Carneiro, Pedro M.; Cruz-Aguayo, Yyannú; Schady, Norbert Rüdiger

Working Paper

Where the girls are not: Households, teachers, and the gender gap in early math achievement

IDB Working Paper Series, No. IDB-WP-807

Provided in Cooperation with:
Inter-American Development Bank (IDB), Washington, DC

Suggested Citation: Carneiro, Pedro M.; Cruz-Aguayo, Yyannú; Schady, Norbert Rüdiger (2017) : Where the girls are not: Households, teachers, and the gender gap in early math achievement, IDB Working Paper Series, No. IDB-WP-807, Inter-American Development Bank (IDB), Washington, DC, https://doi.org/10.18235/0000700

This Version is available at:
https://hdl.handle.net/10419/173871

IDB WORKING PAPER SERIES Nº IDB-WP-807

Where the Girls Are Not: Households, Teachers, and the Gender Gap in Early Math Achievement

Pedro Carneiro
Yyannú Cruz-Aguayo
Norbert Schady

Inter-American Development Bank
Social Sector - SCL/SCL

April, 2017

IDB
Inter-American Development Bank

Where the Girls Are Not: Households, Teachers, and the Gender Gap in Early Math Achievement

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

*University College London
**Inter-American Development Bank

April, 2017

Cataloging-in-Publication data provided by the
Inter-American Development Bank
Felipe Herrera Library
Carneiro, Pedro.
Where the girls are not: households, teachers, and the gender gap in early math achievement / Pedro Carneiro, Yyannú Cruz Aguayo, Norbert Schady.
p. cm. — (IDB Working Paper Series ; 806)
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-806

http://www.iadb.org

Where the Girls Are Not: Households, Teachers, and the Gender Gap in Early Math Achievement

Pedro Carneiro
Yyannú Cruz-Aguayo
Norbert Schady

April 24, 2017[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 (RES-589-28-0001) and the ERC through grant ERC-2015-CoG-682349. We thank Jere Behrman, Gregory Elacqua, Costas Meghir, and Andrew Morrison for their comments, Nicola Dehnen, 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. We find 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, which suggests that maternal education specifically, and home environments generally, are important. There is no evidence that teacher quality affects the boy-girl differences in early test scores, regardless of whether we measure quality on the basis of classroom effects, teacher effects, or the observed interactions between teachers and children.

JEL classification: I21

Keywords: test scores, teachers, gender

1. Introduction

Early math test scores are highly predictive of 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). Early math achievement has been shown to be more strongly associated with the likelihood of high school completion and college attendance than early reading, or the incidence of antisocial behavior, inattention, or anxiety. Children who score in the lowest quartile of the distribution of math achievement at ages 6, 8, and 10 years are 13 percentage points less likely to graduate from high school, and 34 percentage points less likely to attend college (Duncan and Magnuson 2011).

Widespread deficits in early math achievement, and low levels of math achievement in adulthood among the poor, may also have negative effects on aggregate productivity and growth. Using data for a large number of (mainly) developed countries, Hanushek and Kimko (2000) and Hanushek and Woessman (2009) estimate that countries that have one standard deviation higher test scores in math and science have annual GDP growth rates that are 2 percentage points higher.

A growing body of research in psychology and economics has sought to identify the determinants of math achievement at young ages. This research, mainly from the United States, has established two stylized facts. First, there are steep socioeconomic gradients in early math test scores. For example, using the National Longitudinal Survey of Youth (NLSY), Carneiro and Heckman (2003) report that, at age 6 years, there is a difference of 15 percentile score points in math achievement between children in the lowest and highest income quartiles; by age 12 years this gap has increased to almost 25 points. Using the Kindergarten Cohort of the Early Childhood Longitudinal Study (ECLS-K), Duncan and Magnuson (2011) estimate a difference of 1.34 standard deviations in 1st grade math test scores between children in the top and bottom quintiles of socioeconomic status (SES); this difference increases only modestly, to 1.38 standard deviations, by 5th grade.

The second stylized fact is that there are substantial differences in math achievement between boys and girls (with higher scores among boys). The average score of girls on the math component of the Scholastic Aptitude Test (SAT) is 30-40 points below that of boys, and there has

been no substantive change in this gender gap for almost 50 years (Perry 2016). The boy-girl gap in math achievement is largest in the right tail of the distribution.[2]

Most research on gender differences in math achievement has focused on high school and college, and on the dearth of women in technical fields like engineering and the hard sciences. Research on boy-girl differences in math achievement at early ages is less conclusive. 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. In 5th grade, there are almost three times as many boys as girls in the top 5 percent of the distribution of math achievement. On the other hand, in their thoughtful review of the literature in psychology, Halpern et al. (2007) argue that differences between boys and girls in early math achievement are small (and sometimes favor girls) or nonexistent (see Robinson et al. 1996 and Spelke 2005 for opposing results).

Much less is known about socioeconomic gradients and gender differences in math achievement in developing countries. Data from the Programme for International Student Assessment (PISA), a test applied to 15-year olds, show that boys outperform girls in most developing (and developed) countries, with particularly large differences favoring boys in Latin America (OECD 2014).[3] Results from a regional test applied to 6th graders in Sub-Saharan Africa indicate that boys have higher scores than girls in 12 out of 15 countries (Bethell 2016).[4] In Peru, boys outperform girls in math in a nationwide test of 2nd graders (Berlinski and Schady 2015). In a careful study with data from Chile, Bharadwaj et al. (2012) show that, in 4th grade, boys have math test scores that are 0.09 standard deviations higher than girls; by 8th grade, these gender differences have increased substantially, to 0.21 standard deviations.

In this paper, we study early math achievement in Ecuador, a middle-income country in South America. Our estimates are based on unique data from a cohort of approximately 10,400 children who were assigned to kindergarten teachers within schools with a rule that is as-good-as-

[2] 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.

[3] Eight Latin American countries (Argentina, Brazil, Chile, Colombia, Costa Rica, Mexico, Peru, and Uruguay) participated in the 2012 PISA. In all eight, boys outperformed girls. Colombia, Chile, and Costa Rica had the largest, third-largest, and fourth-largest boy-girl differences, respectively, of the 64 countries participating in PISA. In some developing countries, including in Jordan, Qatar, Malaysia, Thailand, Singapore and Bulgaria, girls outperformed boys.

[4] Participating countries included Botswana, Kenya, Lesotho, Malawi, Mauritius, Mozambique, Namibia, Seychelles, South Africa, Swaziland, Tanzania, Uganda, Zanzibar, and Zimbabwe.

random. These children were then re-assigned to different teachers in 1st grade, and once again re-assigned to teachers in 2nd grade.[5,6] Compliance with the as-good-as random assignment was very high: 98.2 percent, in kindergarten, 99.6 percent in 1st grade, and 98.0 percent in 2nd grade.[7] At the end of each grade, age-appropriate math tests were applied to children. In addition, all teachers were filmed teaching for (at least) one day, and the video was coded using a rubric known as the Classroom Assessment Scoring System (CLASS, Pianta et al. 2007). The CLASS is a measure of the quality of the interactions between teachers and students.

The fact that we have rich, longitudinal data on children and their teachers, and the as-good-as-random assignment of children to teachers in three consecutive grades, allows us to make several important contributions to the literature on early math achievement. Our first set of results focuses on the relationship between SES, gender, and math scores. Like others, we find that children in low-SES households have substantially worse math test scores: 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 a primary school education. We also show that girls have substantially lower math scores than boys, 0.13 standard deviations on average.

We next look at possible interactions between gender and SES. Research from the United States has generally found that boys are more sensitive than girls to environmental influences, including home environments and neighborhood quality (Autor et al. 2016a; Bertrand and Pan

[5] 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 first divided by gender, ordered by their first name and last name, and then assigned in alternating order. We present a variety of tests showing that these rules are as good as random assignment in Appendix A.

[6] To the best of our knowledge, ours is the first research project that has assigned children to teachers with a random or as-good-as-random rule in three consecutive years. In Project STAR, children were randomly assigned to teachers between kindergarten and 3rd grade. However, once children had been assigned, they were meant to stay with the same teachers and peers for all four grades (Krueger 1999; Chetty et al. 2011). In our case, there were three rounds of as-good-as-random assignment—to kindergarten, 1st grade, and 2nd grade. Thus, there are three exogenous, orthogonal shocks to classroom quality for each child. This feature of our data allows us to separate any effects of the classroom environment on math achievement in kindergarten, 1st grade, and 2nd grade.

[7] These numbers refer to compliance for children in the balanced panel that we use for estimation throughout the paper, as discussed below. However, compliance rates for the largest possible sample of children in each grade are very similar. 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 look at the effects of teacher 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).

2013; Chetty et al. 2016). A corollary is that SES gradients in a variety of outcomes in adolescence and adulthood tend to be steeper for boys than girls. We show that precisely the opposite pattern occurs in our data: Among girls, the difference in math achievement between children of mothers with university education and those with only primary education is 0.58 standard deviations, while the comparable difference for boys is 0.47 standard deviations. Moreover, we find that boys of mothers with university education do *not* outperform girls of similarly-educated mothers, on average. We are agnostic about the extent to which the "protective" effect of maternal university education on the math achievement of girls is correlational or causal—in all likelihood, it is some of both (Carneiro et al. 2013).

Having established the basic facts about SES, gender, and test scores, we turn to an analysis of how teachers affect the boy-girl gap in math achievement. Our most novel results test for differences in the returns to teacher quality by gender.[8] The data we use, in particular the as-good-as-random assignment of children to classrooms within schools in three subsequent grades, with very high rates of compliance, are particularly well-suited for this purpose.

We build on our earlier work on teacher quality in kindergarten (Araujo et al. 2016) and, more generally, on the large literature on teachers in the United States.[9] We begin by estimating "classroom effects" in kindergarten, 1st grade, and 2nd grade, separately for boys and girls. These classroom effects measure the extent to which there was 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 never statistically different from each other.

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. These teacher effects are somewhat larger for girls than boys, suggesting that, if anything, girls benefit more from better kindergarten teachers in our sample than boys. However, as with the classroom effects, we cannot reject the null hypothesis that the teacher effects for boys and girls are the same.

[8] The literature from the United States has found mixed evidence on the relative benefits of school quality for boys and girls. Using data from lottery assignment in the Charlotte-Mecklenburg school district, Hastings et al. (2006) and Deming et al. (2014) argue that attending a first-choice school benefits girls more than boys. On the other hand, Autor et al. (2016b) use data from Florida to document larger benefits from school quality for boys than girls.

[9] 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.

Finally, we turn to the data on the quality of teacher-child interactions, as measured by the CLASS. In our earlier work (Araujo et al. 2016), we showed that children who were randomly assigned to a teacher with a one standard deviation higher CLASS score had 0.08-0.09 higher math test scores at the end of kindergarten. In this paper, we first show that the CLASS is more strongly associated with learning outcomes in kindergarten than in 1st grade, and more strongly in 1st grade than in 2nd grade. However, we find no evidence that the association between teacher CLASS scores and math achievement is significantly larger (or smaller) for boys than girls in any grade.

In sum, gender differences in the responsiveness to teacher or classroom quality do not appear to contribute in an important way to the boy advantage in early math test scores in our data. It is possible that there are behaviors, uniform across all teachers in Ecuador, that are biased against girls, and our data are not well-suited to study these.[10] However, the fact that the boy-girl gap is large for children of mothers with primary school education, but not for children of mothers with university education, suggests to us that differences in the home environment are likely to be important (even though we cannot establish how much of the association between maternal university education and the absence of significant boy-girl differences in test scores is causal, as discussed above).

The rest of the paper proceeds as follows. In section 2, we describe the setting and our data. Section 3 discusses identification and presents our main results. We conclude in Section 4.

2. 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 202 schools in the 2012 school year with an assignment rule that is as-good-as-random.[11] These children were reassigned to 1st grade teachers in 2013, and to

[10] For example, teachers may call more on boys than girls in class, or give different feedback to boys and girls (as argued by Sadker and Sadker 1995 for the United States, and Bassi et al. 2016 for a sample of 4th grade students in Chile). It is also possible that we underestimate the effect of schools on the gender gap in math achievement, either because we ignore cross-school variation in classroom and teacher quality, or because schools affect the gender gap through other channels, not involving teachers or classrooms.

[11] 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.

2nd grade teachers in 2014. Compliance with our assignment rule was very high—98.6 percent on average.[12] As a result, for the main cohort of children we follow, we have three exogenous, orthogonal shocks to classroom quality. In addition, a second cohort of kindergarten children was assigned to kindergarten 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 the possible purposeful placement of students to teachers that can arise in a non-experimental setting (Chetty et al. 2014a; Rothstein 2010). We provide further details on the assignment rules, compliance, and changes in the sample that arise as children transfer into or out of our sample of schools in Appendix A.

At the end of kindergarten, 1st grade, and 2nd grade, we applied age-appropriate math tests to children. In every grade, these tests covered three domains. The first domain is *number recognition, addition, and subtraction*. In this domain we asked children to recognize numbers, and carry out simple additions and subtractions. The second domain is *number sense*. Here, children were asked to fill in a missing number in a sequence, place numbers on a number line, or make comparisons of numerical quantities. The third domain is *word problems*, in which children were asked to solve simple problems that were read to them. In each case, within a domain, the questions increase in difficulty as children get older.

In our main results, we score each domain by Item Response Theory (IRT), and standardize the score in each domain so it has mean zero and unit standard deviation. We then calculate a total math score as the sum of the scores in the three domains (which each domain receiving the same weight). However, as we show in Appendix B, our results are very robust to alternative ways of scoring the tests—for example, if instead of using IRT we simply record the number of correct responses—and alternative ways of weighting the three domains—for example, if we aggregate the scores on the three domains by factor analysis, or use the approach proposed by Anderson (2008).

[12] 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).

In two of the three test domains (*number recognition, addition and subtraction*, and *word problems*) there was some overlap in the questions in different grades.[13] We use these common questions to calculate "grade equivalents", defined as the average difference in test scores between children in two adjacent grades. We find that the average difference in achievement between children in kindergarten and 1st grade is 1.80 standard deviations, and 1.39 standard deviations between children in 1st and 2nd grades.[14] Further details on the math tests, including on aggregation and the calculation of grade equivalents, are given in Appendix B.

Table 1, Panel A, summarizes the characteristics of children and their families. Children were approximately 5 years of age on the first day of kindergarten. Just under 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). Sixty-two percent of children in the study sample attended preschool. At the beginning of kindergarten, we also tested children with the *Test de Vocabulario en Imágenes Peabody* (TVIP), the Spanish version of the widely used Peabody Picture Vocabulary Test (PPVT) (Dunn et al. 1986). Performance on this test at early ages has been shown to predict important outcomes in a variety of settings, including in Ecuador.[15] The average child in the 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.[16] The baseline TVIP score of girls in our sample is lower than that of boys, a difference of about 0.1 standard deviations.

[13] For example, in the *number recognition, addition, and subtraction* domain, the hardest numbers we asked children to identify at the end of kindergarten were the easiest numbers we asked them to identify in the 1st grade test.

[14] These values are somewhat larger than those typically found in the United States. For example, Hill et al. (2008) analyze six major standardized tests in math, and report an increase in average math achievement of 1.14 standard deviations between kindergarten and 1st grade, and 1.03 standard deviations between 1st and 2nd grade.

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

[16] 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).

Table 1, Panel B, summarizes the characteristics of teachers in our sample. Teachers are in their mid-40s, on average, and 93 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 38 students.

We measure the quality of teacher-student interactions using the CLASS (Pianta et al. 2007). 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.

The CLASS measures the quality of the interactions in three domains: emotional support, classroom organization, and instructional support. Within each of these domains, there are a number of CLASS dimensions. The behaviors that coders are looking for in each dimension are quite specific—see Appendix C for an example. For each of these behaviors, the CLASS protocol gives coders concrete guidance on whether the score given should be ‘‘low’’ (scores of 1–2), ‘‘medium’’ (3–5), or ‘‘high’’ (6–7). In practice, in our application of the CLASS (as well as in others), scores across different dimensions are highly correlated with each other. For this reason, we focus on a teacher's total CLASS score.[17] 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 teachers for a full day of classes (from approximately 8 in the morning until 1 in the afternoon); they did not know on what day they would be filmed until the day itself. Further details on the process of CLASS filming and coding are given in Appendix C.

Figure 1 graphs univariate densities of the distribution of total CLASS scores for teachers in our study, separately by grade. The average score is 3.3. A few teachers in our sample have CLASS scores in the ‘‘low’’ range (scores of 1 or 2), but the vast majority, more than 80 percent, have scores between 3 and 4. In the figure we also graph the distribution of CLASS scores in a

[17] We follow the CLASS protocols to construct the aggregate CLASS scores. Specifically, each dimension gets the same weight to calculate the score for each domain, and the three domains are added, with equal weight given to each, for the construction of the overall CLASS score. For ease of interpretation, in our main analysis we standardize the CLASS to have mean zero and unit standard deviation (separately by grade).

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.

3. Estimation strategy and results

A. SES, gender, and math achievement

To motivate our results, Figure 2 presents grade-specific means of math achievement by maternal education (Panel A), gender (Panel B), and the difference in test scores between boys and girls, by maternal education (Panel C). Panel A shows that there are substantial maternal education gradients in math test scores. Most of the difference in achievement between children of mothers with more or less education is already apparent at the end of kindergarten.[18] Panel B shows a substantial difference between the test scores of boys and girls, especially by the end of 2nd grade: At the end of kindergarten boys have test scores that are 0.08 standard deviations higher than girls, but the gap has doubled by the end of 2nd grade. Panel C shows that the boy-girl differences increase for all three maternal education groups. However, the gap in test scores between boys and girls is much larger for mothers with primary or secondary school education (0.14 and 0.19 standard deviations, respectively, by the end of 2nd grade) than for mothers with university education (0.07 standard deviations by the end of 2nd grade).

To more formally investigate the patterns in Figure 2, we run regressions of the following form:

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

where Y_{ihgs} is the math test score of child i in household h and grade g in school s; $E2_{ihgs}$ and $E3_{ihgs}$ are indicator variables for children of mothers with at least some secondary education and at least some university education, respectively; $Female_{ihgs}$ is an indicator variable for girls; ($E3*Female_{ihgs}$) refers to the daughters of mothers with some university education; Age_{ihgs} is a vector of dummies for single months of child age; and ε_{ihgs} is the error term. The omitted category

[18] As we discuss in Appendix B, there is some evidence that the distribution of test scores in kindergarten (but not in 1st or 2nd grade) is truncated on the left. To the extent that truncation plays a role, we would be underestimating the true differences in kindergarten between children of mothers with more or less education, and between girls and boys, and overestimating changes between kindergarten and 2nd grade.

in this regression corresponds to boys whose mothers have only primary school education. Standard errors are clustered at the school level. We are agnostic about the extent to which the β coefficients in (1) have a causal interpretation or are simply conditional associations.

Our first set of regression results are in Table 2. The table confirms that girls have significantly lower test scores than boys, by 0.13 standard deviations on average. The boy-girl gap increases as children progress through the school system—by the end of 2nd grade it is 0.17 standard deviations, about 0.12 grade equivalents. The next two rows in the table show that children of mothers with more education have higher test scores—relative to children of mothers with primary school education, boys of mothers with secondary school or university have 0.26 standard deviation (0.19 grade equivalents) and 0.47 standard deviation (0.34 grade equivalents) higher scores, respectively. The coefficient on the interaction term is always positive and, with the exception of the 2nd grade regression, is of roughly the same magnitude (but opposite sign) as the main effect for girls. The last line of the table shows the p-value for an F-test of the null hypothesis that the test scores of boys and girls are the same among children of mothers with university education. Regardless of whether we look at data pooled across different grades, or at grade-specific data, we never reject this hypothesis.

The most interesting result in Table 2 is the absence of a gender gap in math achievement among children of mothers with university education. To explore this result further, we create similar indicator variables for the education of fathers (rather than mothers), as well as 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.[19] We partition the wealth data in this way to closely mimic the distribution of education—in our sample, 39.2 percent of children have mothers with at most primary education, 50.9 percent have mothers with secondary school education, and 9.9 percent have mothers with at least some university education. We then run regressions that include only the indicator variables for paternal education and the interaction between girls and paternal university education; only the indicator variables for the middle and top wealth groups, and the interaction between girls and the richest wealth group; or multiple measures

[19] 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.

of SES (education of both parents and wealth) and the interactions. These results are reported in Table 3.

The table shows that, as with maternal education, there are substantial gradients in paternal education and wealth. The association between paternal and maternal education and the math achievement of boys is very similar: A boy whose father has university education has test scores that are 0.50 standard deviations higher, on average, than a boy of comparable age whose father only has primary school education.[20] However, in households in which fathers have university education, or households with relatively high wealth, girls have significantly lower test scores than boys. When only the paternal education categories and interaction are included as regressors, the coefficient on the interaction between university education and girls is small and insignificant (0.040, with a standard error of 0.076); when only the wealth categories are included as regressors, the coefficient on the interaction between being in the top wealth category and girls is also small and insignificant (-0.032, with a standard error of 0.065);[21] when all measures of SES are included, the coefficient on the interaction between maternal university education and girls is similar to that estimated in the regression without paternal education and wealth (0.097 rather than 0.106), although it is estimated more imprecisely (a standard error of 0.072 rather than 0.052).

Although this evidence is by no means definitive, these results suggest that it is something about high levels of maternal education, not SES generally, that protects the math achievement of girls from deviating too much from that of boys.[22] In interpreting these results, we note that in Ecuador, as in other Latin American countries (and unlike many countries in South and East Asia), there are no gender differences in infant mortality, malnutrition, school attendance, or the likelihood of graduation from secondary school.

We next estimate the magnitude of the boy-girl gap in test scores in our data separately for the three domains of math achievement: *number recognition*, *addition, and subtraction*, *number*

[20] The data on paternal education is missing for 22.2 percent of children. If we run the regressions in column (1) of Table 3 only for the sample of children for whom paternal education is not missing the coefficient on the indicator variable for girls is 0.136 (0.021); the coefficients on the indicator variables for children of mothers with secondary school and university education are 0.292 (0.025) and 0.488 (0.054), respectively; and the coefficient on the interaction between mothers with university education and girls is 0.085 (0.066).

[21] We obtain similar results if, instead, we partition the data into wealth quartiles or quintiles, and interact the variable for girls with the indicator variable for the richest wealth group, defined in this alternative way.

[22] To investigate whether what matters is the *relative* schooling of mothers and fathers, we also ran regressions that included as regressors maternal education, the difference between maternal and paternal education, and the interaction between this difference and the dummy variable for girls. In this regression, the coefficient on the interaction is -0.023, with a standard error of 0.026. We conclude that it is high levels of maternal education, rather than the relative education of mothers and fathers, which protect the math achievement of girls.

sense, and *word problems*. The literature from the United States shows that the largest girl disadvantage in math test scores is generally found on tasks that require visuo-spatial skills, as well as in math tasks that are more abstract and are not taught directly in the school curriculum.[23] Table 4 shows that the girl disadvantage in test scores is largest in the *number sense* domain (a boy-girl gap of 0.14 standard deviations), precisely the domain that tested children on tasks that are mostly not covered explicitly in the school curriculum in Ecuador.[24] Table 4 also shows that the smallest boy-girl gap is found in *word problems*. We note that this occurs in spite of the fact that the baseline TVIP score of girls is lower than that of boys, as discussed above.[25]

B. Differences in the distribution of test scores

To begin the analysis of differences in the distribution of test scores in our data, we first graph the density of test scores of boys and girls. Figure 3 suggests that the differences between boys and

[23] Halpern et al. (2007) show there are differences in the kinds of math tasks at which males and females excel. Females tend to outperform males in basic computational tasks, while males do better than females on tests that cover material that is not directly related to what is 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 0.9 to 1.0 standard deviations. 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).

[24] For example, in 2nd grade, we gave children an unmarked number line with the values 0 and 50 on each end, and asked them to place five randomly generated numbers between 1 and 49 on this line. Siegler and his coauthors have shown that, at early ages, children tend to place numbers on a logarithmic scale, so that the numbers on the lower end of the line are placed far apart (for example, the distance between the number 3 and 5 is larger than it should be), while numbers at the end of the scale are compressed (for example, the distance between 45 and 49 is smaller than it should be) (Siegler and Booth 2004; Siegler and Opfer 2003). There is evidence that the ability of children to accurately place numbers on a number line is highly predictive of later math achievement (Halberda et al. 2008; Sasanquie et al. 2012; Siegler and Booth 2004). Boys in our sample substantially outperform girls on this number line task (by 0.19 standard deviations). On the other hand, the 2nd grade difference between boys and girls in simple additions and subtractions, a skill that is an explicit focus of the curriculum, is much smaller (0.08 standard deviations).

[25] Women are generally believed to be more risk-averse than men (Dohmen et al. 2011; Buser et al. 2014), although it is not clear whether this is a result of inherent gender traits or socialization (Booth and Nolen 2012). It is in principle possible that the reason girls have lower math scores than boys in our data is that they are more reluctant to answer a question incorrectly. A priori, we do not believe that this is likely, as the tests were applied to children individually and, when a child did not respond, (s)he was prompted to try to give an answer. However, there is an exception to this general testing format with the test of additions and subtractions that was applied to children in 1st and 2nd grades. In this test, each child was given 90 seconds (in 1st grade) or 3 minutes (in 2nd grade) to answer as many simple addition problems as possible and, separately, the same amount of time to answer as many subtraction problems as (s)he could. Thus, on this test, we can see whether girls answered fewer questions overall, and whether the proportion of correct responses given by girls is no lower (or possibly higher) than is the case for boys. We find no evidence that this is the case. For example, on the 1st grade addition task, boys attempted more questions than girls on average (8.0, rather than 7.5, out of a possible 16 questions), but were also more likely to give a correct answer to a question they attempted (a ratio of correct responses to total responses of 0.73 for boys, and 0.70 for girls). Both of these differences are significant at the 1 percent level.

girls in the left tail of the distribution are modest. On the other hand, the boy-girl gap appears to be substantial at high test scores.

To analyze boy-girl differences in the distribution of math achievement more formally, we generate indicator variables for children whose scores are below a given cutoff (below the 5th, 10th, 25th, and 50th percentiles), or above a given cutoff (above the 75th, 90th, 95th, and 99th percentiles), and run regressions comparable to those in (1) above, but using these indicators (rather than the mean score) as the dependent variable. These results, reported in Table 5, confirm the pattern observed in Figure 3. There are a higher-than-expected number of girls in the left tail of the distribution,[26] but the boy-girl differences are modest. Girls have a 1.3 percentage points higher probability of being below the 10th percentile of the distribution than boys (with a standard error of 0.005), and a 0.3 percentage point higher probability of being below the 5th percentile (with a standard error of 0.004). At the bottom of the distribution, girls and boys of mothers with university education have similar math test scores—indeed, a comparison of the main effect for girls and the interaction term indicates that, among children of mothers with university education, there are *more* boys than girls at these very low test scores. On the other hand, there are a fewer-than-expected girls in the right tail of the distribution, and the boy-girl difference is large, especially at the higher achievement levels. Boys have a 3.3 percentage points higher probability of being above the 90th percentile of the distribution (with a standard error of 0.004), and a 2.1 percentage points higher probability of being above the 95th percentile (with a standard error of 0.003).[27] Importantly, at these high achievement levels, there are significantly fewer girls than boys even among children of mothers with university education.

C. Teachers and the distribution of test scores

The most novel aspect of our data is the as-good-as-random assignment of children to teachers, with very high levels of compliance, in three consecutive grades. This allows us to carefully study whether the increase in the boy-girl gap in math achievement we observe is a result of differences in the returns to teacher quality by gender.

[26] In the sense that, for example, more than 10 percent of girls in our sample score below the 10th percentile of the math distribution.

[27] The ratio of boys to girls above the 95th percentile of the distribution is 1.57. Bharadwaj et al. (2012) report a ratio of 1.35 in 4th grade in their data from Chile.

If boys are more sensitive to teacher (or school) quality than girls, we would expect that over time the variance of test scores of boys would increase relative to the variance of test scores of girls.[28] To motivate our analysis, we therefore begin by calculating the ratio of the boy-girl variance of test scores, separately by grade. We find that the variance of test scores of boys is always larger than that of girls, a fact that has been well-documented for the United States in the literature from psychology (Halpern et al. 2007). Importantly, however, the ratio of the variances does *not* increase as children move through the school system: The ratio is 1.093 in kindergarten, 1.081 in 1st grade, and 1.102 in 2nd grade.

We next 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, indicators for child age, and classroom indicators;[29] (2) construct residualized math scores by subtracting the estimated effects of lagged math scores and age; (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, and some variance. In principle, the square root of this variance 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 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.[30] 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.

[28] To see this, suppose test scores evolve in the following way: $Y_{igs} = \beta_1 Y_{ig-1s} + \beta_{2g} Q_{gs} + \varepsilon_{igs}$, where Y_{igs} is the test score of student i in grade g in school s, and Q_{gs} is the quality of school s in grade g. Suppose that depreciation of knowledge, measured by β_1, does not vary by gender, but sensitivity to quality, measured by β_{2g}, is larger for boys. Assume also that Q_{gs} is independent over time (or, at least, is not negatively correlated over time). Then, even if the variance of lagged test scores (Y_{ig-1s}) and the variance of the residual (ε_{igs}) are the same for boys and girls, since β_{2g} is larger for boys, the variance of current test scores (Y_{igs}) will be larger for boys.

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

[30] 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).

Our first set of results is in Panel A of Table 6. The table has three rows, with each row corresponding to a different grade, and four columns, with each column corresponding to estimates of classroom effects (and the associated standard errors) for the whole sample, girls, boys, and the p-value of a test of differences in the classroom effects for boys and girls. Classroom effects are somewhat larger for kindergarten than for 1st and 2nd grade, although the differences are modest, and are not statistically different from zero. In kindergarten and first grade, the estimated classroom effects are 0.12 for boys and 0.12 for girls, and in second grade the estimated effects are 0.10 for both genders. We cannot reject the null that the estimated effects are the same for boys and girls in every grade.

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.[31] We use these data to calculate kindergarten teacher effects, once again separately for boys and girls. Panel B of Table 6 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 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.[32] To explore this, we calculate the 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.[33] These

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

[32] Dee (2005) argues that, in the United States, boys learn more from male than female teachers, while the reverse is true for girls. We cannot carry out such an analysis with our data because 99 percent, 93 percent, and 87 percent of teachers in kindergarten, 1st grade, and 2nd grade, respectively, are women.

[33] 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

estimates are reported in Table 7. 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 0.24 for girls, and 0.11 for boys.

Table 7 provides no evidence that some 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 0.18 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 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 0.11 are statistically indistinguishable from each other).

As a final step in our analysis of the role of teacher quality, we turn to the data on teacher-child interactions, as measured by the CLASS. To provide some context, we first regress math achievement on the CLASS and school fixed effects, without disaggregating the data by gender.[34] These results, in Panel A of Table 8, suggest that the association between the CLASS and math achievement declines as children become older—from 0.063 in kindergarten (with a standard error of 0.017), to 0.051 in 1st grade (with a standard error of 0.017), and to 0.033 in 2nd grade (with a standard error of 0.020).[35]

Next, we regress math test scores on school fixed effects, the indicator variable for girls, the indicator variables for children of mothers with secondary school and university education, respectively, the main effect in the CLASS, and the interaction between girls and the CLASS. These results, reported in Panel B of Table 8, show that the CLASS is associated with the learning

---

of the gender-specific standard deviations of teacher quality reported in Table 6. 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.

[34] The regressions that pool the data for all three grades include school-by-grade fixed effects (the level at which random assignment took place) rather than school fixed effects.

[35] The coefficient on the CLASS in Table 8 is a little larger for the kindergarten regressions than for the 1st grade regressions in part because we standardize the CLASS coefficients to have mean zero and unit standard deviation, separately by grade, and the CLASS has somewhat higher variance in kindergarten (0.081) than in 1st grade (0.054) or 2nd grade (0.059). If we use the raw CLASS scores instead of the standardized CLASS scores in the regressions in Panel A of Table 8, the coefficients on the CLASS for kindergarten, 1st grade, and 2nd grade are 0.221 (0.059), 0.222 (0.075), and 0.137 (0.084), respectively.

of boys and girls in equal measure. For example, when we pool the data for the three grades, the coefficient on the CLASS is 0.047 (with a standard error of 0.013), and the coefficient on the interaction is smaller than 0.0005 (with a standard error of 0.010).

In sum, we find no evidence that boys in our sample are more sensitive to variations in teacher quality than girls, no matter whether we measure quality on the basis of classroom effects, teacher effects, or the quality of the interactions between teachers and children.

4. Conclusion

Early math achievement is associated with later math achievement, with completed schooling, and with labor market outcomes. In spite of this, there is much to be learned on the determinants of early differences in math test scores between children in rich and poor families, and between boys and girls, especially in developing countries.

In this paper, we study math achievement between kindergarten and 2nd grade in Ecuador. Our data are unique in a number of ways, in particular in the as-good-as-random assignment of children to teachers in three consecutive grades, with very high compliance with the assignment rule. The quality of the experiment and the data we use allows us to make three important contributions to the literature on early math learning outcomes.

First, we show that there are steep socioeconomic gradients and substantial differences in math achievement between boys and girls. Differences between children of parents with more or less education are already apparent at the end of kindergarten and increase only modestly by 2nd grade. The difference between boys and girls, on the other hand, roughly doubles in two years.

Second, we find that, among children of mothers with university education, there is no difference in the average math achievement of boys and girls. This suggests that the home environment is an important determinant of the gender gap in early math test scores. Further research to better understand how SES generally, and maternal education specifically, affects math learning at early ages seems to us an important priority.

Third, we show that in our data there is no evidence that boys and girls respond differently to as-good-as-random variations in teacher quality, no matter whether we measure quality on the basis of classroom effects, teacher effects, or the quality of the interactions between teachers and young children. We cannot rule out that there are teacher behaviors, uniform for all teachers, that lead to the lower math achievement of girls, especially as they spend more years in school. There

would be high returns to policy experimentation, and careful evaluation, of 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.

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. 2016a. "School Quality and the Gender Gap in Educational Achievement." *American Economic Review, Papers and Proceedings* 106(5): 289-95.

----------. 2016b. "Family Disadvantage and the Gender Gap in Behavioral and Educational Outcomes." NBER Working Paper 22267.

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.

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

Bethell, George. 2016. *Mathematics Education in Sub-Saharan Africa: Status, Challenges, and Opportunities*. Washington, D.C.: World Bank.

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.

Booth, Alison L., and Patrick Nolen. 2012. "Gender Differences in Risk Behavior: Does Nurture Matter?" *Economic Journal* 122(558): F56-F78.

Buser, Thomas, Muriel Niederle, and Hessel Oosterbeek. 2014. "Gender, Competitiveness, and Career Choices." *Quarterly Journal of Economics* 129(3): 1409-47.

Carneiro, Pedro, and James J. Heckman. 2003. "Human Capital Policy." NBER Working Paper 9495.

Carneiro, Pedro, Costas Meghir, and Matthias Parey. 2013. "Maternal Education, Home Environments, and the Development of Children and Adolescents." *Journal of the European Economic Association* 11(Supplement 1): 123-60.

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

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.

Chetty, Raj, Nathaniel Hendren, Frina Lin, Jeremy Majerovitz, and Benjamin Scuderi. 2016. "Childhood Environment and Gender Gaps in Adulthood." *American Economic Review, Papers and Proceedings* 106(5): 282-88.

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.

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

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.

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jurgen Schupp, and Gert G. Wagner. 2011. ''Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences.'' *Journal of the European Economic Association* 9(3): 522–50.

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.

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.

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.

Halberda, Justin, Michele Mazzocco, and Lisa Feigenson. 2008. "Individual Differences in Nonverbal Number Acuity Correlate with Maths Achievement." *Nature* 455: 665-68.

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.

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

Hanushek, Eric, and Dennis D. Kimko. 2000. "Schooling, Labor Force Quality, and the Growth of Nations." *American Economic Review* 90(5): 1184–1208.

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

Hanushek, Eric, and Ludgar Woessmann. 2008. ‘‘The Role of Cognitive Skills in Economic Development.’’ *Journal of Economic Literature* 46(3): 607–68.

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.

Hill, Carolyn J., Howard S. Bloom, Alison Rebeck Black, and Mark W. Lipsey. 2008. “Empirical Benchmarks for Interpreting Effect Sizes in Research.” *Child Development Perspectives* 2(3): 172-77.

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.

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.

Krueger, Alan B. 1999. “Experimental Estimates of Education Production Functions.” *Quarterly Journal of Economics* 114(2): 497-532.

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.

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.

OECD. 2014. *PISA 2012 Results: What Students Know and Can Do – Student Performance in Mathematics, Reading and Science (Volume I, Revised edition, February 2014*). PISA: OECD Publishing.

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/

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

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

Robinson, Nancy M., Robert D. Abbott, Virginia W. Berninger, and Julie Busse. 1996. "The Structure of Abilities in Math-Precocious Young Children: Gender Similarities and Differences." *Journal of Educational Psychology* 88(2): 341–52.

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.

Sasanguie, Delphine, Bert De Smedt, Emmy Defever, and Bert Reynvoet. 2012. "Association between Basic Numerical Abilities and Mathematics Achievement." *British Journal of Developmental Psychology* 30(2): 344-57.

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.

Siegler, Robert S., and Julie L. Booth. 2004. "Development of Numerical Estimation in Young Children." *Child Development* 75(2): 428-44.

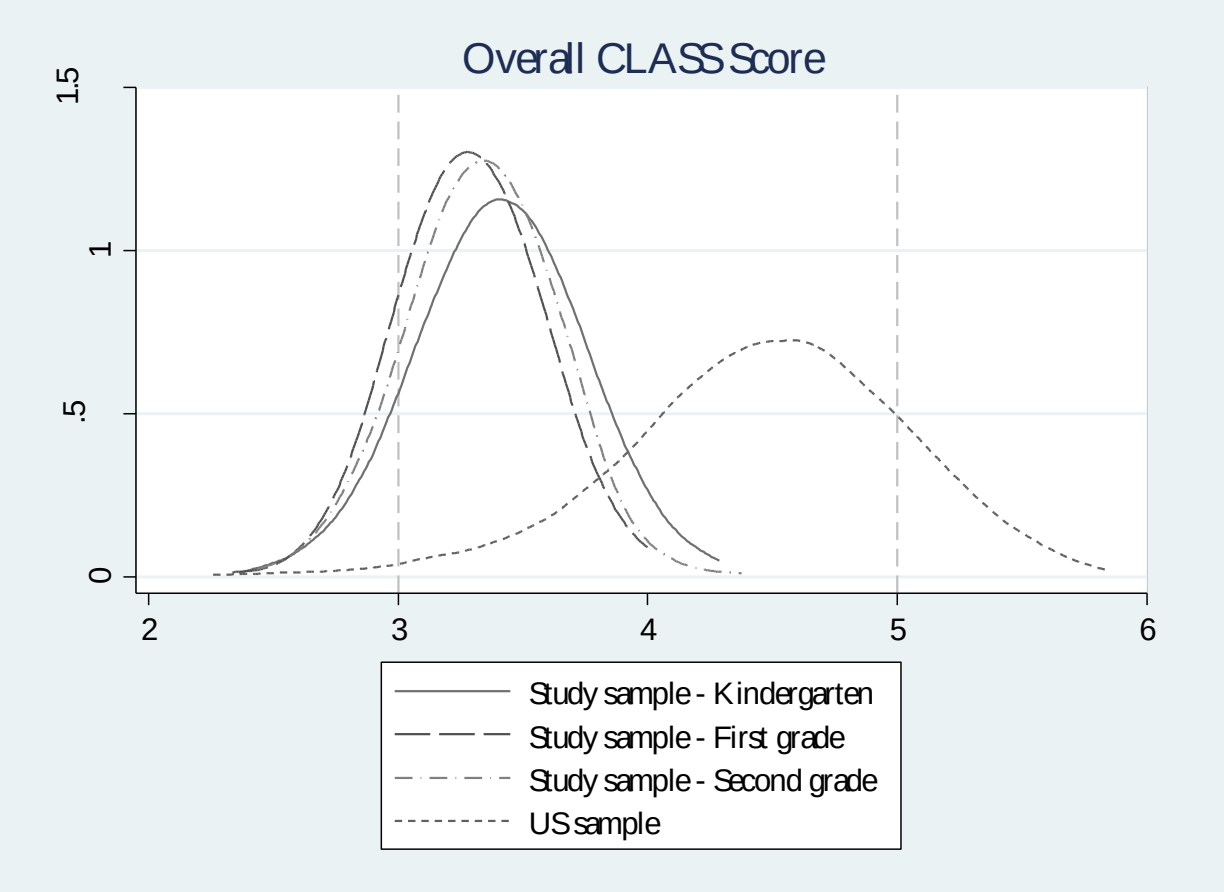
Siegler, Robert S., and John E. Opfer. 2003. "The Development of Numerical Estimation: Evidence for Multiple Representations of Numerical Quantity." *Psychological Science* 14(3): 237-43.

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

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

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.

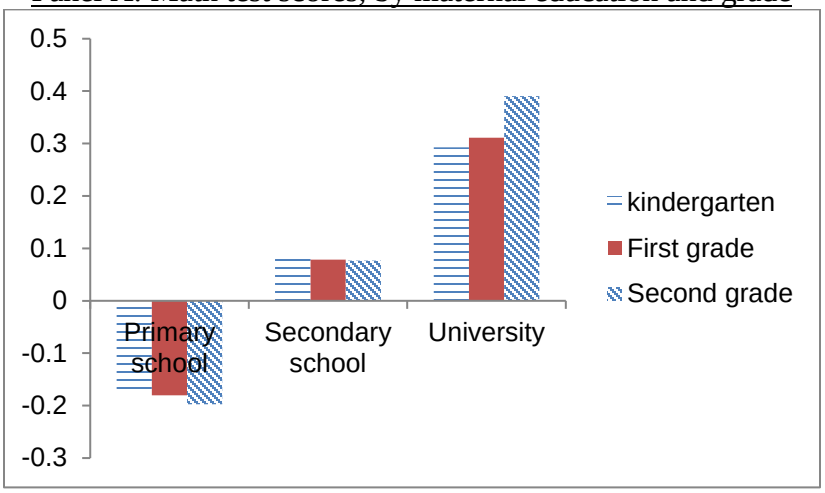
Figure 1: Distribution of CLASS scores, Ecuador and US data



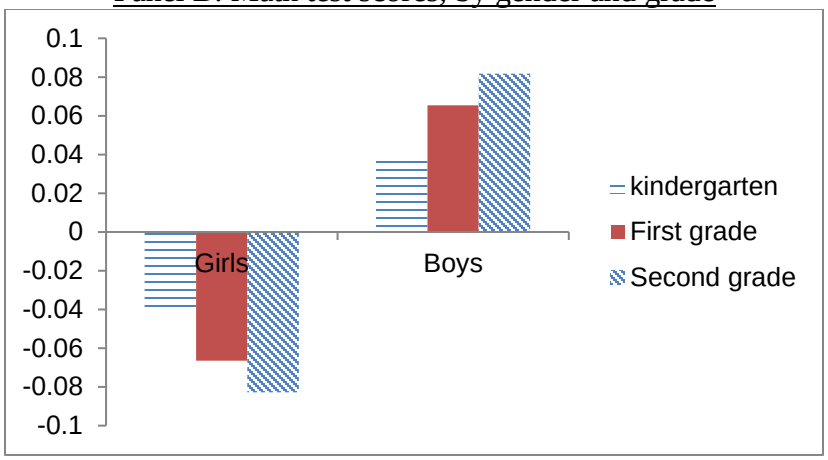
Note: The figure graphs univariate densities of the CLASS score of kindergarten, 1st grade, and 2nd grade teachers in Ecuador, and in a nationally representative sample of kindergarten classrooms in the United States (Clifford et al. 2003). The CLASS is scored on a 1–7 scale; scores of 1–2 indicate poor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

Panel A: Math test scores, by maternal education and grade



Panel B: Math test scores, by gender and grade



Panel C: Difference in test scores between boys and girls, by maternal education and grade

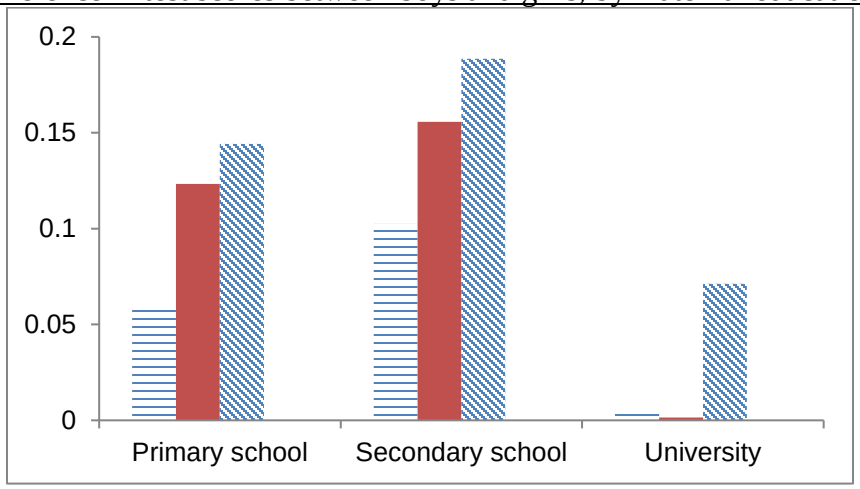


Figure 3: Density of math test scores, by gender

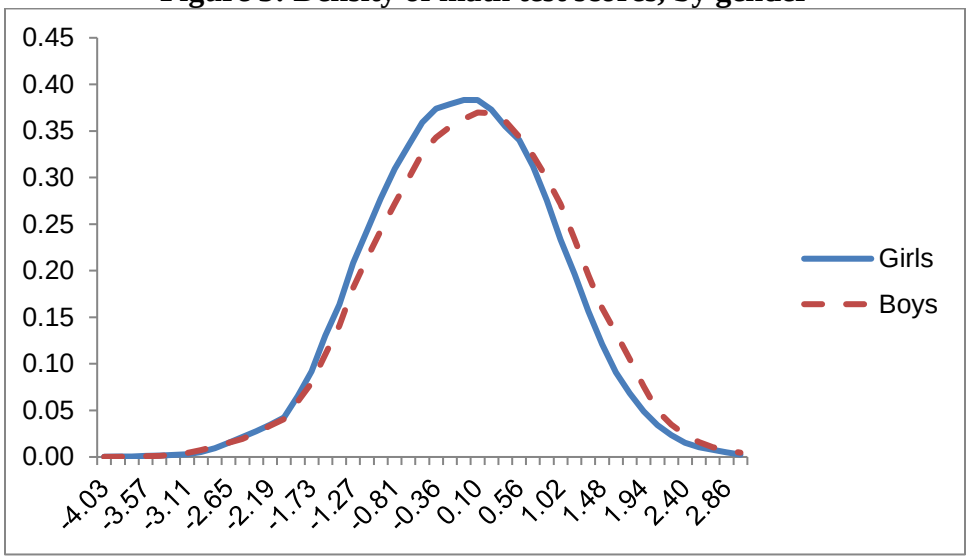


Table 1: Summary Statistics

	Mean	S.D.	Obs.
Panel A: Children			
Age (months)	60.26	4.62	10,521
Proportion female	0.50	0.50	10,671
TVIP	83.23	15.80	10,102
Mother's age	30.36	6.57	10,670
Father's age	34.68	7.92	8,318
Mother's years of schooling	8.82	3.78	10,671
Father's years of schooling	8.49	3.79	8,304
Attended preschool	0.62	0.49	10,643
Panel B: Teachers			
Age	43.95	10.1	1,331
Proportion female	0.93	0.3	1,354
Experience	18.14	10.4	1,354
Proportion tenured	0.75	0.4	1,349
Class size	37.52	7.9	1,355

Notes: 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, gender, and math achievement

	All grades	Kindergarten	First grade	Second grade
Dummy: girls	-.131***	-.072***	-.134***	-.165***
	(.019)	(.021)	(.022)	(.021)
Dummy: Secondary school	.264***	.288***	.271***	.278***
	(.023)	(.027)	(.028)	(.028)
Dummy: University	.465***	.479***	.448***	.552***
	(.044)	(.054)	(.051)	(.054)
Interaction: University*girls	.106**	.071	.131**	.093**
	(.052)	(.069)	(.056)	(.061)
F-test (p-value)	0.60	0.99	0.96	0.20

Note: Sample size is 31,398 in the all grades regression, and 10,466 in each of the grade-specific regressions. All regressions include age in months and its square. F-test is the p-value on an F-test that the sum of the coefficients on the dummy for girls and the interaction between girls and the dummy for mothers with university education is zero. Standard errors corrected for clustering at the school level. *, **, and ***, significant at the 10 percent, 5 percent and 1 percent, respectively.

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

	(1)	(2)	(3)	(4)
Dummy: girls	-.131***	-.130***	-.121***	-.131***
	(.019)	(.020)	(.018)	(.021)
Dummy: Mother secondary school	.264***			.177***
	(.023)			(.025)
Dummy: Mother university	.465***			.217***
	(.044)			(.060)
Interaction: Mother university*girls	.106**			.097
	(.052)			(.072)
Dummy: Father secondary school		.263***		.150***
		(.023)		(.022)
Dummy: Father university		.495***		.254***
		(.055)		(.058)
Interaction: Father university*girls		.040		0.015
		(.076)		(.082)
Wealth-Middle			.221***	.122***
			(.023)	(.027)
Wealth-Top			.510***	.327***
			(.045)	(.055)
Interaction: Wealth-Top*girls			.030	-.032
			(.056)	(.065)
F-test 1 (p-value)	0.60			0.63
F-test 2 (p-value)		0.22		0.16
F-test 3 (p-value)			0.10	0.02
Number of observations	31,398	24,504	31,398	24,504

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 fathers with university education is zero. F-test 3 is test that the sum of the coefficients on girls and the interaction between girls and the dummy for the top wealth decile is zero. Standard errors corrected for clustering at the school level. *, **, and ***, significant at the 10 percent, 5 percent and 1 percent, respectively.

Table 4: Heterogeneity by test

	TOTAL	Test score: *Number recognition, addition, and subtraction*	Test score: *Number sense*	Test score: *Word problems*
Dummy: girls	-.131***	-.118***	-.144***	-.056***
	(.019)	(.018)	(.016)	(.016)
Dummy: Secondary school	.264***	.247***	.217***	.175***
	(.023)	(.024)	(.021)	(.019)
Dummy: University	.465***	.423***	.375***	.333***
	(.044)	(.040)	(.040)	(.039)
Interaction: University*girls	.106**	.087*	.084*	.083*
	(.052)	(.048)	(.047)	(.050)
F-test (p-value)	0.60	0.49	0.20	0.57

Note: Sample size is 31,398 in all regressions. All regressions include child age in months and its square. Standard errors corrected for clustering at the school level. *, **, and ***, significant at the 10 percent, 5 percent and 1 percent, respectively.

Table 5: Distributional effects

	$< 5^{th}$ pctile	$< 10^{th}$ pctile	$< 25^{th}$ pctile	$< 50^{th}$ pctile	$> 75^{th}$ pctile	$> 90^{th}$ pctile	$> 95^{th}$ pctile	$> 99^{th}$ pctile
Dummy: girls	.003	.013***	.036***	.055***	-.054***	-.033***	-.021***	-.005***
	(.004)	(.005)	(.007)	(.008)	(.007)	(.004)	(.003)	(.001)
Dummy: Secondary school	-.031***	-.051***	-.087***	-.104***	.080***	.041***	.027***	.007***
	(.004)	(.006)	(.009)	(.010)	(.008)	(.005)	(.003)	(.001)
Dummy: University	-.048***	-.078***	-.137***	-.186***	.163***	.082***	.046***	.011**
	(.006)	(.009)	(.015)	(.019)	(.017)	(.012)	(.010)	(.006)
Interaction: University*girls	-.007	-.024***	-.057***	-.060**	.018	.008	-.003	-.003
	(.006)	(.010)	(.018)	(.025)	(.023)	(.015)	(.011)	(.006)
F-Test (p-value)	0.45	0.19	0.19	0.83	0.11	0.10	0.03	0.18
Boy-girl ratio	0.96	0.90	0.89	0.91	1.26	1.42	1.57	1.74

Note: Sample size is 31,398 in all regressions. All regressions include child age in months and its square. Standard errors corrected for clustering at the school level. *, **, and ***, significant at the 10 percent, 5 percent and 1 percent, respectively.

Table 6: Classroom and teacher effects, by gender

	All children	Girls	Boys	Difference (p-value)
PANEL A: Classroom effects				
Kindergarten	0.12 (0.01)	0.12 (0.02)	0.12 (0.02)	0.79
1st grade	0.11 (0.01)	0.12 (0.01)	0.12 (0.02)	0.67
2nd grade	0.09 (0.02)	0.10 (0.01)	0.10 (0.01)	0.91
PANEL B: Teacher effects				
Kindergarten	0.09 (0.02)	0.11 (0.03)	0.07 (0.03)	0.27

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. Teacher effects are given by the square root of the covariance of the classroom effects for the two kindergarten cohorts. Standard errors, in parentheses, are calculated with a block bootstrap, with blocks equal to schools. The p-value in the fifth column corresponds to a test that the difference in the classroom or teacher effects for boys and girls is zero.

Table 7: Correlations in learning gains across cohorts

		Cohort 2		
		All	Girls	Boys
Cohort 1	All	0.29 (0.10)	0.24 (0.11)	0.17 (0.12)
	Girls	0.24 (0.11)	0.24 (0.09)	0.13 (0.09)
	Boys	0.23 (0.11)	0.18 (0.08)	0.11 (0.09)

Note: The table reports the 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.

Table 8: CLASS scores, gender, and learning outcomes

	All grades	Kindergarten	First grade	Second grade
PANEL A				
CLASS score	.046***	.063***	.051**	.033
	(.011)	(.017)	(.017)	(.020)
PANEL B				
Dummy: girls	-.123***	-.074***	-.128***	-.167***
	(.011)	(.018)	(.019)	(.017)
Dummy: Secondary school	.213***	.211***	.213***	.218***
	(.014)	(.021)	(.023)	(.023)
Dummy: University	.462***	.444***	.451***	.501***
	(.020)	(.034)	(.035)	(.036)
CLASS score	.047***	.055**	.063**	.037
	(.013)	(.018)	(.020)	(.024)
Interaction: CLASS*girls	.000	.020	-.016	-.009
	(.010)	(.016)	(.017)	(.019)

Note: Sample size is 31,398 in the all grades regression, and 10,466 in each of the grade-specific regressions. All regressions include school fixed effects, and child age in months and its square. Standard errors corrected for clustering at the school level. *, **, and ***, significant at the 10 percent, 5 percent and 1 percent, respectively.

Appendix A: Randomization checks, and changes in the composition of the sample

Randomization checks: A major strength of our paper is that, in all three 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: 98.2 percent in kindergarten, 99.6 percent in 1st grade, and 98.0 percent in 2nd grade.[36]

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; and in 2nd grade, they were divided by gender, ordered by their first name and last name, and then assigned in alternating order.

We argue that these three instances of as-good-as-random assignment of children to teachers produce three exogenous, orthogonal shocks to teacher quality. This has three testable implications. First, the predetermined characteristics of teachers should be uncorrelated with the characteristics of children in their classrooms. Second, the rank that a child received by the assignment rule in one grade should be orthogonal with the rank she received by the assignment rule in the other grades.[37] Third, when teacher quality is defined as quality relative to the school mean, the quality of the teacher a child receives in one grade should be orthogonal with the quality of the teacher she receives in the other two grades. We provide strong evidence that all three implications of as-good-as-random assignment hold in our data.

Table A1 presents the results of regressions of three predetermined teacher characteristics (years of experience teaching, whether a teacher is tenured and, for 1st and 2nd grade, teacher gender),[38] on lagged test scores, child age, gender, maternal education, and wealth. We report the coefficients from 49 separate regressions. None is significant at the 5 percent level or higher.

[36] 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 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).

[37] For example, in kindergarten, a child named María Arce would have a very low rank number, and a child named Emiliano Zapata would have a very high rank number. In 1st grade, a child who is 87 months old (at the 95th percentile of the distribution of age) would have a very low rank, while a child who is 75 months old (at the 5th percentile) would have a very high rank. There is no reason to suppose that a ranking that is based on the last and first name would be correlated with a ranking that is based on child age.

[38] Recall that 99 percent of kindergarten teachers are women, so we cannot test whether children of different characteristics are more or less likely to be assigned to male or female teachers in kindergarten.

We next correlate the within-school ranking that each child had by the assignment rules in kindergarten, 1st grade, and 2nd grade.[39] The correlations are 0.0034 (for kindergarten rank and 1st grade rank), -0.0044 (for kindergarten rank and 2nd grade rank), and 0.0151 (for 1st grade rank and 2nd grade rank). None of these correlations are significant at the 10 percent level or higher.

Finally, we report the correlations of within-school, cross-classroom differences in the quality of teachers a child was assigned to in kindergarten, 1st grade, and 2nd grade. For this purpose, we use two distinct measures of teacher quality: the CLASS, and "classroom effects", as defined in the main body of the paper. In the case of the classroom effects, the correlations are -0.0114 (kindergarten and 1st grade classroom effects), -0.0071 (kindergarten 2nd grade classroom effects), and -0.0046 (1st and 2nd grade classroom effects). In the case of the CLASS, the correlations are -0.0072 (kindergarten and 1st grade CLASS), 0.0003 (kindergarten and 2nd grade CLASS), and -0.0041 (1st and 2nd grade CLASS). None of these correlations are significant at the 10 percent level or better.

We conclude from these checks that our as-good-as-random assignment rules did in fact produce three separate, exogenous shocks to teacher quality, as intended.

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 data on math test scores at the end of kindergarten, 1st grade, and 2nd grade. Conditioning on the availability of data on math achievement in all three grades ensures that our analysis of changes in the boy-girl test score gap is not driven by changes in the sample. In addition, we impose two further restrictions. First, we limit the sample to children for whom we have beginning-of-kindergarten TVIP scores. We do this because, in our analysis of kindergarten classroom and teacher effects, we follow our earlier work (Araujo et al. 2016) and look at within-school, cross-classroom differences in math achievement controlling for TVIP scores. (In 1st and 2nd grade, we control for math achievement at the end of the previous grade.) Second, we limit the sample to children for whom we have data on maternal education. We do this because we study SES gradients in math achievement, and the protective effect of maternal university education on girl math achievement.

Although it makes sense to limit the sample in this way for reasons of comparability, it 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. Specifically:

1. Estimates of the boy-girl gap in kindergarten exclude all children who left our sample of schools at any point in time between the end of kindergarten and the end of 2nd grade.

[39] In larger schools, the maximum rank by any of the assignment rules is higher than in smaller schools. To avoid a mechanical correlation in the ranking by the three rules, we first divide the rank each child had in each year by the highest value of the rank in that school in that grade.

2. Estimates of the boy-girl gap in 1st grade exclude all children who entered 1st grade after the end of kindergarten or left our sample of schools between the end of 1st grade and the end of 2nd grade.
3. Estimates of the boy-girl gap in 2nd grade exclude all children who entered our sample of schools in 1st grade or 2nd grade.
4. Estimates of the boy-girl gap in all three grades exclude children for whom we are missing baseline data on the TVIP or data on maternal education.

In practice:

1. 28 percent of all children with kindergarten test score data are not in the balanced panel: In order of importance, these are children who left the study schools after the end of kindergarten (15 percent) or after the end of 1st grade (7 percent), or are missing data on either baseline TVIP or maternal education (6 percent).
2. 36 percent of all children with 1st grade test score data are not in the balanced panel. In order of importance, these are new arrivals between kindergarten and 1st grade (25 percent), children who left the study schools after 1st grade (6 percent), and children who are missing data on either baseline TVIP or maternal education (5 percent).
3. 38 percent of all children with 2nd grade test score data are not in the balanced panel. In order of importance, these are new arrivals between kindergarten and 1st grade (19 percent), new arrivals between 1st and 2nd grade (13 percent), and children who are missing data on either baseline TVIP or maternal education (5 percent).

In sum, our calculations exclude between 28 percent (in kindergarten) and 38 percent (in 2nd grade) of all children in a given grade. We carry out a number of additional calculations to test how this may affect the interpretation of the coefficients that we report.

Children who enter sample ("new arrivals"):

1. *Children who enter the sample between kindergarten and 1st grade*: In a regression of end of 1st grade scores on an indicator variable for children who arrived in our sample of schools at some point between kindergarten and 1st grade, the coefficient on these "new arrivals is -0.107 (with a standard error of 0.027); in a comparable regression of the indicator variable for girls on the indicator variable for new arrivals, the coefficient is -0.024 (with a standard error of 0.010).
2. *Children who enter the sample between 1st grade and 2nd grade*: In a regression of end of 2nd grade scores on an indicator variable for children who arrived in our sample of schools at some point between 1st and 2nd grade, the coefficient on new arrivals is -0.069 (with a standard error of 0.030); in a comparable regression of the indicator variable for girls on the indicator variable for new arrivals, the coefficient is -0.048 (with a standard error of 0.012).

Children who exit sample ("leavers"):

1. *Children who exit the sample between kindergarten and 1st grade*: In a regression of end of kindergarten scores on an indicator variable for children who left our sample of schools at some point between the end of kindergarten and the end of 1st grade, the coefficient on these "leavers" is -0.128 (with a standard error of 0.037); in a comparable regression of the indicator variable for girls on the indicator variable for leavers, the coefficient is -0.024 (with a standard error of 0.011).
2. *Children who exit the sample between 1st grade and 2nd grade*: In a regression of end of 1st grade scores on an indicator variable for children who left our sample of schools at some point between the end of 1st grade and the end of 2nd grade, the coefficient on these "leavers" is -0.447 (with a standard error of 0.052); in a comparable regression of the indicator variable for girls on the indicator variable for leavers, the coefficient is -0.048 (with a standard error of 0.012).

Children who are missing baseline data on TVIP or maternal education:

1. In a regression of end-of-kindergarten scores on an indicator variable for children who are missing baseline data, the coefficient is -0.112 (0.032); in a comparable regression of the indicator variable for girls on the indicator variable for children missing baseline data, the coefficient is -0.006 (with a standard error of 0.015).
2. In a regression of end-of-1st grade scores on an indicator variable for children who are missing baseline data, the coefficient is -0.129 (0.024); in a comparable regression of the indicator variable for girls on the indicator variable for children missing baseline data, the coefficient is -0.023 (with a standard error of 0.009).
3. In a regression of end-of-2nd grade scores on an indicator variable for children who are missing baseline data, the coefficient is -0.126 (0.021); in a comparable regression of the indicator variable for girls on the indicator variable for children missing baseline data, the coefficient is -0.036 (with a standard error of 0.009).

In sum, the estimates we report in the main body of the paper refer to a sample of children who are on average brighter than the average child in the relevant grade. Moreover, fewer girls than boys are missing from the balanced panel.

The main robustness check we carry out is, where feasible, to reproduce the results in the paper for the largest possible sample of children in a given grade, and to compare these with estimates from the balanced panel.

Boy-girl gap in achievement: All children in a grade versus children in the balanced panel

We begin by reporting the average boy-girl gap in math achievement in each grade.

1. *Kindergarten*: In kindergarten, the boy-girl gap for the full sample of kindergarten students is -0.043 (with a standard error of 0.017), while the gap in the balanced panel is -0.078 (with a standard error of 0.021).
2. *1st grade*: In 1st grade, the boy-girl gap for the full sample of kindergarten students is -0.121 (with a standard error of 0.016), while the gap in the balanced panel is -0.125 (with a standard error of 0.020).
3. *2nd grade*: In 2nd grade, the boy-girl gap for the full sample of kindergarten students is -0.170 (with a standard error of 0.017), while the gap in the balanced panel is -0.160 (with a standard error of 0.020).

In sum, we find some evidence that, relative to the full sample of children in kindergarten, those that focus on children in the balanced panel may overstate the boy-girl gap in achievement, and may therefore understate the increase in the gap between kindergarten and 1st grade. In the other two grades, the difference in the boy-girl gap between the two samples is very small.[40]

Effect of teachers on learning outcomes of boys and girls: All children in a grade versus children in the balanced panel

We next turn to the robustness of the estimates we report in the paper on the effects of classroom and teacher quality on the math achievement of boys and girls.

In Table A2 below we reproduce the results in Table 6 in the paper, but include all children in a given grade in these estimates. These results show that the classroom effects we estimate (for boys and girls together, as well as for boys and girls separately) are very similar if we use the largest possible sample of children in a grade, or if we use only the balanced panel.

We also look at the association between the CLASS and math achievement, once again separating by grade and gender, for the largest sample of children in a grade and children in the balanced panel. As can be seen in Table A3, the basic pattern we observe in the balanced panel—smaller associations between the CLASS and math achievement in 2nd grade than in kindergarten, associations between the CLASS and math achievement that are of similar magnitude for boys and girls—are observed in both the largest possible sample of children in each grade and the balanced panel sample.

[40] For most grades, we cannot carry out calculations comparable to those in Table 2 in the main body of the paper for the full sample of children because we do not have data on baseline TVIP or maternal education on any "new arrivals". Kindergarten is an exception but it is not useful for this purpose because it is the grade in which the boy-girl gap is smallest, as is the estimate of the protective effect of university education on the math achievement of girls. Moreover, as we discuss above, this is also the grade in which there is a difference in the boy-girl gap in the two samples (all kindergarten children versus children in the balanced panel).

We conclude from these robustness checks that, although changes in the composition of the sample are always a concern with panel data, the main results we report in the paper are broadly similar if, instead of using the sample for the balanced panel, we use the sample that corresponds to the largest number of children in a given grade.

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

	Mean	Lagged test score (t-1)	Lagged test score (t-2)	Lagged test score (t-3)	Gender	Age	Maternal education	Wealth
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
					Kindergarten			
Years of experience	14.6	.033 (.047)			-.163 (.102)	-.005 (.009)	-.008 (.015)	.040 (.058)
Teacher is tenured	0.64	.002 (.003)			-.001 (.006)	-.000 (.001)	.000 (.001)	.006 (.004)
Teacher is female	0.99							
					1st grade			
Years of experience	19.2	.047 (.060)	.026 (.061)		-.011 (.114)	-.001 (.005)	.000 (.016)	.137* (.070)
Teacher is tenured	0.72	-.000 (.002)	-.001 (.002)		.007* (.004)	-.000 (.000)	.000 (.001)	.002 (.003)
Teacher is female	0.93	-.000 (.001)	-.001 (.001)		-.004 (.003)	.000 (.000)	-.000 (.001)	.001 (.001)
					2nd grade			
Years of experience	20.2	-.139* (.073)	-.014 (.067)	-.071 (.066)	.005 (.033)	-.009 (.006)	-.004 (.021)	-.089 (.083)
Teacher is tenured	0.89	-.004* (.002)	-.004* (.002)	-.001 (.002)	.000 (.001)	-.000 (.000)	.001 (.001)	.004 (.003)
Teacher is female	0.87	.002 (.002)	.001 (.002)	.002 (.002)	-.001 (.001)	.000 (.001)	.000 (.001)	.001 (.002)

Note: Each cell in columns (1) through (7) corresponds to a separate regression All regressions are based on children in the balanced panel sample. All regressions include school fixed effects. Standard errors clustered at the school level. The wealth aggregate is the first principal component of all the housing characteristics and asset ownership variables collected in the household survey. It includes whether the household has piped water and (separately) sewerage in the home; three variables for the main material of the floor, walls, and roof of the house, respectively; and whether the household owns a television, computer, fridge or washing machine (four separate variables). * significant at 10%, ** at 5%.

Table A2: Classroom effects, by gender, largest possible sample in each grade versus balanced panel

	Boys and girls		Girls only		Boys only	
	Balanced panel	Largest sample	Balanced panel	Largest sample	Balanced panel	Largest sample
Kindergarten	0.12	0.12	0.12	0.13	0.12	0.12
	(0.01)	(0.01)	(0.02)	(0.01)	(0.02)	(0.01)
1st grade	0.11	0.09	0.12	0.11	0.12	0.10
	(0.01)	(0.02)	(0.01)	(0.01)	(0.02)	(0.02)
2nd grade	0.09	0.10	0.10	0.10	0.10	0.09
	(0.02)	(0.03)	(0.01)	(0.04)	(0.01)	(0.04)

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 A3: CLASS scores, gender, and learning outcomes, largest possible sample in each grade versus balanced panel

	Boys and girls		Girls only		Boys only	
	Balanced panel	Largest sample	Balanced panel	Largest sample	Balanced panel	Largest sample
Kindergarten	.062***	.060***	.053**	.049**	.073***	.073***
	(.016)	(.015)	(.022)	(.020)	(.020)	(.018)
1st grade	.049***	.058***	.031	.053***	.068***	.061***
	(.017)	(.014)	(.019)	(.019)	(.021)	(.017)
2nd grade	.033	.024	.033	.017	.034	.030
	(.020)	(.017)	(.022)	(.021)	(.028)	(.020)

Note: All regressions include school fixed effects. Standard errors corrected for clustering at the school level. *, **, and ***, significant at the 10 percent, 5 percent and 1 percent, respectively.

Appendix B: Math tests in kindergarten, 1st grade, and 2nd grade

This Appendix describes the math tests we applied in kindergarten, 1st grade, and 2nd grade, and discusses issues of comparability, possible truncation of scores, and the construction of grade equivalents.

All of the tests were applied to a child individually (a single enumerator with a single child). In most tests, the enumerator would ask a child one question at a time, before moving on to the next question. Most tests also have a "stopping rule": When a child makes three mistakes in a row, or states that she does not know the answer to three questions in a row, that section of the test is stopped.

Description of tests: In each grade, we tested children on three domains: (1) *number recognition, addition, and subtraction*; (2) *number sense*; and (3) *word problems*.

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

Domain 2: *Number sense*: In kindergarten and 2nd grade, children were shown sequences of numbers; in each case, one number was missing, and children were asked to name the missing number. In 1st and 2nd grades, we used the number line test developed by Siegler and his coauthors (Siegler and Booth 2004; Siegler and Opfer 2003).[41] In 2nd grade, 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]

Test comparability: One issue that arises when making comparisons of math achievement as children age is that the math tasks which children should be able to carry out change as they progress through the school system.

[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 grade version of the test was similar, with a number line from 1 to 50. In both 1st and 2nd 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).

The fact that the content of tests changes raises difficult questions about comparability. The literature from developed countries suggests that boys outperform girls on certain math tasks (including visuospatial tasks, abstract problems, and material not covered in the school curriculum), but not on other tasks (including computations and, more generally, material that is covered in school). Thus, if a kindergarten test is weighted heavily towards the tasks on which girls generally perform well, while a 2nd grade test is weighted heavily toward tasks on which boys do well, it might appear that the boy-girl gap grows even if the observed increase in the gap is driven primarily by the change in the content of the test (more items that favor boys).[43]

There is no perfect solution to address this concern. In practice, we address it in three ways. First, in all three grades, children were tested on the same three domains, as discussed above, and each domain was given the same weight. In principle, this should help limit the extent to which changes in the material covered in the tests in different grades drives our results.[44] Second, we test the robustness of our results to alternative ways of aggregating the questions within each domain, and the weights given to each domain. Third, we make use of the fact that, for two domains, *number recognition, addition, and subtraction*, and *word problems*, there are some "anchor" questions that are the same across adjacent grades. As an additional robustness test, we reproduce our main results using only these common questions.

In Table B2, we present the results from estimating equation (1) in the paper, with four alternative aggregations of the tests.[45]

- Aggregation of individual answers within each domain done by IRT, total score given by sum of the three domains, with each domain receiving one-third of the total weight (column 1). This corresponds to the results we present in the main body of the paper.
- Score of each domain is given by simple count of correct answers, total score given by sum of the three domains, with each domain receiving one-third of the total weight (column 2).
- Aggregation of individual answers within each domain done by IRT, aggregation of scores for the three domains into total score by factor analysis (column 3).
- Aggregation of individual answers within each domain done by IRT, aggregation of scores for the three domains into total score with the procedure recommended by Anderson (2008) (column 4).[46]

[43] Gibbs (2010) argues that this explains much of the increase in the boy-girl gap in math achievement as children age in the United States.

[44] Of course, if the questions *within* a domain are more heavily weighted towards tasks that boys (or girls) excel at in one grade than in another one, this would also complicate comparisons of the boy-girl gap in achievement across grades.

[45] To keep the number of results manageable, we focus on the stacked regressions that include children in all three grades.

[46] The steps for constructing the 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 N*N matrix. Call that COV; (2) calculate the inverse of COV, call that INVCOV. That should also be a symmetric N*N 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 just 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.

The results in Table B2 below make clear that the coefficients we estimate are insensitive to alternative ways of aggregating the responses within each domain, and to alternative ways of aggregating the scores from the three domains. The coefficient on the dummy for girls ranges from -0.123 (0.018) to -0.139 (0.019), and the coefficient on the interaction between girls and mothers with university education ranges from 0.104 (0.050) to 0.113 (0.055).

As a final robustness check, we make use of the fact that, for the *number recognition, addition and subtraction* and the *word problems* domains, there are some questions that are common to the tests we applied in kindergarten and 1st grade, and others that are common to the tests we applied in 1st and 2nd grade. (There are no common questions for the second domain, *number sense*.) To test whether changes in the questions across grades have a substantive effect on our results, we first calculate new "total" math scores which are based on all of the questions that are available in the *number recognition, addition and subtraction* and *word problems* domains, and alternative total scores that are calculated only with the questions that are common to adjacent grades. If the increase in the boy-girl gap in achievement we observe as children age were mainly driven by changes in the questions asked in different grades, we would expect there to be substantive differences between the results that use all questions and those that are based only on the questions that are common across grades.

Table B3 shows that in all three grades, the coefficient on the indicator for girls is somewhat smaller than the estimates in the main body of the paper. This is not surprising, as the results in Table B3 do not include the *number sense* domain—precisely the domain where we find the biggest differences in the test scores of boys and girls, as shown in Table 4 in the main body of the paper. A comparison of the upper and lower panels in the table shows, however, that the results we obtain when we include all the questions in the *number recognition, addition and subtraction* and *word problems* domains are close in magnitude to those we obtain when we only consider the questions that are common across grades.

We conclude from these robustness checks that, although changes in the content of the tests are always a challenge for research that focuses on changes in achievement as children age, they are unlikely to be the main reason we observe an increase in the boy-girl achievement in our data.

Truncation of test scores: It is generally hard to develop appropriate math tests for very young children. Figure B1 below graphs the distribution of test scores, for boys and girls, separately by grade. It shows that there is some evidence that, for kindergarten, the distribution is truncated on the left. This occurs because our kindergarten test has floor effects: A substantial number of children could answer very few questions correctly.[47] The fact that the kindergarten test appears to be left-truncated, could lead us to underestimate the boy-girl gap in kindergarten, and to overestimate the increase in the gap between kindergarten and 2nd grade.

Construction of grade equivalents: We estimate "grade equivalents" to put the magnitude of the SES gradients and boy-girl differences in math achievement in context. These grade equivalents are calculated on the basis of the questions that are common across grades. Specifically, we proceed as follows:

[47] For example, the average child recognized only 3 of the 10 numbers on the number recognition test for kindergarten.

- Step 1: Limit questions to those that are common to kindergarten and 1st grade. Note that this only includes questions in the *number recognition, addition, and subtraction* and *word problems* domains.
- Step 2: Calculate the total kindergarten score using only common questions (as described in section on "test comparability" above).
- Step 3: Calculate the total 1st grade score using only common questions, but use the kindergarten data to standardize the 1st grade scores. Specifically, instead of subtracting the mean of the 1st grade score, we subtract the mean of the kindergarten score, and instead of dividing by the standard deviation of the 1st grade score, we divide by the standard deviation of the kindergarten score. The difference between the mean test score of children in kindergarten and 1st grade calculated in this way is one estimate of the grade equivalent for these two adjacent grades.
- Step 4: Carry out steps 1-4 but, instead of standardizing the 1st grade score with the kindergarten mean and standard deviation, standardize the kindergarten score with the 1st grade mean and standard deviation. The difference between the mean test score of children in kindergarten and 1st grade calculated in this way is an alternative estimate of the grade equivalent for these two adjacent grades.
- Step 5: Take average of two alternative procedures to calculating grade equivalents between kindergarten and 1st grade.
- Step 6: Carry out steps 1-4 to calculate grade equivalents between 1st and 2nd grade (with questions common to these two grades).

Based on these calculations, we estimate a grade equivalent of 1.80 standard deviations between kindergarten and 1st grade, and 1.39 standard deviations between 1st and 2nd grade.[48]

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.

Gibbs, Benjamin G. 2010. "Reversing Fortunes or Content Change? Gender Gaps in Math-Related Skill through Childhood." *Social Science Research* 39: 540-69.

Hill, Carolyn J., Howard S. Bloom, Alison Rebeck Black, and Mark W. Lipsey. 2008. "Empirical Benchmarks for Interpreting Effect Sizes in Research." *Child Development Perspectives* 2(3): 172-77.

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

[48] These values are somewhat larger than those typically found in the United States. For example, Hill et al. (2008) analyze six major standardized tests in math, and report an increase in average math achievement of 1.14 standard deviations between kindergarten and 1st grade, and 1.03 standard deviations between 1st and 2nd grade.

Appendix Table B1: Sample questions

Easy question	Hard question
Panel A: Number recognition, addition, and subtraction	
Kindergarten	
Easy 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?"	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?"
1st grade	
Easy 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?" Easy question: *addition and subtraction* 7+1	Hard question: *number recognition* Child is shown a sheet with 4 numbers (94, 200, 105, 513). Enumerator points at the number 105, and asks: "What number is this?" Hard question: *addition and subtraction* 11-1
2nd grade	
Easy question: *addition and subtraction* 20+10	Hard question: *addition and subtraction* 492+213
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. 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: *number sequences* Child is shown the sequence 530, 532, --, 536 and is asked to name the missing number. 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.
Panel C: Word problems	
Kindergarten	
Easy 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?"	Hard question: *word problems* 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* 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.	Hard question: *word problems* 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* Child is shown a picture with 6 buttons and is asked: "If you take away 3 buttons, how many are left?"	Hard question: *word problems* 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?"

Appendix Table B2: Robustness to alternative aggregations

	IRT, equal weights (1)	Simple count, equal weights (2)	IRT, factor aggregate (3)	IRT, Anderson aggregate (4)
Dummy: girls	-.131***	-.139***	-.136***	-.123***
	(.019)	(.019)	(.019)	(.018)
Dummy: Secondary school	.264***	.270***	.266***	.254***
	(.023)	(.023)	(.024)	(.022)
Dummy: University	.465***	.467***	.464***	.449***
	(.044)	(.046)	(.044)	(.043)
Interaction: University*girls	.106**	.113**	.108**	.104**
	(.052)	(.055)	(.051)	(.050)
F-test (p-value)	0.60	0.61	0.55	0.68

Note: Sample size is 31,398 in all regressions. All regressions include controls for child age in months and its square. Standard errors corrected for clustering at the school level. *, **, and ***, significant at the 10 percent, 5 percent and 1 percent, respectively.

Appendix Table B3: Using common questions across grades to test for effects of changes in test content

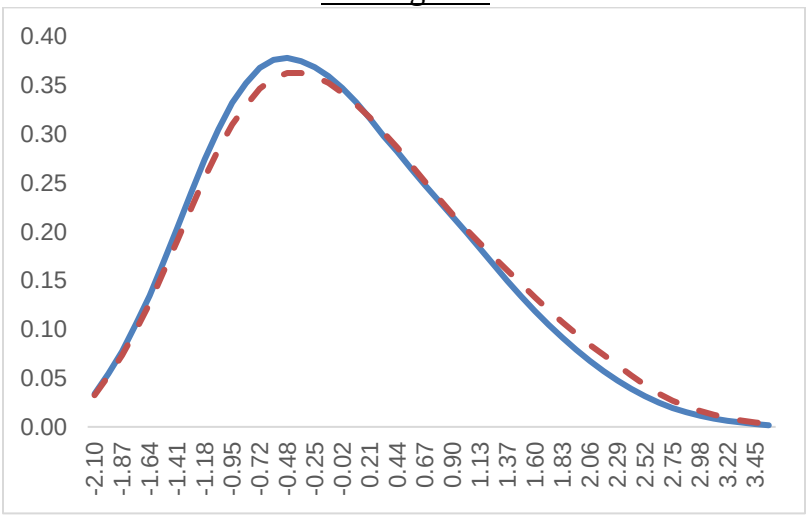
		All questions	
	Kindergarten	1st grade	2nd grade
Dummy: girls	-.049**	-.117***	-.114***
	(.021)	(.022)	(.021)
Dummy: Secondary school	.254***	.250***	.276***
	(.025)	(.028)	(.027)
Dummy: University	.429***	.433***	.534***
	(.052)	(.051)	(.053)
Interaction: University*girls	.055	.107**	.109*
	(.065)	(.059)	(.065)
F-test (p-value)	0.92	0.85	0.94

		Common questions only		
	Kindergarten	1st grade		2nd grade
		k-1	1-2	
Dummy: girls	-.051**	-.099***	-.135***	-.114***
	(.021)	(.023)	(.023)	(.021)
Dummy: Secondary school	.257***	.239***	.215***	.249***
	(.025)	(.027)	(.027)	(.027)
Dummy: University	.432***	.390***	.385***	.392***
	(.052)	(.049)	(.055)	(.055)
Interaction: University*girls	.059	.107*	.078	.157**
	(.065)	(.057)	(.065)	(.066)
F-test (p-value)	0.90	0.89	0.34	0.46

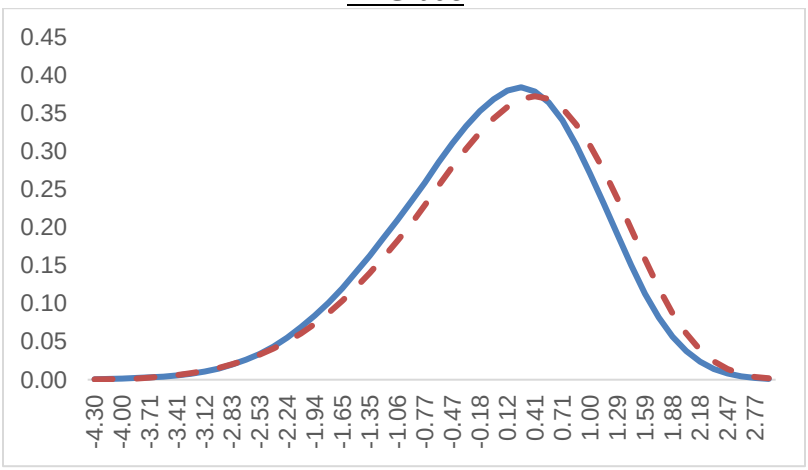
Note: Sample size is 10,466 in all regressions. All regressions include controls for child age in months and its square. There are two scores based on common questions for 1st grade—one based on questions that are common to the kindergarten and 1st grade tests, and another based on questions that are common to 1st and 2nd grade. Standard errors corrected for clustering at the school level. *, **, and ***, significant at the 10 percent, 5 percent and 1 percent, respectively.

Figure B1: Distribution of test scores for boys and girls, by grade

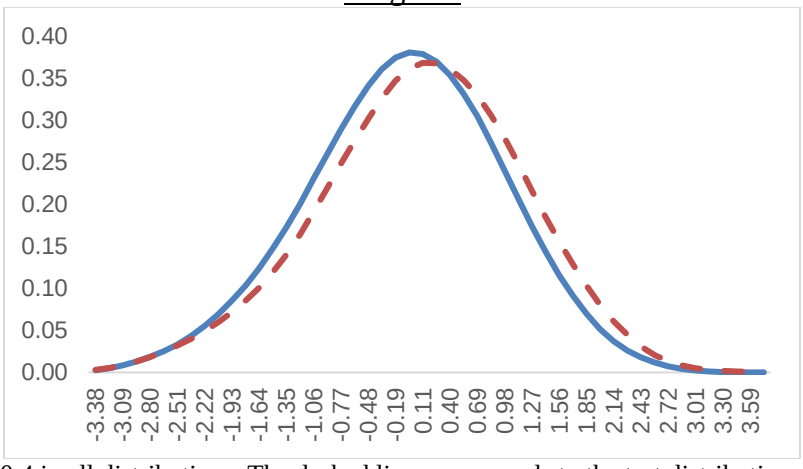
Kindergarten



1st Grade



2nd grade



Note: bandwidth is 0.4 in all distributions. The dashed line corresponds to the test distribution of boys, the solid line to girls.

Appendix C: Application of the CLASS in Ecuador

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

The *behaviors* that coders are looking for in each dimension are quite specific—see Appendix Table C1 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.[49]

[49] 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).

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' 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).[50]

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.[51]"

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 C2.[52] 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

[50] 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).

[51] 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).

[52] These and other results in this Appendix refer to kindergarten teachers. However, results for 1st and 2nd grade are qualitatively very similar.

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 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 C3. 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.

Appendix Table C1: CLASS scores for Behavior Management dimension

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

Source: Pianta et al. (2007).

Appendix Table C2: 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.

Appendix Table C3: 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