(4475): 1262-64.

----------. 1983. "Sex Differences in Mathematical Ability: More Facts." *Science* 222(4627): 1029-31.

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

Bharadwaj, Prashant, Giacomo De Giorgi, David Hansen, and Christopher Neilson. 2012. "The Gender Gap in Mathematics: Evidence from Low- and Middle-Income Countries." NBER Working Paper 18464.

Bond, Timothy N., and Kevin Lang. 2013. "The Evolution of the Black-White Test Score Gap in Grades K-3: The Fragility of the Results." *The Review of Economics and Statistics* 95(5): 1468-79.

Casey, M. Beth, Ronald Nuttall, Elizabeth Pezaris, and Camilla Benbow. 1995. "The Influence of Spatial Ability on Gender Differences in Mathematics College Entrance Test Scores across Diverse Samples." *Developmental Psychology* 31(4): 697–705.

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

Chetty, Raj, John Friedman, and Jonah Rockoff. 2014a "Measuring the Impact of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review* 104(9): 2593–632.

———. 2014b. "Measuring the Impact of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review* 104(9): 2633–679.

Clifford, Dick, Donna Bryant, Margaret Burchinal, Oscar Barbarin, Diane Early, Carollee Howes, Robert Pianta, and Pam Winton. 2003. "National Center for Early Development and Learning Multistate Study of Pre-Kindergarten, 2001–2003." Available at http://doi.org/10.3886/ICPSR04283.v3.

Crowley, Kevin, Maureen A. Callanan, Harriet R. Tenenbaum, and Elizabeth E. Allen. 2001. "Parents Explain More Often to Boys Than to Girls During Shared Scientific Thinking." *Psychological Science* 12(3): 258–61.

Dee, Thomas S. 2005. "A Teacher Like Me: Does Race, Ethnicity, or Gender Matter?" *American Economic Review* 95(2): 158-65.

Deming, David J., Justine S. Hastings, Thomas J. Kane, and Douglas O. Staiger. 2014. "School Choice, School Quality, and Postsecondary Attainment." *American Economic Review* 104(3): 991–1013.

Duflo, Esther. 2003. "Grandmothers and Granddaughters: Old-Age Pensions and Intrahousehold Allocation in South Africa." *World Bank Economic Review* 17(1): 1-25.

Duflo, Esther, Pascaline Dupas, and Michael 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.

Duncan, Greg J., Chantelle J Dowsett, Amy Claessens, Katherine Magnuson, Aletha C. Huston, Pamela Klebanov, Linda S. Pagani, Leon Feinstein, Mimi Engel, Jeanne Brooks-Gunn, Holly Sexton, Kathryn Duckworth, and Crista Japel. 2007. "School Readiness and Later Achievement." *Developmental Psychology* 43(6): 1428-46.

Duncan, Greg J., and Katherine Magnuson. 2011. "The Nature and Impact of Early Achievement Skills, Attention Skills, and Behavior Problems." In Greg J. Duncan and Richard J. Murnane, eds., *Whither Opportunity: Rising Inequality, Schools, and Children Life Chances* (pp. 47-70). New York: Russell Sage Foundation.

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

Else-Quest, Nicole M., Janet Shibley Hyde, and Marcia C. Linn. 2010. "Cross-National Gender Differences in Mathematics: A Meta-Analysis." *Psychological Bulletin* 136(1): 103-27.

Ellison, Glenn, and Ashley Swanson. 2010. "The Gender Gap in Secondary School Mathematics at High Achievement Levels: Evidence from the American Mathematics Competitions." *Journal of Economic Perspectives* 24(2): 109-28.

Fryer, Ronald G., and Steven D. Levitt. 2010. "An Empirical Analysis of the Gender Gap in Mathematics." *American Economic Journal: Applied Economics* 2(2): 210-40.

Gallagher, Ann M., and Richard DeLisi. 1994. "Gender Differences in Scholastic Aptitude Test-Mathematics Problem Solving Among High-Ability Students." *Journal of Educational Psychology* 86(2): 204–11.

Gallagher, Ann M., Jutta Levin, and Cara Calahan. 2002. "Cognitive Patterns of Gender Differences on Mathematics Admissions Tests." ETS Report 02-19. Princeton, NJ: Educational Testing Service.

Geary, David C. 1996. "Sexual Selection and Sex Differences in Mathematical Abilities." *Behavioral and Brain Sciences* 19(2): 229–84.

Graham, Bryan S. 2008. "Identifying Social Interactions Through Conditional Variance Restrictions." *Econometrica* 76(3): 643-60.

Grossman, Pam, Susanna Loeb, Julia Cohen, Karen Hammerness, James Wyckoff, Donald Boyd, and Hamilton Lankford. 2013. "Measure for Measure: The Relationship between Measures of Instructional Practice in Middle School English Language Arts and Teachers' Value Added Scores." *American Journal of Education* 119(3): 445-70.

Guiso, Luigi, Ferdinando Monte, Paolo Sapienza, and Luigi Zingales. 2008. "Culture, Gender, and Math." *Science* 320(5880): 1164-65.

Halpern, Dianne F. 2000. *Sex Differences in Cognitive Abilities* (3rd ed.). Mahwah, NJ: Erlbaum.

Halpern, Diane F., Camilla P. Benbow, David C. Geary, Ruben C. Gur, Janet Shibley Hyde, and Morton Ann Gernsbacher. 2007. "The Science of Sex Differences in Science and Mathematics." *Psychological Science in the Public Interest* 8(1): 1-51.

Hanushek, Eric, and Steven Rivkin. 2012. "The Distribution of Teacher Quality and Implications for Policy." *Annual Review of Economics* 4: 131–57.

Hanushek, Eric, John F. Kain, Jacob M. Markman, and Steven G. Rivkin. 2003. "Does Peer Ability Affect Student Achievement?" *Journal of Applied Econometrics* 18(5): 527-44.

Hastings, Justine S., Thomas J. Kane, and Douglas O. Staiger. 2006. "Gender and Performance: Evidence from School Assignment by Randomized Lottery." *American Economic Review* 96(2): 232–36.

Howes, Carollee, Margaret Burchinal, Robert Pianta, Donna Bryant, Diane Early, Richard Clifford and Oscar Barbarin. 2008. "Ready to Learn? Children's Pre-Academic Achievement in Pre-Kindergarten Programs." *Early Childhood Research Quarterly* 23(1): 27–50.

Hoxby, Caroline. 2000. "Peer Effects in the Classroom: Learning from Gender and Race Variation." NBER Working Paper 7867.

Imberman, Scott A., Adriana D. Kugler, and Bruce I. Sacerdote. "Katrina's Children: Evidence on the Structure of Peer Effects from Hurricane Evacuees." *American Economic Review* 102(5): 2048-82.

Jackson, Kirabo, Jonah Rockoff, and Douglas Staiger. 2014. "Teacher Effects and Teacher-Related Policies." *Annual Review of Economics* 6: 801–825.

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

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

Lavy, Victor, and Edith Sand. 2015. "On the Origins of Gender Human Capital Gaps: Short and Long Term Consequences of Teachers' Stereotypical Biases." NBER Working Paper 20909.

McCaffrey, Daniel, Tim Sass, J.R. Lockwood, and Kata Mihaly. 2009. "The Intertemporal Variability of Teacher Effect Estimates." *Education Finance and Policy* 4(4): 572–606.

Muralidharan, Karthik, and Ketki Sheth. 2016. "Bridging Education Gender Gaps in Developing Countries: The Role of Female Teachers." *Journal of Human Resources* 51(2): 269-97.

Naslund-Hadley, Emma, and Rosangela Bando. 2015. *All Children Count Overview Report: Early Mathematics and Science Education in Latin America and the Caribbean.* Washington, D.C.: Inter-American

Development Bank.

Niederle, Muriel, and Lise Vesterlund. 2010. "Explaining the Gender Gap in Math Test Scores: The Role of Competition." *Journal of Economic Perspectives* 24(2): 129–144.

Nollenberger, Natalia, Nuria Rodríguez-Planas, and Almudena Sevilla. 2016. "The Math Gender Gap: The Role of Culture." *The American Economic Review*, *Papers and Proceedings* 106(5): 257-61.

Nosek, Brian A., Frederick L. Smyth, N. Sriram, Nicole M. Lindner, Thierry Devos, Alfonso Ayala, Yoav Bar-Anan, Robin Bergh, Huajian Cai, Karen Gonsalkorale, Selin Kesebir, Norbert Maliszewski, Félix Neto, Eero Olli, Jaihyun Park, Konrad Schnabel, Kimihiro Shiomura, Bogdan Tudor Tulbure, Reinout W. Wiers, Mónika Somogyi, Nazar Akrami, Bo Ekehammar, Michelangelo Vianello, Mahzarin R. Banaji, and Anthony G. Greenwald. 2009. "National Differences in Gender–Science Stereotypes Predict National Sex Differences in Science and Math Achievement." *Proceedings of the National Academy of Sciences* 106(26): 10593-97.

OECD. 2016. *Pisa 2015 Results (Volume I): Excellence and Equity in Education*. PISA, OECD Publishing: Paris.

Perry, Mark J. 2016. "2016 SAT Results Confirm Pattern that's Persisted for 50 Years: High School Boys are Better at Math than Girls." Available at https://www.aei.org/publication/2016-sat-test-results-confirm-pattern-thats-persisted-for-45-years-high-school-boys-are-better-at-math-than-girls/. Accessed on March 29, 2018.

Pianta, Robert, Karen LaParo, and Bridgett Hamre. 2007. *Classroom Assessment Scoring System—CLASS*. Baltimore: Brookes.

Rangel, Marcos. 2006. "Alimony Rights and Intrahousehold Allocation of Resources: Evidence from Brazil." *The Economic Journal* 116(513): 627-58.

Rivkin, Steven, Eric Hanushek, and John Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica* 73(2): 417–58.

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

Sadker, Myra, and David Sadker. 1995. *Failing at Fairness: How America's Schools Cheat Girls*. New York, NY: Scribner's Sons/MacMillan Publishing Co.

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

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

Spelke, Elizabeth S. 2005. "Sex Differences in Intrinsic Aptitude for Mathematics and Science? A Critical Review." *American Psychologist* 60(9): 950–58.

Spencer, Steven J., Claude M. Steele, and Diane M. Quinn. 1999. "Stereotype Threat and Women's Math Performance." *Journal of Experimental Social Psychology* 35(1): 4–28.

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

Stoet, Gijsbert, and David C. Geary. 2015. "Sex Differences in Academic Achievement Are Not Related to Political, Economic, or Social Equality." *Intelligence* 48: 137-51.

Tenenbaum, Harriet R., Catherine E. Snow, Kevin A. Roach, and Brenda Kurland. 2005. "Talking and Reading Science: Longitudinal Data on Sex Differences in Mother-Child Conversations in Low-Income Families." *Applied Developmental Psychology* 26(1): 1-19.

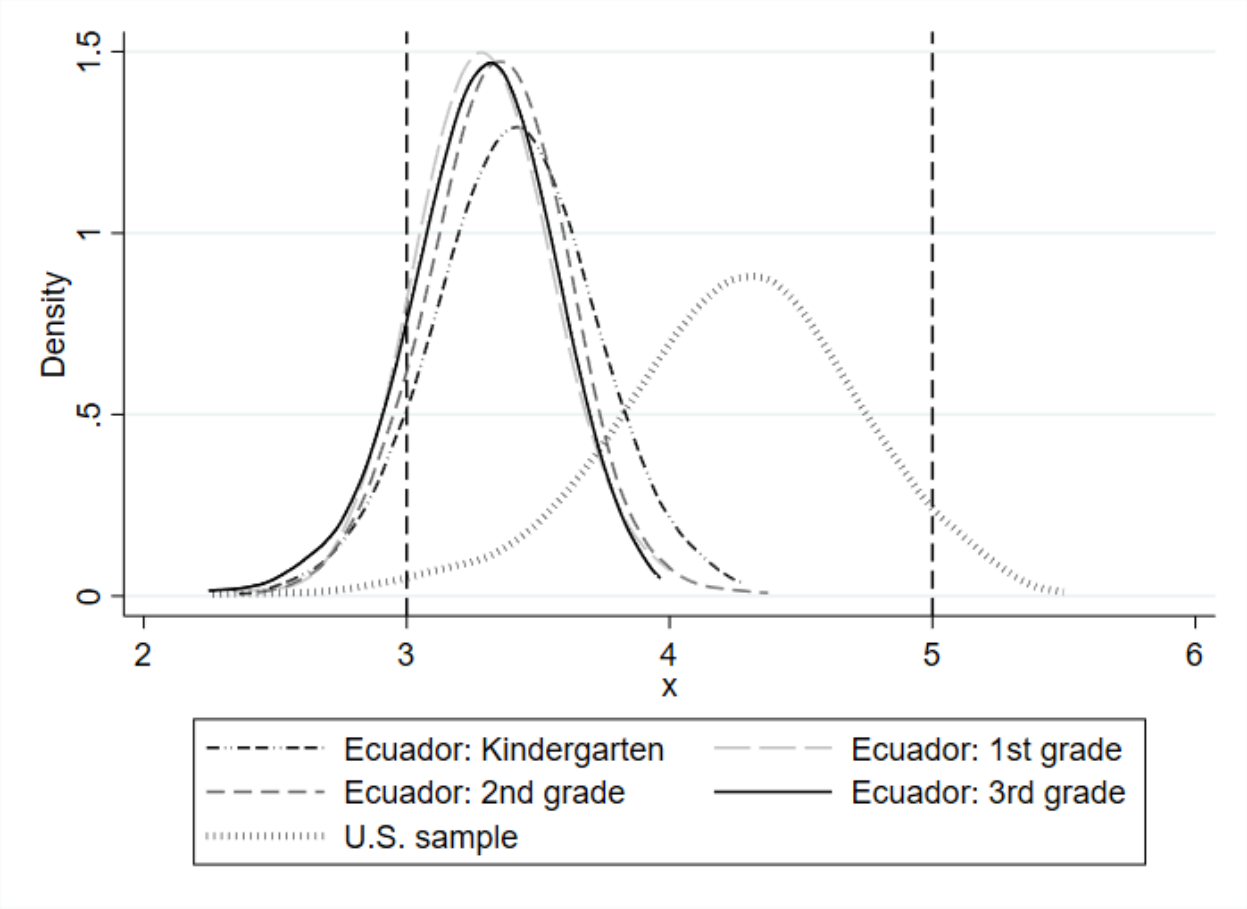
Van Goozen, Stephanie H.M., Peggy T. Cohen-Kettenis, Louis J.G. Gooren, Nico H. Frijda, and Nanne E. Van de Poll. 1994. "Activating Effects of Androgens on Cognitive Performance: Causal Evidence in a Group of Female-to- Male Transsexuals." *Neuropsychologia* 32(10): 1153–57.

Van Goozen, Stephanie H.M, Peggy T. Cohen-Kettenis, Louis J.G. Gooren, Nico H. Frijda, and Nanne E. Van de Poll. 1995. "Gender Differences in Behaviour: Activating Effects of Cross-Sex Hormones." *Psychoneuroendocrinology* 20(4): 343–63.

Wai, Jonathan, Megan Cacchio, Martha Putallaz, and Matthew C. Makel. 2010. "Sex Differences in the Right Tail of Cognitive Abilities: A 30-Year Examination." *Intelligence* 38: 412-23.

Whitmore, Diane. 2005. "Resource and Peer Impacts on Girls' Academic Achievement: Evidence from a Randomized Experiment." *American Economic Review* 95(2): 199-203.

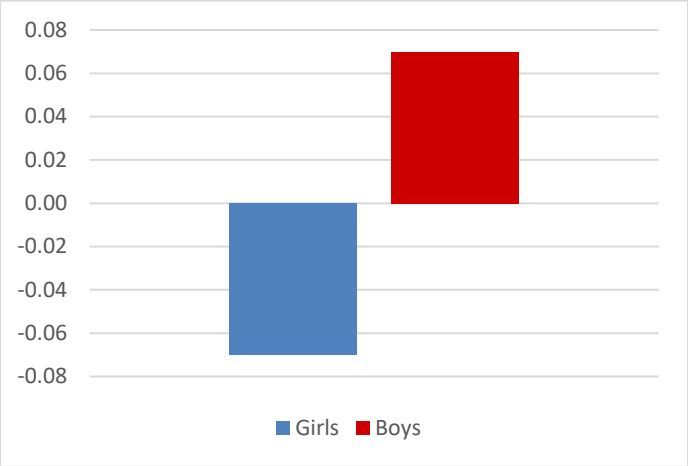
Figure 1: Distribution of CLASS scores, Ecuador and US data



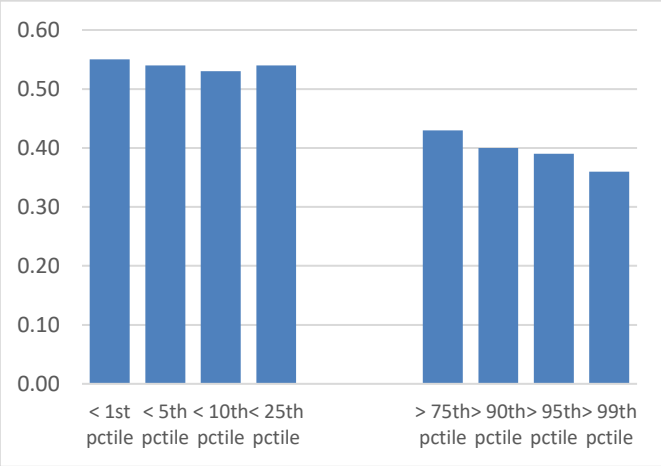
Note: The figure graphs univariate densities of the CLASS score of kindergarten, 1st grade, 2nd grade, and 3rd grade teachers in Ecuador, and in a nationally representative sample of kindergarten classrooms in the United States (Clifford et al. 2003). The CLASS is scored on a 1–7 scale; scores of 1–2 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.

Figure 2: Differences in Math Achievement, by Gender

Panel A: Mean differences



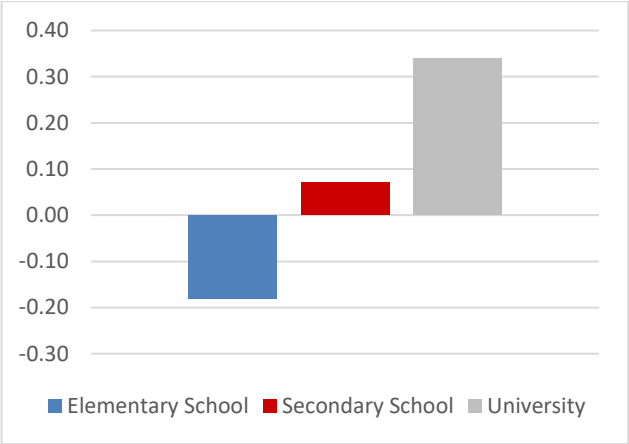
Panel B: Proportion female at different points in the distribution



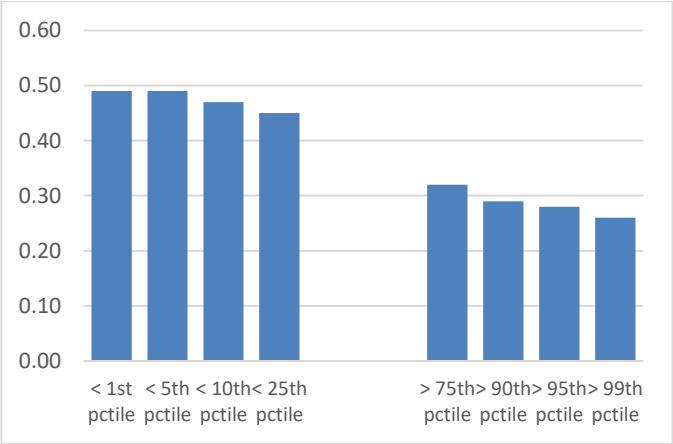
Note: Sample size is 9,013. The math test score is given by an equally-weighted sum of the individual scores on the three sections of the test in each grade: *number recognition and basic arithmetic*, *number sense*, and *word problems*. The questions within each section are aggregated by Item Response Theory (IRT).

Figure 3: Differences in Math Achievement, by Maternal Education

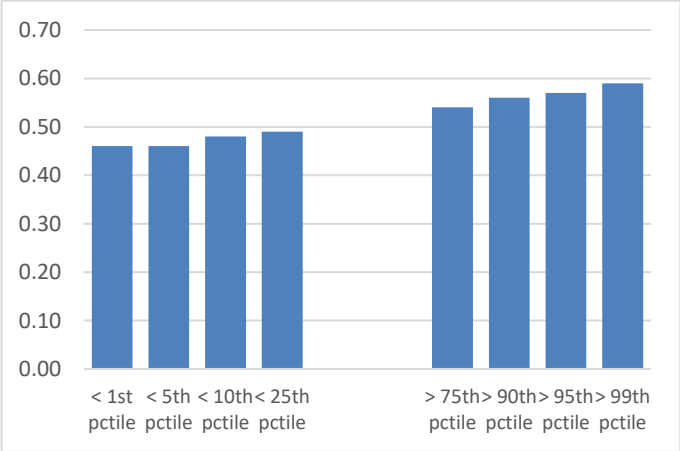
Panel A: Mean differences



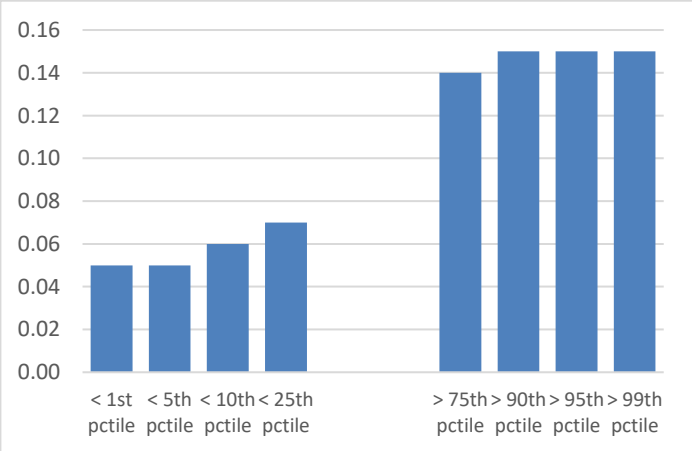
Panel B: Proportion of children of mothers with elementary school education at different points in the distribution



Panel C: Proportion of children of mothers with secondary school education at different points in the distribution



Panel D: Proportion of children of mothers with university education at different points in the distribution



Note: Sample size is 9,013. The math test score is given by an equally-weighted sum of the individual scores on the three sections of the test in each grade: *number recognition and basic arithmetic*, *number sense*, and *word problems*. The questions within each section are aggregated by Item Response Theory (IRT).

Table 1: Summary Statistics

	Mean	S.D.	Obs.
Panel A: Children			
Age (months)	60.22	4.62	9,013
Proportion female	0.50	0.50	9,013
TVIP	84.03	16.78	8,933
Mother's age	30.49	6.57	9,012
Father's age	34.74	7.91	7,110
Mother's years of schooling	8.89	3.76	9,013
Father's years of schooling	8.53	3.78	7,099
Panel B: Teachers			
Age	44.37	10.38	1,699
Proportion female	0.89	0.31	1,716
Experience	18.18	10.58	1,710
Proportion tenured	0.77	0.42	1,704
Class size	37.44	7.50	1,719

Note: 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 2: Mother's education and math achievement

	Test score: *Number recognition and basic arithmetic*	Test score: *Number sense*	Test score: *Word problems*	IRT, equal weights	Simple count, equal weights	IRT, factor aggregate	IRT, Anderson aggregate
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Bivariate associations between math achievement and gender							
Girls	-0.115	-0.162	-0.066	-0.143	-0.150	-0.157	-0.136
	(0.017)	(0.015)	(0.016)	(0.018)	(0.018)	(0.018)	(0.018)
Panel B: Multivariate associations between math achievement and gender							
Girls	-0.114	-0.162	-0.063	-0.141	-0.148	-0.156	-0.135
	(0.018)	(0.016)	(0.016)	(0.019)	(0.019)	(0.019)	(0.019)
Secondary school	0.236	0.227	0.192	0.271	0.277	0.273	0.266
	(0.018)	(0.016)	(0.016)	(0.019)	(0.019)	(0.019)	(0.019)
University	0.402	0.385	0.363	0.473	0.476	0.475	0.466
	(0.037)	(0.036)	(0.039)	(0.042)	(0.044)	(0.043)	(0.042)
University*girls	0.115	0.122	0.091	0.137	0.144	0.130	0.137
	(0.050)	(0.048)	(0.052)	(0.056)	(0.058)	(0.057)	(0.056)
F-test (p-value)	0.99	0.37	0.57	0.94	0.93	0.64	0.96

Note: Sample size is 36,052 in all regressions. All regressions include controls for child age in months and its square. F-test is test that the sum of the coefficients on girls and the interaction between girls and mothers with university education is zero. Standard errors clustered at the student level.

Table 3: Mother's education, wealth, and math achievement

	(1)	(2)	(3)	(4)
Girls	-0.141	-0.137	-0.133	-0.138
	(0.019)	(0.019)	(0.019)	(0.019)
Secondary school	0.271	0.170		0.201
	(0.019)	(0.021)		(0.020)
University	0.473	0.237		0.301
	(0.042)	(0.045)		(0.045)
Interaction: University*girls	0.137	0.132		0.143
	(0.056)	(0.056)		(0.060)
Wealth-Middle			0.240	0.164
			(0.019)	(0.020)
Wealth-Top			0.514	0.360
			(0.041)	(0.044)
Interaction: Wealth-Top*girls			0.042	-0.013
			(0.056)	(0.059)
Polynomial in Wealth	N	Y	N	N
F-test 1 (p-value)	0.94	0.94		0.94
F-test 2 (p-value)			0.09	0.01
Number of observations	36,052	36,020	36,020	36,020

Note: All regressions include child age in months and its square. F-test 1 is test that the sum of the coefficients on girls and the interaction between girls and the dummy for mothers with university education is zero. F-test 2 is test that the sum of the coefficients on girls and the interaction between girls and the dummy for top wealth is zero. Standard errors clustered at the student level.

Table 4: Married and cohabitating mothers: Parental education, wealth, and math achievement

	(1)	(2)	(3)	(4)	(5)	(6)
Girls	-0.147	-0.141	-0.144	-0.138	-0.140	-0.143
	(0.021)	(0.021)	(0.021)	(0.021)	(0.021)	(0.022)
Secondary school-mother	0.303	0.195				0.182
	(0.021)	(0.023)				(0.023)
University-Mother	0.501	0.250				0.228
	(0.049)	(0.052)				(0.056)
Interaction: University-mother*girls	0.105	0.098				0.100
	(0.066)	(0.065)				(0.073)
Secondary school-father			0.268	0.176		0.151
			(0.021)	(0.022)		(0.022)
University-father			0.515	0.289		0.271
			(0.053)	(0.055)		(0.058)
Interaction: University-father*girls			0.063	0.065		0.037
			(0.074)	(0.071)		(0.080)
Wealth-Middle					0.261	0.153
					(0.021)	(0.023)
Wealth-Top					0.567	0.333
					(0.047)	(0.053)
Interaction: Wealth-Top*girls					0.023	-0.028
					(0.064)	(0.069)
Polynomial in Wealth	N	Y	N	Y	N	N
F-test 1 (p-value)	0.50	0.48				0.56
F-test 2 (p-value)			0.25	0.29		0.18
F-test 3 (p-value)					0.05	0.01
Number of observations	28,480	28,456	28,384	28,360	28,456	28,360

Note: All regressions include child age in months and its square. F-test 1 is test that the sum of the coefficients on girls and the interaction between girls and mothers with university education is zero. F-test 2 is test that the sum of the coefficients on girls and the interaction between girls and fathers with university education is zero. F-test 3 is test that the sum of the coefficients on girls and top wealth is zero. Standard errors clustered at the student level.

Table 5: Single mothers: Education, wealth, and math achievement

	(1)	(2)	(3)	(4)
Girls	-0.114	-0.112	-0.089	-0.113
	(0.045)	(0.044)	(0.044)	(0.046)
Secondary school	0.114	0.041		0.068
	(0.047)	(0.051)		(0.049)
University	0.370	0.192		0.261
	(0.092)	(0.098)		(0.098)
Interaction: University*girls	0.239	0.247		0.251
	(0.119)	(0.118)		(0.123)
Wealth-Middle			0.157	0.108
			(0.045)	(0.047)
Wealth-Top			0.351	0.220
			(0.089)	(0.093)
Interaction: Wealth-Top*girls			0.067	0.006
			(0.126)	(0.129)
Polynomial in Wealth	N	Y	N	N
F-test 1 (p-value)	0.26	0.22		0.24
F-test 2 (p-value)			0.85	0.40
Number of observations	6236	6228	6228	6228

Note: All regressions include child age in months and its square. F-test 1 is test that the sum of the coefficients on girls and the interaction between girls and mothers with university education is zero. F-test 2 is test that the sum of the coefficients on girls and the interaction between girls and top wealth is zero. Standard errors clustered at the student level.

Table 6: Distributional effects: Mother's education and math achievement

	q05	q10	q25	q50	q75	q90	q95
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: All children (N=36,052)							
Dummy: girls	-0.063	-0.065	-0.097	-0.147	-0.189	-0.223	-0.201
	(0.026)	(0.025)	(0.022)	(0.021)	(0.021)	(0.027)	(0.029)
Dummy: Secondary school	0.231	0.245	0.253	0.271	0.269	0.270	0.260
	(0.03)	(0.027)	(0.022)	(0.022)	(0.021)	(0.027)	(0.029)
Dummy: University	0.487	0.466	0.476	0.484	0.471	0.435	0.384
	(0.067)	(0.057)	(0.054)	(0.057)	(0.045)	(0.063)	(0.054)
Interaction: University*girls	0.056	0.143	0.136	0.175	0.108	0.083	0.056
	(0.091)	(0.078)	(0.074)	(0.069)	(0.057)	(0.081)	(0.07)
F-test 1 (p-value)	0.77	0.26	0.72	0.69	0.10	0.05	0.02
F-test 2 (p-value)	0.00						
F-test 3 (p-value)	0.43						
F-test 4 (p-value)	0.76						
F-test 5 (p-value)	0.44						
Panel B: Children of single mothers (N=6,236)							
Dummy: girls	-0.124	-0.073	-0.119	-0.126	-0.117	-0.151	-0.168
	(0.065)	(0.07)	(0.055)	(0.053)	(0.052)	(0.059)	(0.071)
Dummy: Secondary school	0.138	0.150	0.161	0.165	0.066	0.072	0.042
	(0.059)	(0.063)	(0.051)	(0.056)	(0.056)	(0.061)	(0.074)
Dummy: University	0.461	0.482	0.361	0.368	0.314	0.360	0.346
	(0.093)	(0.085)	(0.127)	(0.123)	(0.127)	(0.141)	(0.15)
Interaction: University*girls	0.177	0.252	0.282	0.299	0.178	0.202	0.174
	(0.183)	(0.145)	(0.159)	(0.153)	(0.147)	(0.201)	(0.189)
F-test 1 (p-value)	0.98	0.28	0.28	0.19	0.70	0.70	0.98
F-test 2 (p-value)	0.38						
F-test 3 (p-value)	0.36						
F-test 4 (p-value)	0.39						
F-test 5 (p-value)	0.95						

Note: All regressions include child age in months and its square. F-test 1 is test that the sum of the coefficients on girls and the interaction between girls and mothers with university education is zero. F-test 2 is test that the coefficients on girls in quantiles 10 and 90 are the same. F-test 3 is test that the coefficients on secondary school in quantiles 10 and 90 are the same. F-test 4 is test that the coefficients on university in quantiles 10 and 90 are the same. F-test 5 is test that the coefficients on the interaction between girls and university education in quantiles 10 and 90 are the same. Standard errors are calculated with a block bootstrap, with blocks equal to children.

Table 7: Classroom effects, by grade and gender

	All grades	Kindergarten	1st grade	2nd grade	3rd grade
	(1)	(2)	(3)	(4)	(5)
All children	0.113	0.129	0.118	0.111	0.095
	(0.006)	(0.01)	(0.009)	(0.006)	(0.007)
Boys	0.151	0.170	0.161	0.145	0.130
	(0.008)	(0.013)	(0.011)	(0.01)	(0.009)
Girls	0.147	0.174	0.156	0.141	0.117
	(0.008)	(0.012)	(0.011)	(0.008)	(0.01)
Difference, boys-girls (confidence interval)	(-0.008,0.014)	(-0.027,0.02)	(-0.014,0.027)	(-0.014,0.021)	(-0.012,0.027)

Note: Classroom effects are given by the standard deviation of the distribution of the difference in the residualized test scores across classrooms within the same school, corrected for sampling error using an Empirical Bayes estimator, as described in the main text of the paper. Standard errors, in parentheses, are calculated with a block bootstrap, with blocks equal to schools.

Table 8: The effects of teachers on math achievement, by gender

Panel A: Teacher effects, kindergarten

All children	0.09
	(0.02)
Boys	0.07
	(0.03)
Girls	0.11
	(0.03)
Difference (p-value)	0.31

Panel B: Correlations in learning gains across cohorts, kindergarten, by gender

		Cohort2		
		All	Boys	Girls
Cohort 1	All	0.29	0.17	0.24
		(0.1)	(0.1)	(0.09)
	Boys	0.22	0.10	0.17
		(0.1)	(0.09)	(0.08)
	Girls	0.25	0.14	0.24
		(0.09)	(0.09)	(0.09)

Panel C: Effect of having a female teacher, 3rd grade, by gender

All children	0.013
	(0.025)
Boys	0.027
	(0.029)
Girls	-0.001
	(0.028)
Difference (p-value)	0.32

Panel D: CLASS scores and math achievement, by grade and gender

	All grades	Kindergarten	1st grade	2nd grade	3rd grade
All children	0.043	0.054	0.060	0.032	0.023
	(0.006)	(0.013)	(0.012)	(0.014)	(0.01)
Boys	0.038	0.036	0.060	0.024	0.030
	(0.008)	(0.015)	(0.015)	(0.015)	(0.013)
Girls	0.048	0.073	0.060	0.040	0.014
	(0.007)	(0.016)	(0.014)	(0.015)	(0.011)
F-test 1 (p-value)	0.22	0.03	0.97	0.25	0.20
F-test 2 (p-value)	0.06				
F-test 3 (p-value)	0.31				
F-test 4 (p-value)	0.01				

Note: Panel A: Teacher effects are given by the square root of the covariance of the classroom effects for the two kindergarten cohorts. Sample size is 11,943. Standard errors, in parentheses, are calculated with a block bootstrap, with blocks equal to schools. The p-value in the last row corresponds to a test that the effects for boys and girls are the same. Panel B: cross-cohort correlations between the estimated learning gains a given teacher produces for all children, for boys only, and for girls only, and the bootstrapped standard errors of these correlations, in parentheses. Sample size is 11,943. Panel C: Regressions of 3rd grade math achievement on indicator variable for female teachers. All regressions include a fourth-order polynomial in lagged test scores, child age and its square, and school fixed effects. Standard errors clustered at classroom level. Sample size is 9,013. The p-value in the last row corresponds to a test that the effects for boys and girls are the same. Panel D: Regressions of math achievement on teacher's CLASS score. All regressions include a fourth-order polynomial in lagged test scores, child age and its square, and school fixed effects. Standard errors clustered at classroom level. F-test 1 is test that coefficients on CLASS for boys and girls are the same. F-tests 2, 3, and 4 are tests that coefficients on CLASS are the same across grades for all children, boys, and girls, respectively. Sample sizes are 33,600 in the all grades regression, and 8,415 in the grade-specific regressions.

Table 9: Peer Effects

	All grades (1)	Kindergarten (2)	1st grade (3)	2nd grade (4)	3rd grade (5)
Panel A: Effect of peers' lagged test scores on own math achievement					
All children	0.008	0.053	-0.073	0.035	-0.027
	(0.033)	(0.073)	(0.069)	(0.06)	(0.051)
Boys	0.013	0.045	-0.037	0.027	-0.019
	(0.036)	(0.077)	(0.074)	(0.065)	(0.054)
Girls	0.002	0.061	-0.109	0.044	-0.035
	(0.036)	(0.08)	(0.072)	(0.064)	(0.056)
F-test 1 (p-value)	0.65	0.78	0.11	0.71	0.68
F-test 2 (p-value)	0.50				
F-test 3 (p-value)	0.82				
F-test 4 (p-value)	0.29				
Panel B: Effect of proportion of peers who are girls on own math achievement					
All children	0.152	-0.039	0.217	0.305	0.228
	(0.098)	(0.158)	(0.143)	(0.412)	(0.261)
Boys	0.097	-0.109	0.221	0.232	0.171
	(0.107)	(0.177)	(0.167)	(0.423)	(0.271)
Girls	0.208	0.035	0.213	0.387	0.287
	(0.112)	(0.193)	(0.171)	(0.425)	(0.28)
F-test 1 (p-value)	0.25	0.44	0.96	0.42	0.50
F-test 2 (p-value)	0.59				
F-test 3 (p-value)	0.54				
F-test 4 (p-value)	0.79				

Note: Panel A: Regressions of math achievement on lagged peer test scores (leave-*i*-out means). All regressions include a fourth-order polynomial in own lagged test scores, child age and its square, and school fixed effects. Standard errors clustered at classroom level. F-test 1 is test that coefficients on mean lagged test scores of peers for boys and girls are the same. F-tests 2, 3, and 4 are tests that coefficients on mean lagged test scores of peers are the same across grades for all children, boys, and girls, respectively. Sample sizes are 36,052 in the all grades regression, and 9,013 in each of the grade-specific regressions. Panel B: Regressions of math achievement on proportion of peers who are girls. All regressions include a fourth-order polynomial in own lagged test scores, child age and its square, and school fixed effects. Standard errors clustered at classroom level. F-test 1 is test that coefficients on proportion of peers who are girls is the same in the regressions for boys and girls. F-tests 2, 3, and 4 are tests that coefficients on proportion of peers who are girls are the same is the same across grades for all children, boys, and girls, respectively. Sample sizes are 36,052 in the all grades regression, and 9,013 in each of the grade-specific regressions.

Appendix A: Randomization checks

Randomization checks: A major strength of our paper is that, in all four grades, children were assigned to classrooms within schools by a rule that we argue is as-good-as-random. Moreover, compliance with the assignment rule was very high: 98.6 percent in kindergarten, 99.8 percent in 1st grade, 98.5 percent in 2nd grade, and 98.7 percent in 3rd grade.[34]

The as-good-as-random assignment rules we implemented were the following: In kindergarten, all children in a school were ordered by their last name and first name, and were then assigned to teachers in alternating order; in 1st grade, they were ordered by their date of birth, from oldest to youngest, and were then assigned to teachers in alternating order; in 2nd grade, they were divided by gender, ordered by their first name and last name, and then assigned in alternating order; and in 3rd grade, they were divided by gender and then randomly assigned to one or another classroom.

We argue that these four instances of as-good-as-random assignment of children to teachers produce four exogenous, orthogonal shocks to teacher quality. This has two testable implications. First, the predetermined characteristics of teachers should be uncorrelated with the characteristics of children in their classrooms. Second, when teacher quality is defined as quality relative to the school mean, the quality of the teacher a child is assigned to in one grade should be orthogonal with the quality of the teacher she is assigned to in the other three grades; similarly, when peer quality is defined as quality relative to the school mean, the quality of the peers a child is assigned to in one grade should be orthogonal with the quality of the peers she is assigned to in the other three grades. We provide strong evidence that these implications of as-good-as-random assignment hold in our data.

Table A1 presents the results of regressions of child characteristics (age, gender, and the lagged test scores) on three predetermined teacher characteristics (years of experience teaching, whether a teacher is tenured and, for 1st through 3rd grades, teacher gender).[35] We report the coefficients from 33 separate regressions. All of the coefficients are small in magnitude. For example, teachers with 10 more years of experience have students who are between -0.1 and 0.05 months older, on average; tenured teachers have students who are between -0.575 and 0.014 months older on average; female teachers have students whose lagged test scores are between -0.059 and 0.028 standard deviations higher, on average. None of these coefficients are significant once we adjust for multiple hypothesis testing using the step-down procedure proposed by Romano and Wolf (2005).[36]

---

[34] Compliance was measured on the basis of two unannounced school visits, one in the middle of the school year, and another at the end of the school year (when children were tested). A child is taken not to be complying with the rule-based assignment if she was found to be sitting in a classroom other than the one she had been assigned to in *either* one of these two visits. When we analyze the effects of teacher or peer quality on math achievement, we include non-complying 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).

[35] Ninety-nine percent of kindergarten teachers are women, so we cannot test whether children of different characteristics are more or less likely to be assigned to female teachers in kindergarten.

[36] We have carried out Romano-Wolf tests in three different ways. In one, we jointly test the hypothesis in all of the tables in this Appendix (57 hypotheses); in another, we jointly test the hypotheses in each table—33 in Table A1, 12 each in tables A2 and A3; and in yet another we jointly test the hypotheses in each panel of a table—in this case, the number of hypotheses tested ranges from 6 to 9, depending on the panel. No matter which of these tests we carry out, we can

Next, we report the correlations of within-school, cross-classroom differences in the quality of teachers a child was assigned to in all four grades. These results are in Panels A and B of Table A2, corresponding to the cross-grade correlations in classroom effects and the CLASS scores, respectively. Once again, the correlations are very small in magnitude: In the case of the classroom effects, they range from -0.011 to 0.013, and in the case of the CLASS they range from -0.009 to 0.009. None of these correlations are significant after correcting for multiple hypothesis testing. Finally, in Appendix Table A3, we report the correlations of within-school, cross-classroom differences in the quality of peers a child was assigned to in kindergarten, 1st, 2nd, and 3rd grades. Here too we use two measures of peer quality, peers' lagged test scores and the proportion of girls in a classroom who are girls. Panel A shows that the correlations in lagged peers' test scores range from -0.005 to 0.020, and panel B shows that the correlations in the proportion of girls in the classroom range from -0.011 to 0.022. One again, none of these correlations are significant after adjusting for multiple hypothesis testing.

References

Romano, Joseph P., and Michael Wolf. 2005. "Stepwise Multiple Testing as Formalized Data Snooping." *Econometrica* 73(4): 1237-82.

never reject the null that all of the 57 coefficients in Tables A1 through A3 are insignificant at the 10 percent level or higher.

Table A1: Randomization checks: Correlations between child characteristics and predetermined teacher characteristics, by grade

	Mean	Gender	Age	Lagged score
		Kindergarten		
Years of experience	14.65	-0.001	-0.001	0.001
		(0.001)	(0.006)	(0.001)
Teacher is tenured	0.623	-0.002	0.014	0.017
		(0.009)	(0.085)	(0.02)
Teacher is female	0.989			
		1st grade		
Years of experience	18.94	0.000	-0.001	0.001
		(0.000)	(0.003)	(0.001)
Teacher is tenured	0.709	0.018	-0.086	-0.006
		(0.01)	(0.092)	(0.022)
Teacher is female	0.934	-0.032	0.078	-0.005
		(0.021)	(0.092)	(0.032)
		2nd grade		
Years of experience	20.34	0.000	-0.01	-0.002
		(0.000)	(0.005)	(0.001)
Teacher is tenured	0.880	-0.006	-0.575	-0.074
		(0.004)	(0.177)	(0.029)
Teacher is female	0.873	-0.004	0.064	0.028
		(0.004)	(0.177)	(0.028)
		3rd grade		
Years of experience	17.86	0.000	0.005	0.000
		(0.000)	(0.006)	(0.001)
Teacher is tenured	0.822	-0.008	-0.340	0.001
		(0.005)	(0.205)	(0.023)
Teacher is female	0.779	0.001	0.103	-0.059
		(0.005)	(0.153)	(0.022)

Note: All regressions include school fixed effects. Standard errors, in parentheses, clustered at classroom level.

Table A2: Correlations in teacher quality

	Kindergarten	First	Second	Third
		Panel A: Classroom Effects		
Kindergarten	1.000	-0.011	-0.008	0.005
	(0.000)	(0.010)	(0.009)	(0.010)
First		1.000	0.013	-0.008
		(0.000)	(0.010)	(0.010)
Second			1.000	-0.009
			(0.000)	(0.010)
Third				1.000
				(0.000)
		Panel B: Teacher's CLASS score		
Kindergarten	1.000	-0.009	0.002	0.006
	(0.000)	(0.008)	(0.009)	(0.009)
First		1.000	-0.004	-0.007
		(0.000)	(0.008)	(0.010)
Second			1.000	0.009
			(0.000)	(0.008)
Third				1.000
				(0.000)

Note: In Panel B, school-by-grade averages are removed before calculating correlations. Standard errors, in parentheses, are calculated with a block bootstrap, with blocks equal to schools.

Table A3: Correlations in peer quality

	Kindergarten	First	Second	Third
		Panel A: Peers' lagged test score		
Kindergarten	1.000	0.019	0.010	-0.005
	(0.000)	(0.010)	(0.009)	(0.009)
First		1.000	0.008	0.020
		(0.000)	(0.010)	(0.009)
Second			1.000	-0.005
			(0.000)	(0.009)
Third				1.000
				(0.000)
		Panel B: Proportion of girls		
Kindergarten	1.000	0.016	-0.001	-0.011
	(0.000)	(0.008)	(0.008)	(0.007)
First		1.000	0.002	0.022
		(0.000)	(0.008)	(0.008)
Second			1.000	0.003
			(0.000)	(0.010)
Third				1.000
				(0.000)

Note: In both panels, school-by-grade averages are removed before calculating correlations. Standard errors, in parentheses, are calculated with a block bootstrap, with blocks equal to schools.

Appendix B: Changes in the composition of the sample

The estimates we report in the paper are based on a balanced panel of children for whom we have the following data: (i) beginning-of-kindergarten TVIP scores. We impose this restriction because, in our analysis of classroom, teacher, and peer effects in kindergarten, we follow our earlier work (Araujo et al. 2016) and condition on lagged test scores. (In 1st, 2nd, and 3rd grades, we control for math achievement at the end of the previous grade;) (ii) data on maternal education. We impose this restriction so we can analyze any "protective" effects of maternal education on the math achievement of girls, relative to boys; and (iii) math test scores at the end of kindergarten, 1st grade, 2nd grade, and 3rd grade.

Working with a balanced panel ensures that any differences in results across tables or columns within a table are not driven by changes in the composition of the sample. Nevertheless, working with a balanced panel comes at a cost, as it means that the results we report are not based on the full sample of children in a given grade.

It is useful to distinguish between three possible reasons why a child who is attending a given grade could be missing from the sample: (i) new arrivals; (ii) leavers; and (iii) children missing data. We discuss each of these below.[37]

New arrivals: These are children who enter the sample at some point after the beginning of kindergarten.

- New arrivals in kindergarten: Children for whom we have end-of-kindergarten test scores, but are missing baseline TVIP.
- New arrivals in 1st grade: Children for whom we have 1st grade test scores, but are missing *all* of the following: (i) baseline TVIP; (ii) data on maternal education and wealth; (iii) kindergarten test scores.
- New arrivals in 2nd grade: Children for whom we have 2nd grade test scores, but are missing *all* of the following: (i) baseline TVIP; (ii) data on maternal education and wealth; (iii) kindergarten test scores; (iv) 1st grade test scores.
- New arrivals in 3rd grade: Children for whom we have 3rd grade test scores, but are missing *all* of the following: (i) baseline TVIP; (ii) data on maternal education and wealth; (iii) kindergarten test scores; (iv) 1st grade test scores; (v) 2nd grade test scores.

As a first step, we quantify the number of new arrivals, as a proportion of the total number of children in that grade. Note that, because we are interested in the characteristics of children who are in the classroom, but are not included in the balanced panel, the definitions we use for new arrivals in this analysis are *cumulative*: Specifically:

- Kindergarten: Children who are in kindergarten classrooms, but are not included in the balanced panel because they are "new arrivals", are children who arrived over the course of kindergarten. By this definition, 645 kindergarten children are excluded from the balanced panel sample, out of

[37] Note that, as defined, these categories are not mutually exclusive. For example, the same child could be a late arrival—say, she entered the sample in 1st grade—as well as a leaver—she left the sample after the end of 2nd grade.

a total of 14,521 children for whom we have data on end-of-kindergarten test scores (4.4 percent of total).

- 1st grade: Children who are in 1st grade classrooms, but are not included in the balanced panel because they are "new arrivals", include new arrivals in kindergarten as well as new arrivals in 1st grade. By this definition, 4,335 children are excluded from the balanced panel sample, out of a total of 16,432 1st grade children (26.4 percent of total).
- 2nd grade: Children who are in 2nd grade classrooms, but are not included in the balanced panel because they are "new arrivals", include new arrivals in kindergarten, 1st grade, or 2nd grade. By this definition, 5,708 children are excluded from the balanced panel sample, out of a total of 16,949 2nd grade children (33.7 percent of total).
- 3rd grade: Children who are in 3rd grade classrooms, but are not included in the balanced panel because they are "new arrivals", include new arrivals in kindergarten, 1st grade, 2nd grade, or 3rd grade. By this definition, 7,035 children are excluded from the balanced panel sample, out of a total of 17,280 3rd grade children (40.7 percent of total).

Next, to have a better sense of what children are missing from the balanced panel because they are new arrivals, we run regressions of three characteristics—child age, gender, and test scores—as a function of an indicator variable for new arrivals, separately by grade.[38] These regressions include school fixed effects; standard errors are clustered at the classroom level. Results are in Table B1, Panel A.

The main findings are:

- Compared to other children in their classrooms, new arrivals are older. In 1st grade, new arrivals are on average 3.4 months older; in 2nd grade, new arrivals are on average 4 months older; and in 3rd grade, new arrivals are on average 4.1 months older than other children in their classrooms.
- Compared to other children in their classrooms, new arrivals are more likely to be boys. In 1st grade, new arrivals are on average 2.0 percentage points more likely to be boys; in 2nd grade, new arrivals are on average 3.8 percentage points more likely to be boys; and in 3rd grade, new arrivals are on average 3.6 percentage points more likely to be boys.
- New arrivals have lower test scores at the end of the grade than other children in that grade. In 1st grade, new arrivals have test scores that are 0.056 standard deviations lower; in 2nd grade, new arrivals have test scores that are 0.075 standard deviations lower; and in 3rd grade, new arrivals have test scores that are 0.091 standard deviations lower than their classmates.

We return to a discussion of how these changes could affect the main results in the paper at the end of the Appendix.

[38] To test whether new arrivals are smarter or less smart than other children, we use all subsequent test scores. Take, for example, new arrivals in 1st grade. These children, by definition, have 1st grade test data. Unless they are also leavers or are missing data, they will also have 2nd and 3rd grade test data. For these children, as well as their 1st grade classmates, we generate a variable that is the average of 1st, 2nd, and 3rd grade test scores (unless they are missing 2nd or 3rd grade scores, in which the average is over a smaller number of scores). We then run a regression of this test score average as a function of the indicator variable for new arrivals. Results are very similar, however, if we only use the test scores for the grade in question (in the example above, 1st grade test scores).

Leavers: These are children who left our sample of schools at some point after the end of kindergarten.

- Leavers after kindergarten: These are children for whom we have kindergarten test scores, but we are missing *all* of the following: (i) 1st grade test scores; (ii) 2nd grade test scores; (iii) 3rd grade test scores.
- Leavers after 1st grade: These are children for whom we have 1st grade test scores, but we are missing *all* of the following: (i) 2nd grade test scores; (ii) 3rd grade test scores.
- Leavers after 2nd grade: These are children for whom we have test scores at the end of 2nd grade, but we are missing 3rd grade test scores.
- Leavers after 3rd grade: We do not have leavers after 3rd grade, as this is the last year with test score data.

As with new arrivals, we quantify the number of leavers, as a proportion of the total number of children in that grade. Once again, because we are interested in the characteristics of children who are in the classroom, but are not included in the balanced panel, the definitions we use for leavers in this analysis are *cumulative*: Specifically:

- Kindergarten: Children who are in kindergarten classrooms, but are not included in the balanced panel because they are "leavers", include children who left after kindergarten, after 1st grade, or after 2nd grade. By this definition, 4,042 children are excluded from the balanced panel sample, out of a total of 14,521 kindergarten children (30.7 percent of total).
- 1st grade: Children who are in 1st grade classrooms, but are not included in the balanced panel because they are "leavers", include children who left after 1st grade or after 2nd grade. By this definition, 3,241 children are excluded from the balanced panel sample, out of a total of 16,432 1st grade children (19.7 percent of total).
- 2nd grade: Children who are in 2nd grade classrooms, but are not included in the balanced panel because they are "leavers", are children who left after 2nd grade. By this definition, 2,070 children are excluded from the balanced panel sample, out of a total of 16,949 2nd grade children (12.2 percent of total).
- 3rd grade: We do not have leavers after 3rd grade, as this is the last year with test score data.

Here too, to have a better sense of what children are missing from the balanced panel because they are leavers we run regressions of three characteristics—child age, gender, and test scores—as a function of an indicator variable for leavers, separately by grade.[39] These regressions include school fixed effects; standard errors are clustered at the classroom level. Results are in Table B1, Panel B.

[39] To test whether leavers are smarter or less smart than other children, we use all previous test scores. Take, for example, leavers after 1st grade. These children, by definition, have 1st grade test data. Unless they are late arrivals or are missing data, they will also have baseline TVIP and kindergarten test data. To test whether children who left our sample after 1st grade are smarter or less smart than other 1st grade children, we generate a variable that is the average of baseline TVIP, kindergarten, and 1st grade test scores (unless they are missing the baseline TVIP or kindergarten test scores, in which the average is over a smaller number of scores). We then run a regression of this test score average as a function of the indicator variable for leavers. Results are very similar, however, if we only use the test scores for the grade in question (in the example above, 1st grade test scores).

The main findings are:

- Compared to other children in their classrooms, leavers are older. In 1st grade, leavers are on average 0.9 months older; in 2nd grade, leavers are on average 1.6 months older than other children in their classrooms.
- Compared to other children in their classrooms, leavers are more likely to be boys. In kindergarten, leavers are 3.5 percentage points more likely to be boys; in 1st grade, leavers are 3.4 percentage points more likely to be boys than other children in their classrooms.
- At the end of the grade that preceded their departure, leavers had lower test scores than other children in their classrooms, and the differences are large in magnitude. Relative to other kindergarten children, leavers have test scores that are 0.220 standard deviations lower; relative to other 1st grade children, leavers have test scores that are 0.328 standard deviations lower; and relative to other 2nd grade children, leavers have test scores that are 0.262 standard deviations lower

We return to a discussion of how these changes could affect the main results in the paper at the end of the Appendix.

Children with missing data: These are children who are missing data, which we identify by observing an irregular sequence: for example, children who have test scores at the end of kindergarten and 2nd grade, but are missing test scores at the end of 1st grade.

- Missing kindergarten data: Children who have baseline TVIP, but have only *one* of the following: (i) data on wealth and maternal education; (ii) end-of-kindergarten test scores.
- Missing 1st grade data: These are children with kindergarten and 2nd grade test scores, but missing 1st grade test scores.
- Missing 2nd grade data: These are children with 1st and 3rd grade test scores, but missing 2nd grade test scores.

As with new arrivals and leavers, we quantify the number of children with missing data, as a proportion of the total number of children in that grade:

- Kindergarten: There are 924 children, out of a total of 14,521 kindergarten children, who are missing data (6.4 percent of total).
- 1st grade: There are 787 children, out of a total of 16,432 1st grade children, who are missing data (4.7 percent of total).
- 2nd grade: There are 752 children, out of a total of 16,949 2nd grade children, who are missing data (4.4 percent of total).
- 3rd grade: There are 688 children, out of a total of 17,280 3rd grade children, who are missing data (3.9 percent of total).

Here too, to have a better sense of who is missing data, we run regressions of three characteristics—child age, gender, and test scores—as a function of an indicator variable for children with missing data,

separately by grade.[40] These regressions include school fixed effects; standard errors are clustered at the classroom level. Results are in Table B1, Panel C.

The broad picture is as follows:

- There is no clear pattern indicating that children with missing data are consistently older or younger than other children. Kindergarten children missing data are a little older (1.08 months, on average), while 2nd and 3rd grade children missing data are a little younger (0.73 months and 0.98 months, respectively, on average).
- Boys and girls are equally likely to be missing data.
- Children who are missing data have lower test scores. Significant effects range from -0.056 standard deviations in 3rd grade, to -0.81 standard deviations in 1st grade.

Conclusion: Broadly speaking, then, children who are in a grade but are excluded from the balanced panel that is the basis for our estimates are older, are more likely to be boys, and have lower test scores than other children.

To see how excluding these children from the balanced panel sample could affect the conclusions of our paper we reproduce Tables 2 through 9 in the paper, but carry out the calculations for the largest possible sample (as opposed to the balanced panel sample). We now discuss each of these tables in turn.

Table B2 (which corresponds to Table 2 in the paper):

- Panel A: Panel A shows that the boy-girl gap in math achievement in the largest possible sample is somewhat smaller than in the balanced panel sample. For example, in our preferred specification (IRT, with equal weights for each of the three test sections), girls have test scores that are 0.114 (0.012) standard deviations lower than boys in the largest possible sample, compared to 0.143 (0.018) in the balanced panel sample.
- Panel B: The most important result in Panel B is the magnitude of the interaction term between maternal university education and girls, relative to the main effect for girls. In the largest possible sample, the coefficient on girls is -0.105 (0.016), the coefficient on the interaction term is 0.098 (0.047), and the p-value on the F-test that the sum of the two coefficients is zero is 0.87. In the balanced panel sample, the coefficient on girls is -0.141 (0.019), the coefficient on the interaction term is 0.137 (0.056), and the p-value on the F-test that the sum of the two coefficients is zero is 0.94.

We conclude from this comparison that the two main results from this table—that girls have significantly lower math test scores than boys, but that among children of mothers with university

[40] To test whether children with missing data are smarter or less smart than other children in their grade, we use all available test scores. Take, for example, a 1st grade child with missing data. For these calculations, we take all 1st grade children and average all of the available test score data. We then run a regression of this test score average as a function of the indicator variable for children who are missing data. Results are very similar, however, if we only use the test scores for the grade in question (in the example above, 1st grade test scores).

education there is no difference in the math achievement of boys and girls—are apparent in both the balanced panel sample and the largest possible sample.

Table B3 (which corresponds to Table 3 in the paper): The main message from this table is that controlling for wealth in various ways does not change the magnitude of the interaction term between maternal university education and girls. This result is very similar in the largest possible sample and the balanced panel sample.

Table B4 (which corresponds to Table 4 in the paper): The main message from this table is that, among married or cohabitating women, the coefficient on the interaction term between paternal university education and girls is smaller than the corresponding coefficient on the interaction term between maternal university education and girls, and declines when all the measures of socioeconomic status and the interaction terms are included in the regression. This result is apparent in both samples—if anything, the "protective" role of paternal university education on the math achievement of girls, relative to boys, is even smaller in the largest possible sample than in the balanced panel sample. A second result is that the coefficient on the interaction term between maternal university education and girls is not significant at conventional levels, but changes very little with more controls for socioeconomic status. This result is apparent in both samples.

Table B5 (which corresponds to Table 5 in the paper): The main results from this table are two. First, that the girl disadvantage in achievement is smaller among single mothers than among married or cohabitating mothers; second, that the coefficient on the interaction term between maternal university education and girls is large, relative to the main effect for girls. Both results are apparent in both samples.

Table B6 (which corresponds to Table 6 in the paper):

- Panel A: The main messages from Panel A are two. First, that the girl disadvantage in achievement is largest at the top of the distribution—the coefficient on the main effect for girls increases in absolute value as you move from left to right in the table; second, that at the 75th percentile and higher, boys outperform girls even among children of university-educated women. Both results are apparent in both samples.
- Panel B: The main message of Panel B is that, among children of single mothers with university education, there is no difference in the math achievement of boys and girls even in the right tail of the distribution. This is apparent in Table 6 in the main body of the paper, and less clearly in Table B6, although the coefficient on the interaction term between maternal university education and girls at the 90th percentile seems to be an outlier relative to the coefficient on the interaction term at other percentiles of the distribution.

Table B7 (which corresponds to Table 7 in the paper): The main messages of this table are two. First, the magnitude of the classroom effects declines monotonically by grade; second, we cannot reject the null that the classroom effect for boys and girls is the same. Both results are apparent in the two samples.

Table B8 (which corresponds to Table 8 in the paper):

- Panel C: The main message of this panel is that neither girls nor boys benefit more from having a female than a male teacher; this is apparent in both samples.
- Panel D: The main messages from this panel are two. First, that the association between the CLASS and math achievement is larger in kindergarten and 1st grade than in 2nd and 3rd grades; this is apparent in both samples. Second, that the association between a higher teacher CLASS score and math achievement is the same for boys and girls; this is apparent in both samples—if anything, it is clearer in the largest possible sample than in the balanced panel sample.

Table B9 (which corresponds to Table 9 in the paper):

- Panel A: The main message from this panel is that the lagged test scores of peers do not affect own math achievement of boys or girls; this result is apparent in both samples.
- Panel B: The main message from this panel in the main body of the paper is that having more female classmates increases own math achievement of girls—the effect is modest, and only borderline significant (coefficient is 0.208, with a standard error of 0.112). In the largest possible sample, the coefficient is smaller, and the effect is no longer significant (coefficient is 0.100, with a standard error of 0.099).

Conclusions: Our overall conclusions from this extensive analysis on the possible effects of new arrivals, leavers, and children with missing data are as follows:

1. We are excluding from our main analysis sample children who are older, are more likely to be boys, and have lower test scores than other children.
2. In the largest possible sample of children, the gap in math achievement between boys and girls is somewhat smaller than in the balanced panel sample—0.114, rather than 0.143 standard deviations.
3. The "protective" effects of maternal university education on the math achievement of girls, relative to boys, which is particularly apparent among single mothers, is found in both samples.
4. In both samples, we find no evidence that classroom or teacher quality matters more for boys than girls, or vice-versa.
5. The lagged test scores of peers do not predict own math achievement of boys or girls in either sample. In the balanced panel sample, but not in the largest possible sample, the proportion of classmates who are girls has a modest, and borderline significant, effect on the math achievement of girls.

We conclude that, in practice, working with a balanced panel sample, rather than the largest possible sample in each grade, does not substantively affect the results in the paper.

Table B1: Analysis of new arrivals, leavers, and children with missing data

	Panel A: New arrivals			Panel B: Leavers			Panel C: Children with missing data		
	Age	Gender	Score	Age	Gender	Score	Age	Gender	Score
Kindergarten	0.179(0.321)	0.010(0.021)	-0.015(0.044)	-0.027(0.108)	-0.035(0.010)	-0.220(0.018)	1.087(0.206)	-0.007(0.018)	-0.062(0.028)
1st grade	3.362(0.201)	-0.020(0.009)	-0.056(0.019)	0.929(0.160)	-0.034(0.011)	-0.328(0.020)	-0.23(0.207)	0.003(0.018)	-0.081(0.030)
2nd grade	4.020(0.198)	-0.038(0.008)	-0.075(0.017)	1.597(0.260)	-0.015(0.013)	-0.262(0.025)	-0.73(0.232)	-0.004(0.019)	-0.067(0.031)
3rd grade	4.144(0.180)	-0.036(0.008)	-0.091(0.017)				-0.98(0.236)	0.011(0.020)	-0.056(0.032)

Note: Panel A: regressions of indicator variable for new arrivals on child age, gender, and end-of-grade test scores, separately by grade (12 separate regressions). Panel B: regressions of indicator variable for leavers on child age, gender, and end-of-grade test scores in the previous grade, separately by grade (9 separate regressions). Panel C: regressions of indicator variable for children with missing data on child age, gender, and end-of-grade test scores, separately by grade (12 separate regressions). All regressions include school fixed effects. Standard errors are clustered at the classroom level.

Table B2: Mother's education and math achievement

	Test score: *Number recognition and basic arithmetic*	Test score: *Number sense*	Test score: *Word problems*	IRT, equal weights	Simple count, equal weights	IRT, factor aggregate	IRT, Anderson aggregate
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Bivariate associations between math achievement and gender							
Girls	-0.095***	-0.142***	-0.049***	-0.114***	-0.125***	-0.124***	-0.109***
	(0.012)	(0.011)	(0.011)	(0.012)	(0.013)	(0.013)	(0.012)
Panel B: Multivariate associations between math achievement and gender							
Girls	-0.088***	-0.134***	-0.044***	-0.105***	-0.116***	-0.116***	-0.100***
	(0.015)	(0.014)	(0.014)	(0.016)	(0.016)	(0.016)	(0.016)
Secondary school	0.267***	0.246***	0.208***	0.289***	0.294***	0.292***	0.284***
	(0.016)	(0.014)	(0.014)	(0.016)	(0.016)	(0.016)	(0.016)
University	0.473***	0.437***	0.386***	0.519***	0.523***	0.525***	0.510***
	(0.033)	(0.032)	(0.034)	(0.036)	(0.037)	(0.036)	(0.035)
University*girls	0.082*	0.076*	0.087*	0.098**	0.106**	0.086*	0.101**
	(0.043)	(0.042)	(0.045)	(0.047)	(0.049)	(0.047)	(0.047)
F-test (p-value)	0.895	0.139	0.318	0.868	0.830	0.495	0.993

Note: Sample size is 65,169 in all regressions for Panel A. Sample size is 46,332 in all regressions for Panel B. All regressions include controls for child age in months and its square. F-test is test that the sum of the coefficients on girls and the interaction between girls and mothers with university education is zero. Standard errors clustered at the student level.

Table B3: Mother's education, wealth, and math achievement

	(1)	(2)	(3)	(4)
Girls	-0.105***	-0.102***	-0.103***	-0.103***
	(0.016)	(0.016)	(0.015)	(0.016)
Secondary school	0.289***	0.188***		0.218***
	(0.016)	(0.017)		(0.017)
University	0.519***	0.276***		0.340***
	(0.036)	(0.038)		(0.039)
Interaction: University*girl	0.098**	0.098**		0.099**
	(0.047)	(0.046)		(0.050)
Wealth-Middle			0.235***	0.153***
			(0.016)	(0.017)
Wealth-Top			0.543***	0.363***
			(0.034)	(0.037)
Interaction: Wealth-Top*girls			0.036	0.007
			(0.046)	(0.050)
Polynomial in Wealth	N	Y	N	N
F-test 1 (p-value)	0.87	0.94		0.93
F-test 2 (p-value)			0.13	0.05
Number of observations	46,332	46,275	48,710	46,275

Note: All regressions include child age in months and its square. F-test 1 is test that the sum of the coefficients on girls and the interaction between girls and the dummy for mothers with university education is zero. F-test 2 is test that the sum of the coefficients on girls and the interaction between girls and the dummy for top wealth is zero. Standard errors clustered at the student level.

Table B4: Married and cohabitating mothers: Maternal education, wealth, and math achievement

	(1)	(2)	(3)	(4)	(5)	(6)
Girls	-0.115***	-0.111***	-0.109***	-0.104***	-0.108***	-0.111***
	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)
Secondary school-Mother	0.316***	0.210***				0.199***
	(0.018)	(0.019)				(0.020)
University-Mother	0.528***	0.268***				0.247***
	(0.042)	(0.045)				(0.048)
Interaction: University -Mother*girl	0.089	0.087				0.096
	(0.056)	(0.055)				(0.062)
Secondary school-Father			0.264***	0.168***		0.140***
			(0.018)	(0.019)		(0.019)
University-Father			0.527***	0.290***		0.270***
			(0.045)	(0.046)		(0.049)
Interaction: University-Father*girl			0.023	0.026		-0.006
			(0.061)	(0.060)		(0.066)
Wealth-Middle					0.252***	0.139***
					(0.018)	(0.019)
Wealth-Top					0.584***	0.334***
					(0.040)	(0.045)
Interaction: Wealth-Top*girls					0.026	-0.011
					(0.054)	(0.059)
Polynomial in Wealth	N	Y	N	Y	N	N
F-test 1 (p-value)	0.62	0.64				0.81
F-test 2 (p-value)			0.14	0.17		0.08
F-test 3 (p-value)					0.11	0.04
Number of Observations	35,843	35,801	35,749	35,708	35,857	35,667

Note: All regressions include child age in months and its square. F-test 1 is test that the sum of the coefficients on girls and the interaction between girls and mothers with university education is zero. F-test 2 is test that the sum of the coefficients on girls and the interaction between girls and fathers with university education is zero. F-test 3 is test that the sum of the coefficients on girls and top wealth is zero. Standard errors clustered at the student level.

Table B5: Single mothers: Education, wealth, and math achievement

	(1)	(2)	(3)	(4)
Girls	-0.067*	-0.067*	-0.057	-0.068*
	(0.036)	(0.035)	(0.035)	(0.036)
Secondary school	0.176***	0.102**		0.126***
	(0.037)	(0.040)		(0.039)
University	0.491***	0.312***		0.375***
	(0.074)	(0.079)		(0.080)
Interaction: University*girls	0.122	0.132		0.121
	(0.096)	(0.095)		(0.101)
Wealth-Middle			0.182***	0.107***
			(0.036)	(0.038)
Wealth-Top			0.418***	0.232***
			(0.072)	(0.076)
Interaction: Wealth-Top*girls			0.081	0.046
			(0.103)	(0.108)
Polynomial in Wealth	N	Y	N	N
F-test 1 (p-value)	0.54	0.46		0.59
F-test 2 (p-value)			0.81	0.83
Number of observations	8,525	8,510	8,510	8,510

Note: All regressions include child age in months and its square. F-test 1 is test that the sum of the coefficients on girls and the interaction between girls and mothers with university education is zero. F-test 2 is test that the sum of the coefficients on girls and the interaction between girls and top wealth is zero. Standard errors clustered at the student level.

Table B6: Distributional effects: Mother's education and math achievement

	q05	q10	q25	q50	q75	q90	q95
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: All children (N=46,632)							
Girls	-0.005	-0.024	-0.055***	-0.113***	-0.157***	-0.200***	-0.173***
	(0.021)	(0.021)	(0.018)	(0.017)	(0.018)	(0.022)	(0.026)
Secondary school	0.284***	0.265***	0.282***	0.285***	0.281***	0.271***	0.258***
	(0.023)	(0.023)	(0.018)	(0.018)	(0.018)	(0.021)	(0.025)
University	0.507***	0.518***	0.538***	0.533***	0.498***	0.442***	0.379***
	(0.06)	(0.048)	(0.048)	(0.045)	(0.042)	(0.051)	(0.051)
Interaction: University*girls	0.078	0.110*	0.107*	0.117**	0.078	0.069	0.070
	(0.076)	(0.064)	(0.061)	(0.056)	(0.05)	(0.064)	(0.064)
F-test 1 (p-value)	0.29	0.15	0.31	0.97	0.07	0.03	0.09
F-test 2 (p-value)	0.00						
F-test 3 (p-value)	0.91						
F-test 4 (p-value)	0.21						
F-test 5 (p-value)	0.63						
Panel B: Children of single mothers (N=8,525)							
Girls	0.007	0.000	-0.029	-0.078**	-0.095**	-0.166***	-0.172***
	(0.045)	(0.045)	(0.042)	(0.039)	(0.043)	(0.049)	(0.058)
Secondary school	0.191***	0.150***	0.212***	0.210***	0.125***	0.121**	0.092
	(0.043)	(0.045)	(0.042)	(0.041)	(0.047)	(0.049)	(0.056)
University	0.534***	0.516***	0.507***	0.490***	0.415***	0.419***	0.389***
	(0.07)	(0.083)	(0.092)	(0.089)	(0.095)	(0.121)	(0.11)
Interaction: University*girls	0.110	0.169	0.155	0.184	0.086	0.035	0.142
	(0.131)	(0.12)	(0.12)	(0.114)	(0.115)	(0.166)	(0.148)
F-test 1 (p-value)	0.34	0.21	0.17	0.30	0.94	0.08	0.78
F-test 2 (p-value)	0.00						
F-test 3 (p-value)	0.82						
F-test 4 (p-value)	0.61						
F-test 5 (p-value)	0.43						

Note: All regressions include child age in months and its square. F-test 1 is test that the sum of the coefficients on girls and the interaction between girls and mothers with university education is zero. F-test 2 is test that the coefficients on girls in quantiles 10 and 90 are the same. F-test 3 is test that the coefficients on secondary school in quantiles 10 and 90 are the same. F-test 4 is test that the coefficients on university in quantiles 10 and 90 are the same. F-test 5 is test that the coefficients on the interaction between girls and university education in quantiles 10 and 90 are the same. Standard errors are calculated with a block bootstrap, with blocks equal to children.

Table B7: Classroom effects, by grade and gender

	All grades	Kindergarten	1st grade	2nd grade	3rd grade
	(1)	(2)	(3)	(4)	(5)
All children	0.101	0.114	0.100	0.097	0.091
	(0.009)	(0.011)	(0.012)	(0.008)	(0.009)
Boys	0.134	0.149	0.139	0.127	0.120
	(0.01)	(0.012)	(0.012)	(0.012)	(0.009)
Girls	0.130	0.151	0.135	0.124	0.110
	(0.01)	(0.014)	(0.011)	(0.01)	(0.01)
Difference, boys-girls (confidence interval)	(-0.01,0.01)	(-0.02,0.01)	(-0.01,0.02)	(-0.01,0.02)	(-0.01,0.02)

Note: Classroom effects are given by the standard deviation of the distribution of the difference in the residualized test scores across classrooms within the same school, corrected for sampling error using an Empirical Bayes estimator, as described in the main text of the paper. Standard errors, in parentheses, are calculated with a block bootstrap, with blocks equal to schools.

Table B8: The effects of teachers on math achievement, by gender

Panel C: Effect of having a female teacher, 3rd grade, by gender

All children	0.020
	(0.023)
Boys	0.036
	(0.027)
Girls	0.005
	(0.026)
Difference (p-value)	0.231

Panel D: CLASS scores and math achievement, by grade and gender

	All grades	Kindergarten	1st grade	2nd grade	3rd grade
All children	0.045***	0.060***	0.051***	0.034***	0.025***
	(0.006)	(0.011)	(0.01)	(0.012)	(0.01)
Boys	0.045***	0.054***	0.053***	0.030**	0.033***
	(0.007)	(0.013)	(0.012)	(0.013)	(0.012)
Girls	0.045***	0.067***	0.049***	0.039***	0.018*
	(0.007)	(0.014)	(0.012)	(0.013)	(0.01)
F-test 1 (p-value)	0.96	0.38	0.74	0.48	0.21
F-test 2 (p-value)	0.13				
F-test 3 (p-value)	0.51				
F-test 4 (p-value)	0.02				

Note: Panel C: Regressions of 3rd grade math achievement on indicator variable for female teachers. All regressions include a fourth-order polynomial in lagged test scores, child age and its square, and school fixed effects. Standard errors clustered at classroom level. Sample size is 9,965. The p-value in the last row corresponds to a test that the effects for boys and girls are the same. Panel D: Regressions of math achievement on teacher's CLASS score. All regressions include a fourth-order polynomial in lagged test scores, child age and its square, and school fixed effects. Standard errors clustered at classroom level. F-test 1 is test that coefficients on CLASS for boys and girls are the same. F-tests 2, 3, and 4 are tests that coefficients on CLASS are the same across grades for all children, boys, and girls, respectively. Sample sizes are 45,569 in the all grades regression.

Table B9: Peers effects

	All grades (1)	Kindergarten (2)	1st grade (3)	2nd grade (4)	3rd grade (5)
Panel A: Effect of peers' lagged test scores on own math achievement					
All children	0.004	0.027	-0.033	0.022	-0.016
	(0.029)	(0.059)	(0.058)	(0.05)	(0.048)
Boys	0.016	0.024	-0.005	0.027	-0.009
	(0.031)	(0.062)	(0.061)	(0.054)	(0.051)
Girls	-0.008	0.030	-0.060	0.017	-0.023
	(0.031)	(0.064)	(0.061)	(0.054)	(0.052)
F-test 1 (p-value)	0.26	0.90	0.14	0.80	0.71
F-test 2 (p-value)	0.83				
F-test 3 (p-value)	0.95				
F-test 4 (p-value)	0.70				
Panel B: Effect of proportion of peers who are girls on own math achievement					
All children	0.075	-0.041	0.09	0.285	0.193
	(0.085)	(0.134)	(0.117)	(0.338)	(0.235)
Boys	0.052	-0.049	0.125	0.211	0.118
	(0.093)	(0.149)	(0.141)	(0.346)	(0.244)
Girls	0.10	-0.033	0.052	0.362	0.269
	(0.099)	(0.167)	(0.141)	(0.353)	(0.254)
F-test 1 (p-value)	0.60	0.92	0.64	0.37	0.35
F-test 2 (p-value)	0.69				
F-test 3 (p-value)	0.79				
F-test 4 (p-value)	0.62				

Note: Panel A: Regressions of math achievement on lagged peer test scores (leave-*i*-out means). All regressions include a fourth-order polynomial in own lagged test scores, child age and its square, and school fixed effects. Standard errors clustered at classroom level. F-test 1 is test that coefficients on mean lagged test scores of peers for boys and girls are the same. F-tests 2, 3, and 4 are tests that coefficients on mean lagged test scores of peers are the same across grades for all children, boys, and girls, respectively. Sample sizes are 36,052 in the all grades regression, and 9,013 in each of the grade-specific regressions. Panel B: Regressions of math achievement on proportion of peers who are girls. All regressions include a fourth-order polynomial in own lagged test scores, child age and its square, and school fixed effects. Standard errors clustered at classroom level. F-test 1 is test that coefficients on proportion of peers who are girls is the same in the regressions for boys and girls. F-tests 2, 3, and 4 are tests that coefficients on proportion of peers who are girls are the same is the same across grades for all children, boys, and girls, respectively. Sample size is 45,639 in the all grades regression, and 13,087, 11,730, 10,857, 9,965 in each of the grade-specific regressions.

Appendix C: Math tests

This Appendix briefly describes the math tests we applied in each grade. In kindergarten through 2nd grade, tests were applied to a child individually (a single enumerator with a single child). The enumerator would ask a child one question at a time, before moving on to the next question. Most tests also had a "stopping rule": When a child made three mistakes in a row, or stated that she did not know the answer to three questions in a row, that section of the test was stopped. In 3rd grade, tests were applied in a classroom testing setting, with children taking the tests at the same time. There were no stopping rules for specific questions, although there were time limits for different sections of the test.

Description of tests: In each grade, the test had three sections: (1) *number recognition and basic arithmetic*; (2) *number sense*; and (3) *word problems*.

Domain 1: *Number recognition and basic arithmetic*: In kindergarten and 1st grade, children were asked to identify numbers. In 1st, 2nd, and 3rd grades, children were given 16-18 addition problems (mainly single-digit additions in 1st grade, double-digit additions in 2nd grade, double- and triple-digit additions in 3rd grade), and were given 90 seconds (in 1st grade) or 3 minutes (in 2nd and 3rd grade) to solve as many as they could. Also in 1st, 2nd, and 3rd grades, children were given 9-12 subtraction problems (mainly single-digit subtractions in 1st grade, double-digit subtractions in 2nd grade, double- and triple-digit subtractions in 3rd grade) and were given 90 seconds (in 1st grade) or 3 minutes (in 2nd grade) to solve them. Finally, in 3rd grade children were given 3 minutes to solve 20 simple multiplication problems.

Domain 2: *Number sense*: In kindergarten, 2nd and 3rd grades, children were shown sequences of numbers; in each case, one number was missing, and children were asked to name the missing number. In 1st through 3nd grades, we used the number line test developed by Siegler and his coauthors (Siegler and Booth 2004; Siegler and Opfer 2003).[41] In 2nd and 3rd grades, children were also tested on place value—see examples in Table B1 below.

Domain 3: *Word problems*: In this domain, children were given simple word problems, some of which had pictures as visual aids.

Table B1 gives an example of an "easy" question (a question answered correctly by approximately 75 percent of children) and a "hard" question (a question answered correctly by approximately 25 percent of children) for each domain and grade.[42]

References

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

[41] In the 1st grade version of this test, children were shown a number line with a value of 1 on the left, 20 on the right, and no other numbers. They were then asked to place a randomly generated number between 2 and 19 in the appropriate spot on the number line. The 2nd and 3rd grade versions of the test were similar, with a number line from 1 to 50. In both 1st, 2nd and 3rd grade, each child carried out this exercise 5 times, and different children got different random numbers.

[42] In general, there is no single question that is answered correctly by exactly 25 percent or 75 percent of children. In these cases, we pick the question that is answered correctly by the proportion of children that is closest to these values (25 or 75 percent).

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

Siegler, Robert S., and John E. Opfer. 2003. "The Development of Numerical Estimation: Evidence for Multiple Representations of Numerical Quantity." *Psychological Science* 14(3): 237-43.

Table C1: Sample questions

Panel A: Number recognition and basic arithmetic	
Kindergarten	
Easy question: *number recognition*	Hard question: *number recognition*
Child is shown a sheet with 5 numbers (2, 9, 4, 6, and 10). Enumerator points at the number 6, and asks: "What number is this?"	Child is shown a sheet with 5 numbers (17, 14, 33, 58, 187). Enumerator points at the number 17, and asks: "What number is this?"
1st grade	
Easy question: *number recognition*	Hard question: *number recognition*
Child is shown a sheet with 5 numbers (17, 14, 33, 58, 187). Enumerator points at the number 17, and asks: "What number is this?"	Child is shown a sheet with 4 numbers (94, 200, 105, 513). Enumerator points at the number 105, and asks: "What number is this?"
Easy question: *basic arithmetic* 7+1	Hard question: *basic arithmetic* 11-1
2nd grade	
Easy question: *basic arithmetic* 20+10	Hard question: *basic arithmetic* 492+213
3rd grade	
Easy question: *basic arithmetic* 14+3	Hard question: *basic arithmetic* 56*100
Panel B: Number sense	
Kindergarten	
Easy question: *number sequences* Child is shown the sequence 3, 4, --, 6, and is asked to name the missing number.	Hard question: *number sequences* Child is shown the sequence 10, 11, 12, -- and is asked to name the missing number.
1st grade: see description in text of number line test	
2nd grade: also, see description in text of number line test	
Easy question: *number sequences* Child is shown the sequence 133, ---, 135, 136, and is asked to name the missing number.	Hard question: *number sequences* Child is shown the sequence 530, 532, --, 536 and is asked to name the missing number.
Easy question: *place value* Child is shown a page with the number 6 on the left, the number 6 on the right, and is then asked whether the appropriate sign that should be placed between them is >, =, or <	Hard question: *place value* Child is shown a page with the equation 386 < -- < 521, and is asked to choose the right answer from the following 4 options: (a) 297; (b) 334; (c) 410; and (d) 528.
3rd grade: also, see description in text of number line test	
Easy question: *number sequences* Child is shown the sequence 0, 2, 4, ---, and is asked to name the missing number.	Hard question: *number sequences* Child is shown the sequence 3342, 3341, 3340, --, and is asked to name the missing number.
Easy question: *place value*	Hard question: *place value*

Child is shown a page with the number 5 on the left, the number 10 on the right, and is then asked whether the appropriate sign that should be placed between them is >, =, or <	Child is shown 400 + 40 + 8, and is asked to choose the right answer from the following 4 options: (a) 4408; (b)448; (c) 400408; and (d) 4048.
Panel C: Word problems	
Kindergarten	
Easy question: *word problems*	Hard question: *word problems*
Child is shown a picture with 3 apples and 6 apple cores, and is then asked the following question: "How many apples have not been eaten?"	Child is shown picture with 6 cookies and is asked the following question: "If José eats 3 cookies, how many cookies are left?"
1st grade	
Easy question: *word problems*	Hard question: *word problems*
Child is shown a picture with two dogs and three numbers (2, 3, 5) and is asked to point at the number that corresponds to the number of dogs in the picture.	Child is shown a picture with 2 red circles, and is asked the following question: "If you draw 2 more circles, how many circles will there be in total?"
2nd grade	
Easy question: *word problems*	Hard question: *word problems*
Child is shown a picture with 7 crayons and is asked: "If you take away 3 crayons how many are left?"	Child is asked the following question: "Pablo bought two boxes of chewing gum for 50 cents. He had 10 pieces of gum in total and he gave half to his brother. How many pieces of gum does Pablo have left?"
3rd grade	
Easy question: *word problems*	Hard question: *word problems*
Child is shown a picture with 7 crayons and is asked: "If you take away 3 crayons how many are left?"	Child is asked the following question: "If a ladybug has 6 spots. How many spots would 10 ladybugs have?"

Appendix D: Application of the CLASS in Ecuador

In this paper, we measure the quality of teacher-child interactions using the Classroom Assessment Scoring System (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 D1 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.[43]
>
> *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'

[43] 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).[44]

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.[45]"

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 D2.[46] 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 a sample of schools in six US cities (Dallas, Charlotte-Mecklenburg, Hillsborough, Memphis, New York and Denver) 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). Because the scores on the different CLASS dimensions are highly correlated, we focus on a teacher's *total* CLASS score (given by the simple average of her score on the 10 dimensions). We take this score to be a measure of Responsive Teaching (as in Hamre et al. 2014).

To apply the CLASS in Ecuador, we filmed all kindergarten teachers for a full school day (from approximately eight in the morning until one in the afternoon). In accordance with CLASS protocols, we then discarded the first hour of film (when teachers and students are more likely to be aware of, and

[44] 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).
[45] 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).
[46] These and other results in this Appendix refer to kindergarten teachers. However, results for 1st, 2nd, and 3rd grade are qualitatively very similar.

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 per grade. 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 very small. In kindergarten, for example, the mean score is 3.35 in the fourth segment, compared to 3.48 in the first segment. This 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 D3. 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. (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.

Table D1: CLASS scores for Behavior Management dimension

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

Source: Pianta et al. (2007).

Table D2: Pairwise correlation of CLASS dimensions, kindergarten

		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

Note: Table shows the Pairwise Correlation Coefficient 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% confidence level.

Table D3: Sources of measurement error in the CLASS, kindergarten teachers

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