cemmap
centre for microdata methods and practice

When promising interventions fail: personalized coaching for teachers in a middle-income country

Pedro Carneiro
Yyannu Cruz-Aguayo
Ruthy Intriago
Juan Ponce
Norbert Schady
Sarah Schodt

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP05/22

UKRI Economic and Social Research Council

When Promising Interventions Fail: Personalized Coaching for Teachers in a Middle-Income Country[1]

Pedro Carneiro*
Yyannu Cruz-Aguayo†
Ruthy Intriago‡
Juan Ponce‡
Norbert Schady°
Sarah Schodt°°

Abstract

Children in developing countries have deep deficits in math and language. Personalized coaching for teachers has been proposed as a way of raising teacher quality and child achievement. We designed a coaching program that focused on one aspect of teacher quality—teacher-child interactions—that researchers in education and psychology have argued is critical for child development and learning. We implemented the coaching program in Ecuador, with 100 1st grade teachers randomly assigned to treatment and 100 to control. Coaching improved the quality of teacher-child interactions but reduced child achievement. Our results underline the importance of evaluating new forms of professional development for teachers, even those that follow best practice, before these interventions are taken to scale.

* University College London, CEMMAP, IFS, and FAIR-NHH
† Inter-American Development Bank, Washington, D.C.
‡ FLACSO, Quito, Ecuador
° World Bank, Washington, D.C.
°° Independent consultant, Washington, D.C.

[1] We thank Alejandra Campos and Carolina Echeverri for outstanding research assistance, and Jennifer LoCasale-Crouch, Bob Pianta, Hiro Yoshikawa, two referees and the editor for very helpful comments. Carneiro gratefully acknowledges the support of the ESRC for CEMMAP (ES/P008909/1) and the ERC through grant ERC-2015-CoG-682349.

1. Introduction

Teachers are the most important input into the production of learning within schools. Improving teacher quality is a central goal of policymakers in developed and developing countries. In this paper, we evaluate an innovative program that sought to improve teacher effectiveness in a sample of schools in Ecuador, a middle-income country in South America.

Our paper is motivated by four observations from the recent literature on teachers. First, there is substantial variation in teacher quality, even within the same schools, and these differences have important consequences for achievement, college attendance, and labor market outcomes.[2] However, there is considerable uncertainty about policies that can increase the effectiveness of current teachers—see Jackson, Rockoff, and Staiger (2014) and Fryer (2017) for a discussion focused on the U.S., and Evans and Popova (2016), Gaminian and Murnane (2016), and Glewwe and Muralidharan (2016) for evidence from developing countries.

Second, most countries, both developed and developing, spend substantial resources to improve teacher quality. The U.S., for example, spends an estimated $18 billion a year on in-service training for teachers (Education Next 2018). A recent paper (Loyalka et al. 2019) reports that between 2012 and 2017, India's national government allocated US $1.2 billion to teacher professional development programs, and that teachers in Mexico spend an average of 23 days in professional development each year. Yet much of this in-service training is thought to be ineffective because it does not give teachers actionable guidance on how to improve their teaching practices (Popova, Evans, and Arancibia 2016; Popova et al. 2018).

Third, given the perceived shortcomings of traditional in-service training for teachers, policymakers have begun to experiment with alternative approaches, in particular coaching by expert teachers. Coaching has been identified as a promising way of improving classroom quality and learning for young children (Yoshikawa et al. 2013; Evans and Popova 2016), and a recent meta-analysis of coaching programs in the U.S. and other developed countries finds pooled effect sizes of 0.49 SDs on instruction and 0.18 SDs on achievement (Kraft, Blazar, and Hogan 2018). The Biden Administration's American Families Plan, which aims to provide universal preschool to all 3- and 4-year-old children, specifies that this must be accompanied by regular "job-embedded coaching for teachers" (Weiland and

[2] For evidence on teacher effects in developed countries see Rivkin, Hanushek, and Kain (2005), Chetty, Friedman, and Rockoff (2014a) and the reviews in Hanushek and Rivkin (2012), Jackson, Rockoff, and Staiger (2014), and Koedel, Mihaly, and Rockoff (2015). For evidence from developing countries see Araujo et al. (2016) and Bau and Das (2020). For evidence on long-term effects of teacher quality in the U.S., see Chetty et al. (2011) and Chetty, Friedman, and Rockoff (2014b).

Yoshikawa 2021, p. 1). However, the evidence base from developing countries on these more innovative, personalized forms of in-service training for teachers is still thin.

Finally, much recent literature in education and child psychology has emphasized the importance of *interactions* between teachers and children, especially in preschool and the early years of elementary school. Indeed, as Perlman et al. (2016) write, the focus on interactions is "driven by some of the most fundamental theories of developmental psychology", including "attachment theory, Ecological Systems theory's focus on the child's interactions with his/her most immediate environment, and Vygotsky's emphasis on learning through social exchanges by supportive 'experts'". As a result, a handful of pilots that seek to improve the quality of teacher-child interactions have been implemented, especially in the U.S.

With these insights in hand, we designed, implemented, and evaluated a coaching program for 1st grade teachers in Ecuador. The program was directly based on two coaching programs for teachers in the U.S., *Making the Most of Classroom Interactions* (MMCI) and *My Teaching Partner* (MTP). We worked closely with the creators of these programs at the University of Virginia and with officials at the Ministry of Education in Ecuador to adapt them to the Ecuadorean context.

The coaching intervention we study provided 1st grade teachers with bi-weekly, personalized coaching. It had what has been argued are critical elements for success: It was *focused* on a particular determinant of learning—namely, the nature and quality of teacher-student interactions, and how to improve them; it was *semi-structured*, using a curriculum that gave teachers a framework to think about classroom quality, but also providing concrete recommendations to improve classroom practices; and it was *personalized* to what coaches observed in the classrooms of individual teachers.

Coaches were teachers who had been nominated by headmasters and fellow teachers in their schools. They received two weeks of full-time training before the intervention began and were then taken out of their classrooms for a year so that they could work full-time as coaches. The program worked on a two-week cycle and lasted for a full school year. Each teacher in the treatment group received 13-14 personalized feedback sessions from her coach. Earlier research from the MTP coaching program in the U.S. suggests that 8-12 biweekly coaching cycles are the necessary dosage to change teacher behaviors (Downer et al. 2009; Pianta et al. 2014).

We evaluate the impact of the coaching program on the quality of teacher-child interactions, and on achievement in math and language. To measure interactions, we filmed teachers teaching for a full day, and coded the video footage with a much-used classroom observation tool, the Classroom Assessment Scoring System (CLASS; Pianta, LaParo, and Hamre 2007). The CLASS measures the quality

of teacher-child interactions in three broad domains: *Emotional Support*, *Classroom Organization*, and *Instructional Support*.

We first show that, at the end of 1st grade, teachers in the treatment group had higher-quality interactions with their children, about 0.26 SDs. We find, however, that these improvements in classroom quality did not translate into higher achievement: The point estimates from regressions of test scores on an indicator of random assignment to mentoring are generally *negative*, in some cases significantly so.

We discuss various possible explanations for our results. First, as we show, coaching primarily improved *Emotional Support*, a dimension of quality that is only weakly correlated with achievement. Thus, the coaching intervention improved aspects of teaching practices that were unlikely to raise test scores, at least in the short run. Second, coaching may have taken time away from other activities carried out by teachers (like lesson planning). Third, it may be that encouraging teachers to teach in a way that was unfamiliar to them improved teacher-child interactions but disrupted the learning process. Possibly, more time is needed for teachers and children to adjust to a new pedagogical approach.

Our paper contributes to a literature on professional development for teachers. It shows that more innovative—but also more expensive—forms of professional development, like coaching, may not always have the expected results, even when they have a strong foundation in theories of child learning, are carefully designed, and are faithfully implemented.

The coaching program we analyze required teachers to be responsive and adjust to the needs of individual children in their classroom. Pedagogical approaches that are more child-centered are easier to implement where classroom sizes are relatively small and teacher capabilities are high. However, this is rarely the case in developing countries: While average class size in elementary school in OECD countries is 15, it was 31.5 in the 1st grade classrooms in our sample, and is 50 or more, on average, in nine countries in Sub-Saharan Africa.[3] We speculate that, under these circumstances, coaching may be more effective when it focuses on a particular child outcome, gives teachers specific guidance on the steps needed to improve that outcome, and provides teachers with complementary material to guide them as they adjust their classroom practices.

To illustrate this point, we contrast our results with those from two recent evaluations of coaching programs in developing countries. Yoshikawa et al. (2015) analyze *Un Buen Comienzo*, a coaching program for pre-k and kindergarten teachers in Chile. The program focused on the quality of teacher-

[3] Specifically, teacher-student ratios in elementary school are 50 in Angola, 51 in Tanzania, 52 in Guinea-Bissau, 55 in Mozambique and Ethiopia, 57 in Chad, 59 in Malawi, 60 in Rwanda, and 83 in the Central African Republic (World Bank 2021). These are national averages, and hide substantial variability, so it is not unusual for some teachers to have 80 or more children in their classrooms.

child interactions but gave teachers little guidance on specific instructional or classroom practices. The authors show that *Un Buen Comienzo* improved teacher-child interactions, as measured by the CLASS, but did not raise child development or achievement.

Cilliers et al. (2020) analyze a coaching intervention (as well as, separately, more traditional in-service training) for 1st and 2nd grade teachers in a sample of South African schools. The program sought to change how reading was taught to young children. Specifically, it encouraged teachers to switch from reading out loud in front of the class to group reading by the children themselves. The coaching intervention also provided fully scripted lesson plans, encouraged teachers to group children by ability, and promoted frequent assessments of children. Like our paper, and like Yoshikawa et al. (2015), Cilliers et al. (2020) test for changes in classroom practices, but these were practices *related to reading*—for example, the frequency of group reading—rather than teacher-student interactions more broadly defined, as measured by the CLASS. The authors show that the intervention changed classroom practices and significantly improved reading outcomes.

Our paper also adds to a recent literature in economics on the challenges inherent in replication and scale-up of promising interventions (see the collection of papers in List, Suskind, and Supplee, 2021, which focus on interventions for young children; also, Al-Ubaydli, List, and Suskind 2017; Banerjee et al. 2017).

The rest of the paper proceeds as follows. We describe the setting, the coaching intervention, and data in section 2. Section 3 presents results on the cross-sectional association between the CLASS and achievement. We describe our identification strategy in section 4. Our main results on the effects of coaching are in section 5, and we conclude in section 6.

2. Setting, pilot, and data

A. Setting and intervention

Ecuador is a middle-income country in South America. The elementary school cycle runs from kindergarten to 6th grade. The overwhelming majority of children attend public (rather than private) schools. Enrollment in elementary school is essentially universal. The key educational challenge in elementary school is quality: On an international test of 3rd graders, 38 percent of children in Ecuador had the lowest of the four levels of performance on math, very similar to the average for the 15 countries in Latin America that participated in the test (40 percent), but substantially more than higher-performing countries like Costa Rica (18 percent) or Chile (10 percent) (Berlinski and Schady 2015).

Our study took place in 198 schools in the province of Pichincha, the most-densely populated province in the highlands region of Ecuador. We selected a sample of 10 coaches through a process that

included nomination by headmasters and peers, and performance on tests that assessed how suitable teachers would be as coaches.

Specifically, the selection process had three stages. First, the Ministry of Education identified 115 potential coaches in 33 schools in Pichincha province; all potential coaches were tenured and had worked as teachers in k-3rd grade in the last 5 years. Next, we had these teachers take a test ("Ideas about Children") that was meant to assess whether they were receptive to the notion that children, not teachers, should be at the center of the learning process. At this stage, we also asked all other teachers in the 33 schools to name 3 teachers they would go to if they needed guidance on some aspect of teaching practice and calculated the proportion of all votes cast in a school that were cast for each potential coach. Similarly, we asked school principals to identify the 3 teachers in their school who had the most potential to serve as coaches. On the basis of these three pieces of data—score on the test, recommendations from peer teachers, and recommendation by the principal—we calculated an aggregate score for each potential coach.

We invited 24 teachers with the highest scores to a 3-day training on the CLASS, and scored them on class participation, comprehension of the material, and fidelity in scoring a sample of videos according to the CLASS. With this information, we selected 10 coaches, all of whom accepted the offer of employment for a year.

Once coaches had been selected, we matched each coach to the 20 elementary schools that were closest to her place of residence (excluding her own school) and, within these schools, we randomly assigned 10 schools to treatment and 10 to control. In each treated school, the coach was assigned one 1st grade teacher. Control schools continued business as usual. In total, there were 100 1st grade teachers in the treatment group, and 98 in the control group.[4]

Coaches received two weeks of training administrated by expert CLASS trainers, some of whom had also participated in the MMCI and MTP programs that were the basis of the Ecuador coaching pilot. In turn, teachers assigned to the coaching treatment received one week of training on the general framework of the CLASS, with a focus on the importance of teacher behaviors and teacher-student interactions.

The coaching pilot was semi-structured. It followed a two-week cycle. Every cycle focused on a particular topic—for example, how to give feedback to children. Each teacher in the treatment group was recorded teaching for a full day every two weeks. The coach reviewed the video, and looked for specific moments that showed desirable behaviors, as well as those that could be improved—focusing

[4] When there was more than one 1st grade teacher in a school, one was selected at random. The original sample was 200 schools, but two control schools dropped out, so our final sample is 198 1st grade teachers.

mainly, but not exclusively, on the topic for that cycle. She also compared these video clips with selected videos from a large "library" of videos from Ecuador that had been prepared for this purpose. Coach and teacher then had an in-person meeting, viewed the video and the relevant clips from the library together, and agreed on concrete actions that the teacher could take in her classroom.

B. Data

We use the CLASS (Pianta, LaParo, and Hamre 2007) to measure the quality of teacher-child interactions. The CLASS is based on developmental and education theories that argue that the daily interactions between teachers and children are the "primary engine" for child development and learning in preschool and early elementary school (Hamre and Pianta 2007; Leyva et al. 2015).

The CLASS measures teacher behaviors in three broad domains: *Emotional Support*, *Classroom Organization* and *Instructional Support*. The behaviors that coders are looking for in each dimension are quite specific—Appendix Table A1 in Araujo et al. (2016) gives an example. For each of these behaviors, the CLASS protocol gives coders concrete guidance on whether the score given should be "low" (scores of 1–2), "medium" (3–5), or "high" (6–7).

The CLASS has been widely used both for research and policy purposes in the U.S. Head Start grantees need a minimum score on the CLASS to be re-certified for funding, and several states have integrated the CLASS into their quality rating and improvement systems. The CLASS has also been used as a measure of teacher quality in research on several Latin American countries, including Chile (Leyva et al. 2015; Yoshikawa et al. 2015; Bassi, Meghir, and Reynoso 2020), Ecuador (Araujo et al. 2016; Campos et al. 2020), and Peru (Araujo, Dormal, and Schady 2019).

We carefully followed CLASS protocols to code the videos of treated and control teachers recorded at the end of 1st grade.[5] Specifically, each day of film was cut into 20-minute segments. We took the first four segments, and each segment was coded by two separate coders. Coders were blinded to treatment status. The correlation in the scores given by different coders is high—the inter-coder reliability ratio is 0.84, on average. On the other hand, there is more variation in the CLASS scores given to different segments from the same day—the inter-segment reliability ratio between the 1st (earliest) and 4th (latest) segments that were coded is only 0.35, on average. This pattern of results suggests there is substantial measurement error in the CLASS, principally because the behaviors teachers exhibit at any time are not a perfect measure of the behaviors they engage in over the course of a day, leave alone over the course of a school year.

[5] The videos of teachers in the treatment group filmed every two weeks were not formally coded with the CLASS. Thus, we do not have data on the evolution of teaching practices between the beginning and the end of the school year for teachers who received coaching.

In Table 1, we summarize the characteristics of schools (Panel A), teachers (Panel B) and children (Panel C) in our sample. Panel A shows that the average school in the sample had 4.7 teachers between kindergarten and 3rd grade. These values, as well as the proportion of teachers in different pay grades in the salary scale, are similar in treatment and control schools. Panel B shows that essentially all (98 percent) of teachers are women, and most (90 percent) are tenured. On average, teachers have almost 16 years of experience and 31 children in their classrooms. Panel C shows that half of the children in the sample are girls, and average age (measured at the end of the grade, when children took the achievement tests) is 7 years. There are small differences in child age and gender by treatment status: Children in the treatment group are about 0.7 months older than those in the control group and are 4 percentage points more likely to be female. In our estimates, we control for all school, teacher, and child characteristics in Table 1.

We provide further details on the CLASS in the Appendix. Figure A1 graphs univariate densities of the distribution of CLASS scores for teachers in the control group, by domain. The figure shows that CLASS scores are highest in *Classroom Organization*, with teachers distributed in the "medium" and "high" parts of the distribution; somewhat lower in *Emotional Support*, with most teachers in the "medium" range; and lowest in *Instructional Support*, where all teachers have "low" CLASS scores. There are clearly floor effects for *Instructional Support*, but not the other domains. On average, the CLASS scores in this sample are somewhat higher than those found in a nationally representative sample of kindergarten classrooms in Ecuador, but substantially lower than those generally found in U.S. settings (Araujo et al. 2016). Table A1 shows that CLASS scores across domains for the same teachers are positively correlated, with correlations that range from 0.43 for *Emotional Support* and *Instructional Support* to 0.69 for *Emotional Support* and *Classroom Organization*. The fact that the correlation between teacher scores on the different domains of the CLASS are far from unity likely reflects a combination of factors: Different teachers may genuinely excel in the behaviors in different domains, but measurement error would also tend to reduce the magnitude of the correlations.

Within each classroom, we used a random number generator to select a sample of 20 children who would be tested at the end of the year.[6] To measure achievement, we applied two language and two math tests. The language tests were a test of letter and word recognition and a test of receptive vocabulary, while the math tests were a test of number recognition and a test of simple addition and subtraction. To measure receptive vocabulary, we use the *Test de Vocabulario en Imagenes Peabody* (TVIP)

[6] With 198 classrooms and 20 children per classroom, our target number of children with test score data was 3,960. However, there were 9 classrooms that had fewer than 20 children in total. Accounting for these classrooms results in 3,909 children. Our final sample, 3,751 children with valid test score data, is 96 percent of the intended sample.

(Dunn et al. 1986), the Spanish-speaking version of the much-used Peabody Picture Vocabulary Test (PPVT). The TVIP has been used widely to measure development among Latin American children—see Paxson and Schady (2007) for a comparison of vocabulary scores between children in Ecuador and the U.S., and Schady et al. (2015) for evidence on levels and socioeconomic gradients in the TVIP in five Latin American countries, including Ecuador. The other three tests were taken from the Woodcock-Johnson battery of achievement tests (Woodcock and Muñoz-Sandoval 1996) and have been applied in other evaluations of interventions in settings similar to ours, including in Ecuador (Paxson and Schady 2011; Araujo et al. 2016).

Unsurprisingly, performance on the four tests is positively correlated in our data. The lowest correlation, 0.34, is between vocabulary and number identification, and the highest, 0.64, is between number identification and basic arithmetic. As with the CLASS, the magnitude of the correlations likely reflects that different tests measure different dimensions of knowledge or achievement, but also measurement error in the tests.

3. Cross-sectional associations between CLASS and achievement

To motivate our analysis, we first calculate the associations between the CLASS and child test scores. Panel A of Table 2 reports the results from regressions of overall achievement on the CLASS, with the sample limited to children in the control group. Column (1) shows that a 1 SD increase in the CLASS is associated with a 0.097 SD increase in test scores (p-value: <0.001). In the following columns we report the association between individual domains of the CLASS and achievement. These columns, and in particular the specification in column (5), which includes all three CLASS domains at the same time, show that *Classroom Organization* is most strongly associated with test scores. In contrast, the association between *Emotional Support* and achievement is not significant once other CLASS domains are included as controls.

To further explore the associations between the CLASS and achievement, we next make use of data from another experiment, carried out in a different sample of 200 schools in Ecuador (see Araujo et al. 2016; Campos et al. 2020). In that experiment, children were randomly assigned to classrooms within schools in kindergarten and were then reassigned to different classrooms in every grade between 1st and 6th grades. Much as in the coaching pilot, teachers were filmed, and the video was used to calculate teacher CLASS scores. At the end of each grade, children were given a large battery of age-appropriate tests in math and language. In 1st grade, these tests included the same math and language tests as were applied in the coaching pilot (as well as other tests). For the calculations we report below, we limit the sample to teachers and children in 1st grade and use only those tests that were also applied in the coaching pilot, scored and aggregated in the same way.

We first use these data to run similar regressions to those reported in Panel A. The results in Panel B show that the association between the CLASS and achievement in these data is similar to that which we estimate in the control group of the coaching pilot: A 1 SD increase in the CLASS is associated with a 0.082 SD increase in achievement. When we look at different domains of the CLASS simultaneously, only *Classroom Organization* is consistently and significantly associated with test scores.

The results in Panels A and B of Table 2 are associations, not necessarily causal effects. Indeed, if better-off children attend higher-quality schools, as seems likely, the coefficients from these cross-sectional regressions would overstate the importance of teacher behaviors for achievement. In the multi-grade experiment, however, there was random assignment of children to classrooms within schools. As a result, estimates that use only the within-school, cross-classroom variation in the CLASS and achievement are more likely to have a causal interpretation (see Campos et al, 2020, for a detailed analysis of these data).

In Panel C of Table 2, we report the results from regressions of achievement on the CLASS, including school fixed effects. In these school fixed effects regressions, teacher CLASS scores should not be correlated with the observable or unobservable characteristics of the students in their classrooms, although the CLASS could still be correlated with other teacher attributes that affect test scores, as discussed in Araujo et al. (2016). The coefficient on the CLASS in these fixed effects regressions is somewhat smaller than in Panel B—0.062, rather than 0.082. Importantly, however, the last column of Panel C shows that—much as is the case in the results without school fixed effects—only *Classroom Organization*, not *Emotional Support* or *Instructional Support*, consistently predicts achievement.

4. Identification strategy

To estimate the impacts of the coaching program on teaching practices, we run regressions of the following form:

$$CLASS_{tsb} = \delta_b + \rho T_{sb} + \theta_1 Z_{tsb} + \theta_2 X_{sb} + u_{tsb} \quad (1)$$

The dependent variable is the CLASS (or one of its domains) of teacher *t* in school *s* and block *b*, where the blocks refer to the groups of 20 schools that were the basis for the block randomization; δ_b is a set of block fixed effects; T_{tsb} takes on the value of 1 for teachers (in schools) randomly assigned to the coaching intervention, 0 otherwise; Z_{tsb} and X_{sb} are the teacher and school controls in Table 1; and u_{tsb} is the regression error term. The coefficient of interest is ρ. To test for the possibility of differences in effects at the top and bottom of the distribution, we also run regressions in which the dependent variable is an indicator variable which takes on the value of 1 if the CLASS is below the 10th or 20th percentiles, or above the 80th or 90th percentiles, respectively (four separate regressions).

We proceed in a comparable fashion to estimate treatment effects on child achievement. The estimating equation is now:

$$Achievement_{itsb} = \delta_b + \rho T_{isb} + \theta_1 W_{itsb} + \theta_2 Z_{tsb} + \theta_3 X_{tsb} + u_{itsb} \quad (2)$$

where $Achievement_{itsb}$ is achievement on a given test or total achievement, and W_{itsb} consists of child gender and age and its square. Here too we run regressions in which the dependent variable is total achievement or achievement on one of the four tests. As with the CLASS, we also test for the possibility of differences in effects at the top and bottom of the distribution.

Finally, to gain some understanding of the mechanisms whereby coaching affects achievement, we run regressions of the following form:

$$Achievement_{itsb} = \delta_b + \rho_1 T_{isb} + \rho_2 CLASS_{tsb} + \theta_1 W_{itsb} + \theta_2 Z_{tsb} + \theta_3 X_{tsb} + u_{itsb} \quad (3a)$$

as well as:

$$Achievement_{itsb} = \delta_b + \rho_1 T_{isb} + \rho_2 CLASS_{tsb} + \rho_3 (CLASS_{tsb} * T_{isb}) + \theta_1 W_{itsb} + \theta_2 Z_{tsb} + \theta_3 X_{tsb} + u_{itsb} \quad (3b)$$

Equation (3a) is in the spirit of standard mediation analysis. It estimates, under strong assumptions, whether any effect of coaching on achievement can be accounted for by its effect on teacher-student interactions. In equation (3b) we add the interaction between the CLASS and treatment—see Imai, Tingley, and Yamamoto (2013) and Huber (2020) for discussion. In this regression, the coefficient ρ_3 estimates whether the slope of the association between the CLASS and achievement is different in treatment and control groups.

In all regressions, we normalize the CLASS and each of its domains to have zero mean and unit standard deviation. We follow the same procedure with the individual achievement test, and also calculate a measure of total achievement, which gives one-quarter weight to each test. As with the individual tests, total achievement is normalized to have zero mean and unit standard deviation. All regressions are estimated by OLS. Standard errors in (2) and (3) adjust for clustering at the school level.

5. Impacts of coaching on teacher-student interactions and achievement

To motivate our regression results, in Figure 1 we graph the univariate densities of the CLASS (Panel A) and achievement (Panel B) in treatment and control groups. Panel A shows that, relative to the control group, the distribution of CLASS scores for teachers who received coaching is shifted to the right. The biggest difference between the two distributions appears to be in the right tail. Panel B shows a smaller difference in achievement between treatment and control groups, although the distribution of test scores of children in classrooms where the teacher received coaching appears to be shifted to the left.

Our main results are in Table 3. Panel A, which corresponds to equation (1), shows that teachers who were randomly assigned to receive coaching had higher end-of-grade overall CLASS scores, with an effect size of 0.26 SDs (p-value: 0.06). Columns (2) through (5) show that the effects on coaching are concentrated at the top of the distribution: Teachers who received coaching are not significantly less likely to be below the 10th percentile of the distribution (coefficient of -0.034, with a standard error of 0.041), but *are* significantly more likely to be above the 90th percentile (coefficient of 0.089, with a standard error of 0.041). Columns (6) through (8), finally, show that coaching effects are only significant for one CLASS domain, *Emotional Support*, where the regression coefficient is 0.32 (p-value: 0.02). Treatment effects on *Classroom Organization*—the domain most strongly associated with test scores, as seen in Table 2—are small and insignificant: the coefficient is 0.11 (p-value: 0.42).

Panel B of Table 3 corresponds to equation (2). It shows that children in classrooms of teachers who received coaching have lower overall achievement, -0.069 SDs (p-value: 0.04). The negative effects of coaching on test scores are concentrated at the bottom of the distribution: Children in classrooms where teachers received coaching are significantly more likely to be below the 10th percentile of the distribution of achievement (coefficient of 0.017, with a standard error of 0.010), but are no less likely to be above the 90th percentile (coefficient of -0.007, with a standard error of 0.010). Columns (6) through (9) show that the coaching effects are larger for language than for math. However, they are not driven by a single test—rather, there are negative, significant (or borderline significant) impacts of coaching on three of the four individual tests we applied.

In sum, the results to this point show that coaching significantly improved the quality of teacher-child interactions but, surprisingly, significantly lowered test scores, especially at the bottom of the distribution. We now turn to possible explanations for this result.

We begin by discussing two features of the way in which the data for the CLASS were collected. First, teachers were filmed towards the end of the academic year, so the changes in behaviors we observe among teachers who received coaching may be relatively recent—too recent to affect child achievement. Second, it is possible that teachers in the treatment group, who knew what behaviors were "expected" of them, "acted" for the camera on the unannounced day on which they were filmed—a Hawthorne effect.[7] Both of these features of the data could mean that we may overstate the extent to which the coefficient on the CLASS in equation (2) reflects true changes in teacher behaviors over the course of the school year. On their own, however, they cannot account for the significantly *negative* effect of coaching on test

[7] We do not believe that Hawthorne effects are likely to be important. In practice, the difference in CLASS scores between teachers in the treatment and control groups is very similar early on the day—when they were arguably more likely to be acting for the camera—as later in the day (see also Araujo et al. 2016, Appendix B, for a discussion).

scores. Rather, our findings suggest that the coaching intervention disrupted learning in some way, and that this disruption was not picked up by the CLASS.

To further explore this, we first turn to estimates of equation (3a). Columns (1) through (5) of Table 4 show—unsurprisingly, given the positive effect of coaching on the CLASS and the negative effect on test scores—that including the CLASS as an additional ("mediating") variable in the achievement regression increases the negative coefficient on coaching in absolute value (in the regression in which we control for the overall CLASS the coefficient on treatment is -0.079, with a standard error of 0.034).

In columns (6) through (10), we include the interactions between the CLASS and treatment, as in equation (3b). These results show there are remarkable differences between treatment and control groups in the association of the CLASS and test scores. Column (6) shows that at all levels of the CLASS, children in classrooms in which the teacher received coaching had lower achievement—the estimate of ρ_1 is -0.083 (with a standard error of 0.034), very close to the value in column (1). The results in column (6) show, however, that there is also a difference in slopes—indeed, the association between achievement and the CLASS in the treatment group, given by the sum of ρ_2 (0.103, with a standard error of 0.027) and ρ_3 (-0.105, with a standard error of 0.034), is essentially zero. In other words, in the treatment group, children in classrooms of high- CLASS teachers did not learn more than those in classrooms of low- CLASS teachers.

Much the same pattern can be observed for individual domains of the CLASS: In columns (7) through (9) estimates of ρ_1 are always negative and significant, as are estimates of ρ_3. For example, in the case of *Emotional Support*, ρ_1 is -0.076 (with a standard error of 0.034), and ρ_3 is -0.070 (with a standard error of 0.035).[8]

What could account for these effects of coaching? Coaching may have crowded out time that teachers spent on other tasks. Teachers and coaches only met for one hour every two weeks, at the end of the school day, but coaches also gave teachers "assignments" they were meant to complete by their next meeting. It is possible that time on meetings and assignments was time that the teachers would otherwise have spent on other teaching-related activities—for example, lesson planning.

It is also possible that changes in teaching practices disrupted learning, at least temporarily. Indeed, it is likely that teachers who took the coaching intervention most seriously spent more time

[8] We include the estimates in column (5) for completeness, but they are hard to interpret. They suggest that the biggest changes in the association between the CLASS and achievement took place in *Instructional Support*: Controlling for coaching effects on *Socio-Emotional Support* and *Classroom Management*, classrooms in the treatment group saw both downward shift in the intercept, -0.067 (with a standard error of 0.034), and a reduction in the slope for *Instructional Support*, -0.124 (with a standard error of 0.040).

preparing assignments and made more changes to their in-class behaviors than other teachers. These teachers may also have had the largest disruptions in their classrooms, and this, in turn, could account for the fact that the CLASS does not predict achievement among children in the treatment group.

We conclude with two important caveats to our results. First, it could be that it takes time for teachers and children to adapt to a new pedagogical approach—teachers who received coaching could be less effective initially, but more effective eventually. In that case, we might expect to see improvements in achievement when teachers in the treatment group teach the *next* cohort of 1st grade students, perhaps especially if these teachers had received coaching for a second year.

Second, even if coaching lowered achievement in the short run, changes in teacher behaviors may have led children to acquire other skills—for example, higher levels of *Emotional Support* could raise child self-esteem or foster a growth mindset. Plausibly, improvements in unobserved child skills, in turn, could raise child test scores in subsequent years. Indeed, a number of interventions in early childhood have yielded zero or very small effects on cognition and achievement, but have resulted in substantial improvements in outcomes in adulthood—see Garcia et al. (2020), Conti, Heckman, and Pinto (2016), and Heckman, Pinto, and Savelyev (2013) for an analysis of two influential programs in the U.S., and Chetty et al. (2011), who find that the effects of higher classroom quality on test scores in early elementary school fade out quickly, but lead to improvements in wages and other outcomes in adulthood. Unfortunately, we cannot investigate these possibilities because we do not have data on outcomes other than achievement, or on achievement in subsequent years.

6. Conclusion

Learning outcomes of young children in many developing countries are low. By the middle of elementary school, a large fraction of children cannot read, and cannot do basic math operations like single-digit addition. Four out of five students in Mozambique and Nigeria cannot read a simple word of Portuguese and English, respectively, after more than three years of compulsory language learning. In India, only one in four 4th grade students manages tasks—such as basic subtraction—that are part of the curriculum for 2nd grade (Bold et al. 2019), and 31 percent of children in 3rd grade cannot recognize basic words (Kremer, Brannen, and Glennerster 2013). In Latin America, two-thirds of children do not achieve the minimum levels of literacy expected for their age (Busso et al. 2017). Researchers and policymakers have struggled for decades to find ways of raising achievement, including with policies that seek to improve the skills of teachers who are currently in service. Interventions that provide teachers with practical, personalized tools to improve teaching practices, including teacher coaching programs, have been identified as particularly promising.

In this paper, we analyze one such program in Ecuador. The intervention we study provided 1st grade teachers with bi-weekly, personalized coaching using a semi-structured curriculum. The content of the program drew on theories in developmental psychology and best practice, and was faithfully implemented. It had the elements of what has been identified as best practice: The World Bank *World Development Report* on education, for example, states that "to be effective, teacher training needs to be individually targeted and repeated, with follow-up coaching, often around a specific pedagogical technique" (World Bank 2018, p. 131).

After one year, coaching had a modest, positive effect on the quality of teacher-student interactions, albeit primarily in one domain, *Emotional Support*. On the other hand, the program did not raise achievement. Indeed, as we show, the program had a significant negative effect on test scores, especially at the bottom of the distribution. It is possible that the meetings between coaches and teachers crowded out other valuable teacher activities, like lesson planning, or that the changes in pedagogical practices introduced by coaches disrupted learning, at least in the short run.

Recent research on preschool in the U.S. suggests that coaching for teachers may be most effective when it focuses on a single domain of learning—math, language, literacy, socio-emotional skills—rather than on improving classroom quality more broadly (see the discussion and references in Weiland and Yoshikawa 2021). Our results, as well as those from coaching programs for teachers implemented in Chile (Yoshikawa et al. 2015) and South Africa (Cilliers et al. 2020), are consistent with this observation. We also speculate that, in settings where teacher quality may be low and the number of students per teacher is high, coaching programs may have to be highly structured, giving teachers concrete tools like scripted lessons and learning assessments, to guide them through implementation. We believe that this is an important area for future research.

Personalized coaching tends to be expensive, especially if it is in-person, rather than web-based. The intervention we study cost well over 10 percent of the salary of the average teacher.[9] In order to have a positive benefit-cost ratio, coaching programs would likely need to have substantial, positive effects on achievement in the short run, or wages in the long run. The results in this paper, and our read of the evidence from other developing countries, suggests that caution is in order before coaching programs for teachers of young children, perhaps especially programs that focus on improving teacher-child interactions, are taken to scale. They also underline the difficulties inherent in translating policies from one setting, like the U.S., to very different settings.

[9] The coaches in the pilot were teachers who were taken out of the classroom for a full year, and each coach had 10 teachers assigned to her. The cost of the program, without including the costs of filming, travel, and other expenses, was therefore at least 10 percent of teacher salaries. In practice, the salary costs were higher because the teachers who were chosen to be coaches were relatively senior and therefore earned above-average wages.

References

Al-Ubaydli, Omar, John A. List, and Dana J. Suskind. 2017. "What Can We Learn from Experiments? Understanding the Threats to the Scalability of Experimental Results." *American Economic Review* 107(5): 282–86.

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

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

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, and Michael Walton. 2017. "From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application". *Journal of Economic Perspectives* 31(4): 73-102.

Bau, Natalie, and Jishnu Das. 2020. "Teacher Value Added in a Low-Income Country." *American Economic Journal: Economic Policy* 12(1): 62-96.

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

Bold, Tessa, Deon Filmer, Ezequiel Molina, and Jakob Svensson. 2019. "The Lost Human Capital: Teacher Knowledge and Student Achievement in Africa." World Bank Policy Research Working Paper 8849.

Busso, Matias, Julian Cristia, Diana Hincapie, Julian Messina, and Laura Ripani. 2017. *Learning Better: Public Policy for Skills Development*. Washington, D.C.: Inter-American Development Bank.

Campos, Alejandra, Pedro Carneiro, Yyannu Cruz-Aguayo, and Norbert Schady. 2020. "Do Teacher Behaviors Predict Achievement, Executive Function, and Non-Cognitive Outcomes in Elementary School? Evidence from Multiple Rounds of Random Assignment." Unpublished manuscript, Inter-American Development Bank.

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): 2593-2632.

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

Cilliers, Jacobus, Brahm Fleisch, Cas Prinsloo, and Stephen Taylor. 2020. "How to Improve Teaching Practice? An Experimental Comparison of Centralized Training and In-Classroom Coaching." *Journal of Human Resources* 55(3): 926-62.

Conti, Gabriella, James Heckman, and Rodrigo Pinto. 2016. "The Effects of Two Influential Early Childhood Interventions on Health and Healthy Behavior." *The Economic Journal* 126(596): 28-65.

Downer, Jason, Marcia Kraft-Sayre, and Robert Pianta. 2009. "Ongoing, Web-Mediated Professional Development Focused on Teacher-Child Interactions: Early Childhood Educators' Usage Rates and Self-Reported Satisfaction." *Early Education and Development* 20: 321-45.

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

Education Next. 2018. "EdStat: $18 Billion a Year is Spent on Professional Development for U.S. Teachers." Available at https://www.educationnext.org/edstat-18-billion-year-spent-professional-development-u-s-teachers/, accessed on January 26, 2021.

Evans, David, and Anna Popova. 2016. "What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews." *World Bank Research Observer* 31(2): 242-70.

Fryer, Roland G. 2017. "The Production of Human Capital in Developed Countries: Evidence from 196 Randomized Field Experiments." In: Esther Duflo and Abhijit Banerjee, (Eds.), *Handbook of Field Experiments*, Volume 2 (pp. 95-322). Amsterdam: North Holland.

Gaminian, Alejandro, and Richard Murnane. 2016. "Improving Education in Developing Countries: Lessons from Rigorous Impact Evaluations." *Review of Educational Research* 86(3): 719-55.

Garcia, Jorge Luis, James Heckman, Ermini Leaf, and Maria Jose Prados. 2020. "Quantifying the Life-Cycle Benefits of an Influential Early-Childhood Program." *Journal of Political Economy* 128(7): 2502-41.

Glewwe, Paul, and Karthik Muralidharan. 2016. "Improving Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications." In: Eric A. Hanushek, Stephen Machin, and Ludger Woessman, (Eds.), *Handbook of the Economics of Education*, Volume 5 (pp. 653-743). Amsterdam: North Holland.

Hamre, Bridget, and Robert Pianta. 2007. "Learning Opportunities in Preschool and Early Elementary Classrooms." In Robert Pianta, Martha Cox, and Kyle Snow (Eds.), *School Readiness and the Transition to Kindergarten in the Era of Accountability* (pp. 49–84). (Baltimore, MD: Brookes).

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

Heckman, James, Rodrigo Pinto, and Peter Savelyev. 2013. "Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes." *American Economic Review* 103(6): 2052-86.

Huber, Martin. 2020. "Mediation Analysis." In Zimmermann, Klaus F. (Ed.) *Handbook of Labor, Human Resources and Population Economics*, pp. 1-38. Springer.

Imai, Kosuke, Dustin Tingley, and Teppei Yamamoto. 2013. "Experimental Designs for Identifying Causal Mechanisms." *Journal of the Royal Statistical Society*, Series A, 176: 5-51.

Jackson, Kirabo, Jonah Rockoff, and Douglas Staiger. 2014. "Teacher Effects and Teacher-Related Policies." *Annual Review of Economics* 6: 801-25.

Koedel, G., K. Mihaly, J. Rockoff. 2015. "Value-Added Modeling: A Review." *Economics of Education Review* 47: 180-95.

Kraft, Mathew, David Blazar, and Dylan Hogan. 2018. "The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence." *Review of Educational Research* 88(4): 547-88.

Kremer, Michael, Conner Brannen, and Rachel Glennerster. 2013. "The Challenge of Education and Learning in the Developing World." *Science* 340(6130): 297-300.

Leyva, Diana, Christina Weiland, M. Clara Barata, Hirokazu Yoshikawa, Catherine Snow, Ernesto Treviño, and Andrea Rolla. 2015. "Teacher–Child Interactions in Chile and Their Associations with Prekindergarten Outcomes." *Child Development* 86(3): 781-99.

List, John, Dana Suskind, and Lauren Supplee, Eds. Forthcoming. *The Scale-Up Effect in Early Childhood and Public Policy: Why Interventions Lose Impact at Scale and What We Can Do About It.* Routledge.

Loyalka, Prashant, Anna Popova, Guirong Li, and Zhaolei Shi. 2019. "Does Teacher Training Actually Work? Evidence from a Large-Scale Randomized Evaluation of a National Teacher Training Program." *American Economic Journal: Applied Economics* 11(3): 128-54.

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

Paxson, Christina, and Norbert Schady. 2011. "Does Money Matter? The Effects of Cash Transfers on Child Development in Rural Ecuador." *Economic Development and Cultural Change* 59(1): 187-230.

Perlman, Michal, Olesya Falenchuk, Brooke Fletcher, Evelyn McMullen, Joseph Beyene, and Prakesh S Shah. 2016. "A Systematic Review and Meta-Analysis of a Measure of Staff/Child Interaction Quality (the Classroom Assessment Scoring System) in Early Childhood Education and Care Settings and Child Outcomes." *PLoS ONE* 11(12): e0167660. doi:10.1371/journal.pone.0167660.

Pianta, Robert, Jamie DeCoster, Sonia Cabell, Margaret Burchinal, Bridget Hamre, Jason Downer, Jennifer LoCasale-Crouch, Amanda Williford, and Carollee Howes. 2014. "Dose-Response Relations between Preschool Teachers' Exposure to Components of Professional Development and Increases in Quality of Their Interactions with Children." *Early Childhood Research Quarterly* 29: 499-508.

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

Popova, Anna, David K. Evans, and Violeta Arancibia. 2016. "Training Teachers on the Job: What Works and How to Measure It." World Bank Policy Research Working Paper 7834.

Popova, Anna, David K. Evans, Mary E. Breeding, and Violeta Arancibia. 2018. "Teacher Professional Development around the World: The Gap between Evidence and Practice." World Bank Policy Research Working Paper 8572.

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

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

Weiland, Christina, and Hirokazu Yoshikawa. 2021. "Evidence-Based Curricula and Job-Embedded Coaching for Teachers Promote Preschoolers' Learning." *Child Evidence Brief, Society for Research in Child Development* 12.

Woodcock, Richard, and Ana Muñoz-Sandoval. 1996. *Baterria Woodcock-Muñoz: Pruebas de Aprovechamiento-Revisada*. Chicago: Riverside.

World Bank. 2018. *World Development Report 2018: Learning to Realize Education's Promise*. (Washington, D.C.: World Bank).

World Bank. 2021. *World Development Indicators*. Available at https://data.worldbank.org/indicator/SE.PRM.ENRL.TC.ZS?view=chart. Accessed on November 13, 2021.

Yoshikawa, Hirokazu, Diana Leyva, Catherine Snow, Ernesto Treviño, M. Clara Barata, Christine Weiland, Celia J. Gomez, Lorenzo Moreno, Andrea Rolla, Nikhit D'Sa, and Mary Catherine Arbour. 2015. "Experimental Impacts of a Teacher Professional Development Program in Chile on Preschool Classroom Quality and Child Outcomes." *Developmental Psychology* 51(3): 309–22.

Yoshikawa, Hirokazu, Christina Weiland, Jeanne Brooks-Gunn, Margaret R. Burchinal, Linda M. Espinosa, William T. Gormley, Jens Ludwig, Katherine A. Magnuson, Deborah Phillips, and Martha J. Zaslow. 2013. *Investing in Our Future: The Evidence Base on Preschool Education*. Society for Research in Child Development.

Table 1: Baseline characteristics of schools, teachers and children

	Full sample	Treated	Control	Test of differences (p-value)
	Panel A: School characteristics			
# of teachers between kindergarten and 3rd grade	4.67	4.47	4.87	0.563
Proportion of teachers from pay grade "E"	0.43	0.44	0.42	0.667
Proportion of teachers from pay grade "C" or "D"	0.20	0.19	0.21	0.489
Proportion of teachers from category "F"	0.25	0.24	0.26	0.594
Proportion of teachers from category "G" or "J"	0.08	0.1	0.06	0.203
	Panel B: Teacher characteristics			
Gender (female=1)	0.98	0.97	0.99	0.32
Years of experience	15.77	14.84	16.71	0.15
Contract (tenured=1)	0.90	0.91	0.90	0.78
# of children per classroom	31.48	31.83	31.13	0.47
	Panel C: Child characteristics			
Age (months)	84.42	84.09	84.74	0.00
Gender (female=1)	0.50	0.52	0.48	0.01

Note: Table reports means characteristics of teachers and children. P-values refer to a test that the value in the treatment and control groups is the same.

Table 2: Associations between CLASS and achievement

	(1)	(2)	(3)	(4)	(5)
	Panel A: Coaching experiment, control group				
Total CLASS score	0.097***				
	(0.026)				
Emotional Support		0.055**			-0.019
		(0.025)			(0.032)
Classroom Organization			0.103***		0.093***
			(0.026)		(0.034)
Instructional Support				0.079***	0.049*
				(0.024)	(0.027)
F-test					0.012
Observations	1,870	1,870	1,870	1,870	1,870
	Panel B: Alternative experiment, cross-sectional estimates				
Total CLASS score	0.082***				
	(0.008)				
Emotional Support		0.066***			0.006
		(0.008)			(0.011)
Classroom Organization			0.086***		0.080***
			(0.008)		(0.011)
Instructional Support				0.037***	0.008
				(0.008)	(0.008)
F-test					<0.001
Observations	16,254	16,254	16,254	16,254	16,254
	Panel C: Alternative experiment, within-school estimates				
Total CLASS score	0.062***				
	(0.011)				
Emotional Support		0.057***			0.021
		(0.011)			(0.016)
Classroom Organization			0.066***		0.056***
			(0.011)		(0.016)
Instructional Support				0.003	-0.018*
				(0.010)	(0.010)
F-test					<0.001
Observations	16,254	16,254	16,254	16,254	16,254

Notes: Regressions of achievement on the CLASS or its domains. Panel A refers to the control group in the coaching experiment. Panels B and C refer to 1st grade data from the multi-grade experiment analyzed in Araujo et al. (2016) and Campos et al. (2020). Regressions in Panel C include school fixed effects, those in Panel B do not. Standard errors in all regressions are corrected for clustering at the school level. *, **, and *** refer to significance at the 10 percent, 5 percent, and 1 percent, respectively.

Table 3: Effects of coaching intervention on teacher-child interactions and achievement

Panel A: Effects on teacher-child interactions

	Total CLASS	CLASS <10th pctile	CLASS <20th pctile	CLASS >80th pctile	CLASS >90th pctile	Emotional Support	Classroom Organization	Instructional Support
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	0.262*	-0.034	-0.075	0.114**	0.089**	0.323**	0.111	0.220
	(0.136)	(0.041)	(0.059)	(0.055)	(0.041)	(0.138)	(0.138)	(0.142)

Panel B: Effects on achievement

	Total achievement	Achievement <10th pctile	Achievement <20th pctile	Achievement >80th pctile	Achievement >90th pctile	ID numbers	Addition & subtraction	ID letters & words	Vocabulary
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Treatment	-0.069**	0.017*	0.025*	-0.021	-0.007	-0.058*	0.003	-0.077**	-0.081**
	(0.034)	(0.010)	(0.013)	(0.013)	(0.010)	(0.033)	(0.033)	(0.033)	(0.033)

Notes: Panel A: Regressions of teacher CLASS scores on treatment, corresponding to equation (1) in text. All regressions include fixed effects for blocks (10 fixed effects), and all the controls in panels A and B of Table 1. N is 198. Panel B: Regressions of test scores on treatment, corresponding to equation (2) in text. Sample sizes are 3,751. All regressions include fixed effects for blocks (10 fixed effects), and all controls in Table 1. Standard errors in Panel B are corrected for clustering at the school level. *, **, and *** refer to significance at the 10 percent, 5 percent, and 1 percent, respectively.

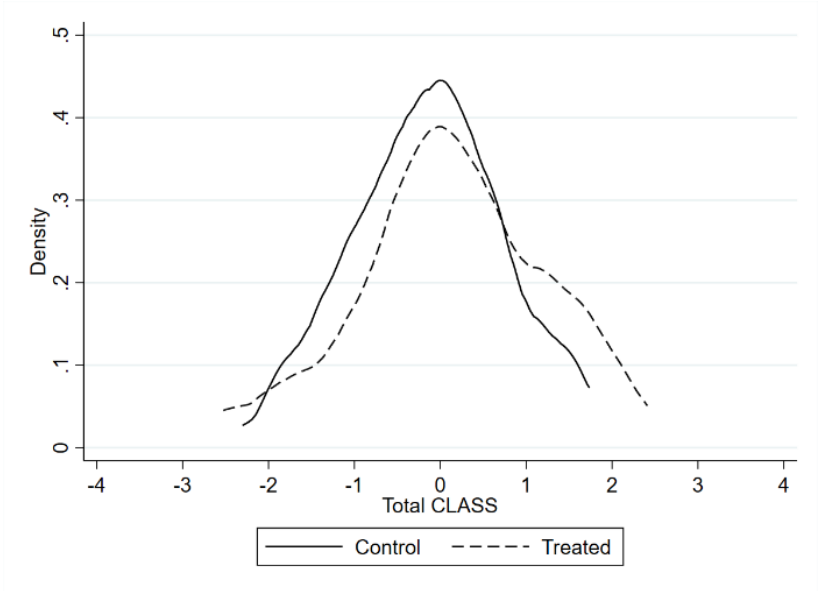
Table 4: Mediation analysis

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Total CLASS score	0.040**					0.103***				
	(0.018)					(0.027)				
Emotional Support		0.017			-0.044*		0.054**			-0.028
		(0.018)			(0.025)		(0.026)			(0.033)
Classroom Organization			0.060***		0.089***			0.112***		0.099***
			(0.018)		(0.025)			(0.027)		(0.035)
Instructional Support				0.020	0.001				0.092***	0.068**
				(0.017)	(0.019)				(0.026)	(0.028)
Treatment	-0.079**	-0.074**	-0.075**	-0.073**	-0.064*	-0.083**	-0.076**	-0.078**	-0.074**	-0.067**
	(0.034)	(0.034)	(0.034)	(0.034)	(0.034)	(0.034)	(0.034)	(0.033)	(0.034)	(0.034)
Treatment*						-0.105***				
Total CLASS score						(0.034)				
Treatment*							-0.070**			-0.028
Emotional Support							(0.035)			(0.051)
Treatment*								-0.088***		-0.007
Classroom Organization								(0.034)		(0.049)
Treatment*									-0.131***	-0.124***
Instructional Support									(0.034)	(0.040)

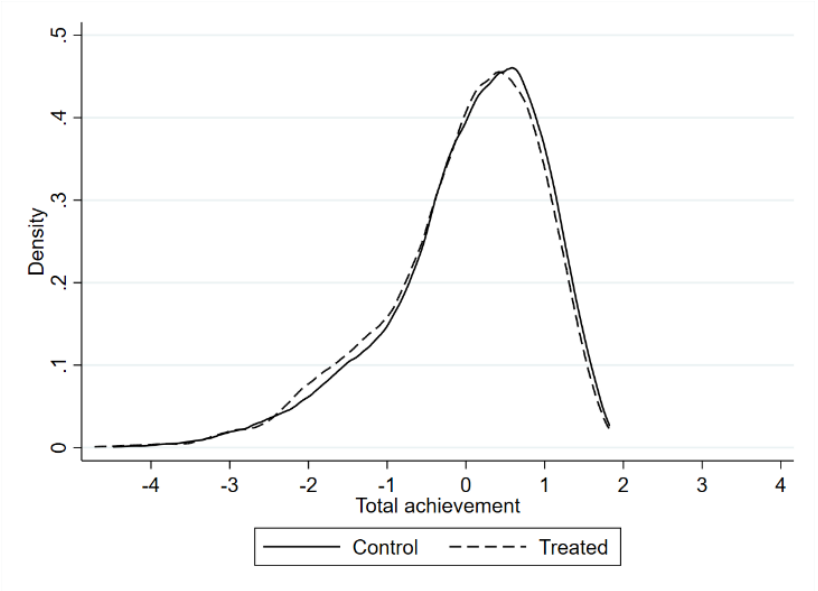
Notes: Regressions of test scores on treatment, the CLASS score (or one of its domains), and the interaction between treatment and the CLASS score (or one of its domains). Sample sizes are 3,751. Specifications (1) through (5) refer to equation (3a) in the text, and specifications (6) through (10) refer to equation (3b). All regressions include fixed effects for blocks (10 fixed effects), and all controls in Table 1. Standard errors are corrected for clustering at the school level. *, **, and *** refer to significance at the 10 percent, 5 percent, and 1 percent, respectively.

Figure 1: Univariate densities

Panel A: Total CLASS score



Panel B: Achievement



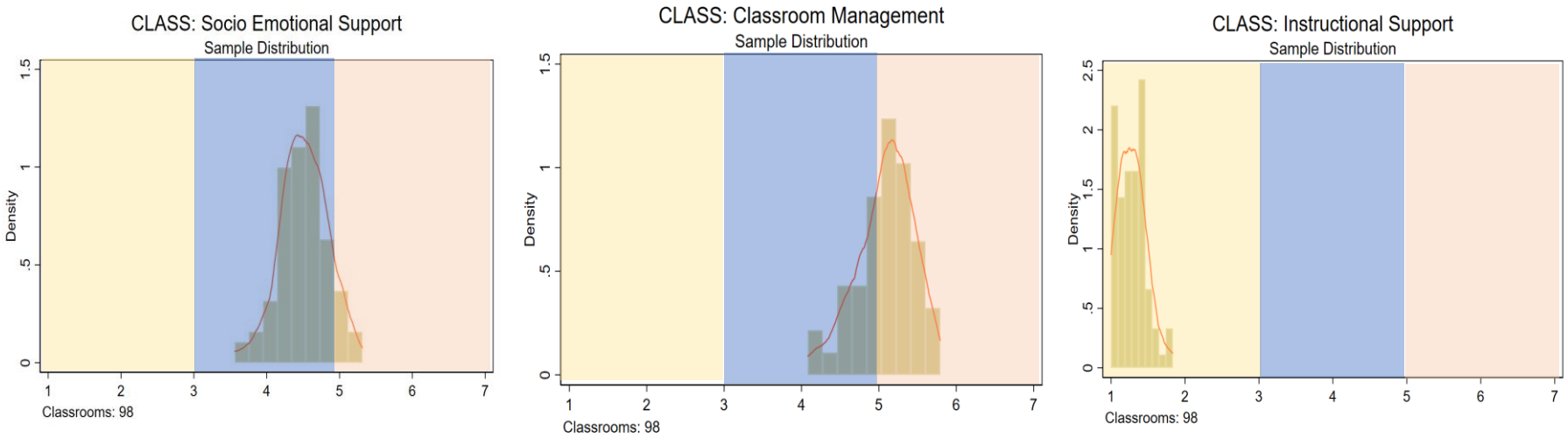
Note: Figure shows univariate densities of total CLASS scores and total achievement.

Appendix Table A1: Correlations among three domains of the CLASS score

	Emotional Support	Classroom Organization	Instructional Support
Socio-Emotional Support	1.000		
Classroom Management	0.6949*	1.000	
Instructional Support	0.4267*	0.4663*	1.000

Note:

Appendix Figure A1: Distribution of CLASS scores



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