Disruption in the classroom: Experimental evidence from Ecuador

Pedro Carneiro
University College London, IFS, CEMMAP

Yyannú Cruz-Aguayo
Inter-American Development Bank

Francesca Salvati
University of Essex, IFS

Norbert Schady
World Bank

May 1, 2024[1]

Abstract

We study how disruptive children affect learning using data from a unique experiment which randomly assigned children to classrooms for seven consecutive grades. Children with persistent behavioral problems lower the math and language achievement of their classmates. There are dosage effects and, when there are multiple children with behavioral difficulties in a classroom, the non-cognitive outcomes (including depression, self-esteem, grit, and growth mindset) of their classmates are also lower. We find indirect evidence that children with persistent behavioral difficulties are passed around schools.

[1] We gratefully acknowledge the support of the InterAmerican Development Bank. Carneiro also gratefully acknowledges the support of the ERC through grant ERC-2015-CoG-682349 and the Spencer Foundation.

1. Introduction

Across OECD countries, teachers spend on average 13 percent of their time (8 minutes per teaching hour) keeping order in the classroom (OECD 2019). If no learning occurs while teachers struggle to maintain order, the economic costs of classroom disruption could imply billions of dollars of foregone earnings in the U.S. alone.[2] Understanding the determinants of student misbehavior and how policies can reduce classroom disruption is an important priority.

In this paper, we study classroom disruption using a unique experiment in Ecuador, a middle-income country in South America. Our analysis is motivated by an influential paper by Lazear (2001), who presents a model of how classroom disruption affects learning. All children can be disruptive, but some are more disruptive than others. Particularly disruptive children have large (negative) spillovers on the learning of their peers. Lazear (2001) also shows that, under these circumstances, the optimal class size is inversely related to the likelihood of disruption.

In stark contrast with the large body of research on the effects of *average* peer quality, however, the empirical literature on how disruptive children affect the achievement of their classmates is small. In large measure, this is because the estimation challenges are considerable. First, disruptive children are unlikely to be placed in classrooms at random. Second, because it is not clear *what* makes a child disruptive, or how to measure it, researchers have generally focused on characteristics of children that are correlated with poor classroom behavior, including exposure to domestic violence (Carrell and Hoekstra 2010, Carrell et al 2018), or whether children have been diagnosed with, and treated for, attention-deficit disorders (Aizer 2008; Kristoffersen et al. 2015).[3]

To analyze the effects of classroom disruption, we use data from 202 schools in the coastal region of Ecuador. Every school had at least two classrooms per grade. A cohort of children entering kindergarten was randomly assigned to classrooms within schools. These children were then randomly reassigned to classrooms in every grade between 1st and 6th grades. Thus, children who did not switch schools were exposed to seven exogenous, orthogonal sets of peers, some of whom may have been

[2] This can be shown with simple back-of-the-envelope calculations. Hanushek and Woessmann (2020) propose a "rough rule of thumb" that, on average, there is 0.3 SDs of learning per grade. If there is no learning when there is misbehavior in the classroom, and teachers spend 13 percent of their time managing misbehavior (the OECD average), learning lost would be ~0.04 SDs (0.13*0.3), or ~1.6 percentiles at the mean of a standard normal distribution. Using data from Project Star, Chetty et al. (2011) estimate that a 1 percentile increase in kindergarten test scores leads to a 0.83 percent increase in earnings, so a 1.6 percentile decline would imply a ~1.3 percent decline in earnings. To translate this into dollars, we use data from the U.S Bureau of Labor Statistics. These data show that, in the first quarter of 2024, there were 119.2 million full-time wage and salary earners in the U.S., making $1,139 per week on average (Bureau of Labor Statistics 2024). A 1.3 percent decline in earnings would therefore amount to a yearly decline of $770 per worker per year, and a total loss of earnings of approximately $92 billion.

[3] See also Figlio (2007), who focuses on boys with names commonly given to girls. He shows that these children are more likely to be disruptive, and that this reduces peer achievement.

particularly disruptive.[4] Compliance with the random assignment was almost perfect, 98.9 percent on average.

At the end of each grade, children were tested in math and language. Between 1st and 4th grades, data on child executive function (EF) was also collected. EF refers to a set of skills that allow individuals to plan, focus attention, remember instructions, and juggle multiple tasks. It includes working memory, inhibitory control, and cognitive flexibility (Center for the Developing Child 2019). Data on child depression, self-esteem, grit, and growth mindset was collected at the end of 6th grade.

To identify disruptive children in the sample, we use information on their classroom behaviors. At the end of each school year, teachers were asked to list 5 children in their classrooms with the most serious behavioral problems and, separately, 5 children who had the biggest difficulties learning. In this paper, we are interested in the effects of children with *persistent* behavioral problems or difficulty learning. For this reason, and to alleviate concerns about measurement error in teacher reports from a single grade, we classify a child as *poorly-behaved* if teachers reported them to be one of the worst-behaved children in their classroom in (all of) the three previous grades, and define *low-achieving* students in a comparable way.[5]

The design of the experiment allows us to address important identification challenges. First, random assignment, with essentially perfect compliance, ensures that our results are not biased by purposeful placement of disruptive children. Second, since our measures are based directly on teacher reports (on the same children, but at different points in time) our analysis focuses precisely on the behaviors that teachers believe disrupt learning.

We study the impact of disruptive peers on learning only from 3rd grade onwards. We start in 3rd grade because, to classify a child as poorly-behaved or low-achieving, we need teacher reports from the three previous grades. For example, to classify a 3rd grade child as poorly-behaved, we use teacher assessments from kindergarten to 2nd grade. Third grade is the first grade for which we have three prior measures of student behaviors.

We first study the extent to which poorly-behaved students depress the learning outcomes of their classmates. Pooling information across grades 3 through 6, having one or more poorly-behaved

[4] Of course, there is also variation in exposure to other measures of classroom quality, including the quality of teachers. However, by design, these are orthogonal to peer quality. Teachers were also assigned randomly to classrooms within schools and grades.

[5] The decision to use three (as opposed to two, four, or any other number) years of problem behaviors (low achievement) to define poorly behaved (low achieving) children is done to balance two goals: identifying persistently disruptive children and estimating the model with reasonable sample sizes. The larger the number of consecutive periods of problem behaviors used in this classification the more restrictive the definition of a poorly behaved child, and the more seriously disruptive the child is likely to be. This leads to fewer but more seriously disruptive children and also fewer grades over which we can measure their impact, which affects the power of our estimates.

children in a classroom lowers classmates' achievement by (on average) -.019 SDs. There is considerable heterogeneity of impacts by grade, with the largest negative effects found among the youngest children: having a poorly-behaved student in 3rd grade reduces achievement by -.034 SDs, while the comparable effect in 6th grade is -.007 SDs. These results are consistent with those in other papers, including papers using this same experiment, which suggest that younger children may be more sensitive to environmental influences than those who are somewhat older, even within elementary school.[6]

In principle, persistently low-achieving children could also have negative effects on the learning of their classmates if teachers spend an inordinate amount of time helping these children catch up. We find no evidence that this is the case. Although poorly-behaved and low-achieving students are treated in the same way in the Lazear (2001) model—both disrupt the learning of their classmates—they are not equivalent in our sample: the former disrupt learning, while the latter do not.[7]

We analyze "dosage" effects and find these to be important. In the sample pooled across grades, having exactly one poorly-behaved student lowers others' achievement by -.011 SDs; having exactly two such students has an effect of -.030 SDs; and three or more such children reduce classmates' achievement by -.051 SDs. The effect of having at least three students with persistent behavioral problems is about one-half as large as that of having a one standard deviation better teacher, estimated for kindergarten teachers in this sample (Araujo et al. 2016).

The fact that we follow the same children over time allows us to study dynamic effects. We show that fade-out is substantial: for 3rd and 4th graders, the contemporaneous effect of having poorly-behaved peers is -.030 SDs, on average, while the effects one and two grades later are -.019 SDs and -.012 SDs, respectively.[8] An important question is what implications this has in the long run. We cannot analyze long-term effects from our experiment, but a number of studies of young children, primarily in the U.S., have found that the effects on achievement of being in a high-quality preschool, or the impact of better teachers, tend to fade out quickly, but reappear in adulthood in the form of better labor market performance or a lower probability of criminal behavior.[9]

---

[6] Carneiro et al. (2024) analyze the effect of within-classroom achievement rank on performance for children in this sample. More highly-ranked children have higher achievement than those with lower rank, with the largest effects found among children in 1st and 2nd grade. Aizer (2008) also finds evidence that the negative effects of peers with ADD are larger for younger children. She suggests that these may be driven by a higher rate of ADD diagnosis, and treatment (which improves these children's behavior), among somewhat older children. Although we have no data on the prevalence of ADD in our sample, the proportion of ADD children who are diagnosed is likely to be quite small, and the probability that they receive effective treatment even smaller.

[7] It is intuitive that children with behavioral difficulties could be disruptive. However, as argued by Lazear (2001), very low-achieving students could also depress the learning of their classmates—for example, by asking questions to which other students already know the answer.

[8] We cannot estimate twice-lagged effects for 5th and 6th grade children because our data ends in 6th grade.

[9] See, for example, Chetty et al (2014), and Jacob et al (2010) for estimates of the fade-out of the effects of teacher quality, measured by teacher value added.

Turning to outcomes other than achievement, we find no impact of poorly-behaved children on classmates' executive function. On the other hand, there is some evidence that having peers with persistent behavioral difficulties negatively affects depression, self-esteem, growth mindset, and grit in 6th grade, although the effects are only significant when there are multiple such students in a classroom. Specifically, having three or more poorly-behaved peers reduces the composite measure of non-cognitive outcomes in 6th grade by -.059 SDs.

Finally, we analyze whether being poorly-behaved, or having a poorly-behaved peer, increases the probability that a child attrits from our sample of schools. This is of interest because differential attrition could present an estimation challenge, but also substantively, if poorly behaved kids are encouraged to move schools. In practice, attrition is no different for children in classrooms with, and without, poorly-behaved children. On the other hand, children with behavioral problems are themselves more likely to leave the sample, and children who transfer from other schools are more likely to have behavioral problems—even years after they first arrive. Below, we discuss how this could affect our estimates.

The design of the experiment allows us to make several contributions. First, we can examine the effect of poorly-behaved students on achievement, but also on other outcomes (executive function, including working memory, inhibitory control, and cognitive flexibility, as well as, separately, non-cognitive outcomes, including self-esteem, growth mindset, grit, and depression). This is important because there is good evidence that EF and various non-cognitive skills are malleable in childhood and can have large effects on adult outcomes.[10]

Second, ours is the first paper that explicitly compares two kinds of children who could potentially have negative effects on the achievement of their classmates: poorly-behaved children, and children with persistently low achievement. Third, we can study fade-out (albeit over a limited time horizon), dosage effects, and differences in the impacts of disruptive peers across grades. In practice, all these considerations are important in our setting.

Fourth, we examine the relationship between student turnover and disruptive behavior, by asking whether disruptive students are more likely to move schools (and also whether children randomly assigned to classrooms with disruptive students are more likely to move schools). Fifth, to the best of our knowledge, ours is the first paper on the effect of disruptive peers in a developing country setting. This is noteworthy because in general class sizes will be larger, teachers will have fewer qualifications,

[10] This is a large literature. On executive function, see Moffitt et al. (2011), and on non-cognitive outcomes, see Heckman et al. (2006), among many references.

and the proportion of children with undiagnosed medical conditions, such as ADD, will be higher in poorer countries.

The remainder of the paper proceeds as follows. We discuss the setting, data, and experimental design in section 2. Section 3 presents our empirical specification and results. Section 4 concludes.

2. Data and experimental design

The data we use come from an experiment in 202 schools in Ecuador, a middle-income country in South America.[11] Schools have at least two classrooms per grade (most have exactly two). A cohort of children entering kindergarten was randomly assigned to classrooms (within schools) in the 2012 school year, and then randomly reassigned them to classrooms in every grade between 1st and 6th grade. Compliance with the assignment rules was very high—98.9 percent on average. We provide further details on the classroom assignment rules and compliance with randomization in Appendix A.

We have baseline data on maternal education, household wealth, whether a child attended preschool, and her vocabulary skills at the beginning of kindergarten. Data on math and language achievement was collected at the end of each grade between kindergarten and 6th grade. For both subjects, tests were a mixture of material that teachers were meant to have covered explicitly in class—for example, in math, addition or subtraction; material that would have been covered, but probably in a somewhat different format—for example, simple word problems; and material that would not have been covered at all in class but that has been shown to predict current and future math achievement—for example, the Siegler number line task (Siegler and Booth 2004). We aggregate responses in math and, separately, language, by Item Response Theory (IRT), and calculate an average achievement score that gives the same weight to math and language.[12]

Child executive function (EF) was assessed in every grade between kindergarten and 4th grade. EF includes a set of basic self-regulatory skills which involve various parts of the brain, but in particular the prefrontal cortex. Low levels of EF are associated with low levels of self-control and "externalizing" behavior, including disruption, aggression, and inability to sit still and pay attention (Séguin and Zelazo 2005). Executive function in childhood has also been shown to predict a variety of outcomes in adulthood, including performance in the labor market, involvement in criminal activities, and health status, even after controlling for socioeconomic status in childhood (Moffitt et al. 2011).

[11] Araujo et al. (2016) discuss in detail the selection of schools in this study. They show that the characteristics of students and teachers in our sample are very similar to those of students and teachers in a nationally representative sample of schools in Ecuador.

[12] Our results are very similar if, instead, we calculate a simple sum of correct responses within blocks of questions on each test and give equal weight to each of these test sections (as in Araujo et al. 2016).

Executive function is generally thought of as having three domains: working memory, inhibitory control, and cognitive flexibility. We separately calculate scores for each of these domains, as well as an average EF score that gives the same weight to each component. In 6th grade, finally, data was collected on child depression, self-esteem, growth mindset, and grit. For each outcome, we aggregate responses by factor analysis, and also calculate an overall non-cognitive score that gives the same weight to each of the individual assessments. Further details on child assessments are provided in Appendix B.

At the end of each grade, teachers were asked to list the 5 children with the most severe behavioral problems and, separately, the 5 children with the lowest achievement in their class. We use these data to classify a child as *poorly-behaved* if teachers in the three previous grades reported them to be one of the worst-behaved children and define *low-achieving* students in a comparable way.

Importantly, our experiment generates considerable variation in exposure to poorly-behaved (and low-achieving) students. This can be seen in Table 1, which shows the number of poorly-behaved students in each grade (column 1); the number of classrooms with poorly-behaved students and the number of total classrooms (columns 2 and 3); the proportion of classrooms with different numbers of poorly-behaved students (columns 4 to 7); the proportion of classrooms (among those with at least one poorly-behaved student) where the poorly-behaved student is rated by the teacher as being one of the five worst-behaved students in the current grade (column 8), as well as those who remain in the bottom five in the subsequent grades (columns 9 to 11).

Although the proportion of poorly-behaved students in our sample is small—between 2.5 and 3.3 percent by grade—most of them are in different classrooms. As a result, roughly half of all classrooms in 3rd through 6th grade have at least one disruptive student. Table 1 also shows there is a high degree of persistence in disruptive behavior: over two-thirds of children who were reported to be among the 5 worst-behaved by their teachers in three consecutive grades are also reported to be among the 5 worst-behaved children in the following grade.[13]

Table 2 provides summary statistics for children in our sample, comparing those who are classified as poorly-behaved in at least one grade, using teacher reports from the three previous grades, and those who are not. It shows that children who are not poorly-behaved were 5 years of age on the first day of kindergarten, on average, and half of them are girls. Mothers were in their early 30s and fathers in their mid-30s. Both parents had on average just under 9 years of schooling, which corresponds

[13] Recall that a child is classified as *poorly-behaved* in *g* if he was rated as being among the 5 worst-behaved students in the classroom at the end of grades *g-1*, *g-2* and *g-3*. One way to validate the informativeness of our measure is to check if children who we classify as disruptive in *g* (based on past information) also exhibit poor behaviors in *g+1*, and Table 1 shows that this is indeed the case. Much the same holds for children who are persistently low-achieving: between 65 percent and 71 percent of children who are listed as having the biggest difficulties learning in *g-2*, *g-1* and *g* are also listed as such by their teachers in *g+1*.

to completed middle school. Araujo et al (2016) report that, at the beginning of kindergarten, the average receptive vocabulary score of children in the sample is 1.7 SDs below the level of children that were used to norm the test.[14]

Turning to the comparison between poorly-behaved and other students, Table 2 shows that children with persistent behavioral problems are overwhelmingly male—over 95 percent of them are boys. They have lower performance on math and language tests than other children, and lower levels of executive function. Poorly-behaved students also have worse depression scores, lower levels of self-esteem, lower levels of grit, and lower values for the measure of growth mindset than other children.[15] Broadly speaking, the socioeconomic status of poorly-behaved and other students appears to be similar: The education levels of fathers of poorly-behaved students are higher, but household wealth is lower, and these differences are small in magnitude. Poorly-behaved students are more likely to have attended preschool than other children, a difference of about 10 percentage points. This may seem surprising, although we note there are several papers which show that prolonged time in daycare can have negative impacts on children's socio-emotional development (see for example Baker et al. 2008, 2019 on a program in Quebec).

3. Empirical specification and results

A. Empirical specification

Our main goal is to estimate whether child *i* in classroom *c*, grade *g*, and school *s* has lower achievement, executive function, or non-cognitive development after she was randomly assigned to classrooms with, or without, poorly-behaved students. For this purpose, we run regressions of the following form:

$$Y_{i,c,g,s} = \beta D_{c,g,s} + \varphi\left(Y_{i,c,g-1,s}\right) + \theta X_{i,c,g,s} + \delta_{g,s} + \varepsilon_{i,c,g,s} \quad (3.1)$$

where $D_{c,g,s}$ is an indicator variable that takes on the value of one if there is one or more disruptive students in a classroom; the function $\varphi(.)$ is a flexible formulation of lagged achievement or executive function;[16] $X_{i,c,g,s}$ includes child age and gender; $\delta_{g,s}$ is a set of school-by-grade fixed effects; and $\varepsilon_{i,c,g,s}$ is a residual. Standard errors are clustered at the school level.

[14] To measure baseline receptive vocabulary, we use the *Test de Vocabulario en Imágenes Peabody* (TVIP) (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, for example Schady et al. (2015).

[15] For the comparisons in Table 2, we use *lagged* achievement and executive function. We cannot do this for the measures of depression, self-esteem, growth mindset and grit, as these were only collected in 6th grade.

[16] In practice, we use a fourth-order polynomial in lagged achievement or executive function, although our estimates are robust to using only lower-order polynomials. Note that because measures of depression, self-esteem, growth mindset, and grit were only collected in 6th grade, the regressions for these non-cognitive outcomes do not include the lags.

Other estimates we report are variants on this basic formulation. Specifically, we estimate (1) regressions that refer to outcomes measured one or two grades after a student was exposed to a disruptive peer, not just those that refer to contemporaneous effects; (2) models in which the coefficients on β are allowed to vary by grade, rather than restricted to be the same across all grades; (3) models that allow effects to vary with the number of disruptive students in a classroom; and (4) models that include separate indicator variables for poorly-behaved and low-achieving children.

B. Results

Our first set of results is in Table 3, where the outcome of interest is the average of math and language achievement in each grade. The first row of Panel A shows estimates of equation (3.1), while other rows of this panel correspond to variants of this equation where achievement is measured at a later point in time than that when children were, or were not, exposed to a poorly-behaved classmate ($D_{c,g,s}$). In the first column of the panel the coefficients of equation (3.1) are restricted to be the same for all grades, while in the remaining columns this restriction is relaxed.

We see that, in the model that restricts coefficients to be the same across grades, having at least one poorly-behaved student in a class lowers the achievement of classmates by -.019 SDs. The effects fall monotonically by grade, and we can reject the null that the average effect for 3rd and 4th graders, and that for 5th and 6th graders are the same (p-value: .055, reported in the last column of this row).[17]

There is also some evidence that the impact of being exposed to poorly-behaved students fades out. The most convincing comparisons are those that look at specific grades—in the pooled sample the number of grades that are included varies across regressions, so we could confound differences in effects by grade with the pattern of fade-out. In 3rd grade, the contemporaneous, once-lagged, twice-lagged, and thrice-lagged effects are -.034, -.017, -.011, and .012, respectively, and we can reject the null that these effects are the same (p-value: .002, reported in the last row of this panel). In 4th and 5th grades, the patterns are less clear, although the number of lags we can consider is also smaller.

Panel B turns to the comparison between the effects of poorly-behaved and low-achieving students. As discussed earlier, these children are treated as equivalent in the Lazear (2001) model. The results in this panel are based on regressions that include indicator variables for whether a classroom had at least one poorly-behaved student and, separately, at least one low-achieving student. The coefficients

[17] The fact that poorly-behaved students have larger, negative effects on their classmates in the earlier grades could occur either because when they are younger, poorly-behaved students engage in behaviors that are more disruptive than when they are older (for example, biting a classmate), or because older children are better able to pay attention than younger children even when there is classroom disruption, or some combination of both.

on low-achieving students are always close to zero. In the pooled regression, we can reject the null that the effects of poorly-behaved and low-achieving students are the same (p-value: .007).

Panel C, finally, focuses on the number of poorly-behaved students in a classroom—what we refer to as dosage effects. For this purpose, we estimate a version of equation (3.1) that expands our explanatory variable, $D_{c,g,s}$, from a single indicator for whether there was at least one disruptive child in the classroom, to three indicators for whether there were 1, 2, or 3 or more disruptive children in the classroom. There is clear evidence that having more poorly-behaved students in the classroom is worse than having less of them: in the estimates that pool across grades, having exactly 1, exactly 2, and 3 or more students with persistent behavioral difficulties lowers learning of other children in the class by -.011, -.030 and -.051 SDs, respectively, and we can reject the null that these effects are the same (p-value: .001, in the last row of the panel).

In sum, Table 3 shows that students with persistent behavioral difficulties lower the achievement of their classmates; that the effects are concentrated among the youngest students; that the negative impacts of poorly-behaved students on achievement fade out over time; that there are dosage effects—the more poorly-behaved children there are in a class, the larger is the negative effect on the achievement of their classmates; and that having students who are persistently low-achieving, as reported by their teachers, does not lower the learning outcomes of their classmates.

We turn to other outcomes in Table 4. Each outcome is in a different column. The first 4 columns correspond to executive function, and we present results that restrict coefficients to be the same across grades 3 and 4 (we do not have measures of EF for grades 5 and 6). The first column aggregates different measures of EF, and individual impacts on inhibitory control, memory and attention, and cognitive flexibility are shown in columns 2 to 4. Similarly, column 5 aggregates non-cognitive skills into a single index, and results for individual components of this index are shown in the subsequent columns. Panel A corresponds to the estimates of equation (3.1), while Panel B considers an extension of equation (3.1) that accounts for dosage effects.

Panel A shows there is no evidence that having poorly-behaved peers in the classroom lowers the scores on the measures of classmates' inhibitory control, working memory, or cognitive flexibility, or on the composite measure of executive function. Panel B shows that this is the case for children exposed to a single, but also multiple, poorly-behaved students.

Other columns in the table focus on the effects of poorly-behaved peers on depression, self-esteem, growth mindset, and grit. In Panel A, the coefficients from these regressions are consistently negative, but they are not significant. Moreover, because we only collected data on these outcomes in 6th grade, we cannot pool data across grades (as we do with achievement) to increase precision. That said,

here too we find evidence of dosage effects, as can be seen in Panel B. In the regression that focuses on the non-cognitive aggregate, the coefficients on 1, 2, and 3 or more poorly-behaved students are -.009, -.013, and -.059, respectively, and the coefficient on 3 or more students is significant at conventional levels.[18] The clearest negative effects of multiple students with persistent behavioral problems are on growth mindset.

Next, in Table 5, we turn to patterns of transfers in and out of our sample of schools. In Panel A we show the impact of having a poorly-behaved peer (column 1), or of being a poorly-behaved student (column 2), on the likelihood of leaving the sample between two consecutive grades. Children in our sample are no more likely to attrit when they are exposed to poorly-behaved students. Therefore, it is unlikely that our estimates are affected by selective attrition. On the other hand, the grade-on-grade attrition rate of poorly-behaved students is higher by 2.4 percentage points (relative to average grade-to-grade attrition of 5.3 percent). The converse is also true: children who move out of the school in a given grade are more likely to be classified as poorly behaved in the previous grade.

Note also that just as students from the schools in our sample are being *sent* to other schools, the schools in our sample are *receiving* students transferring from elsewhere. In panel B we estimate how the probability of being poorly behaved differs between new entrants and children who were already in a given school. In column 1 of panel B, we regress an indicator for being listed by a teacher in a given grade as a child with behavioral problems on an indicator for being a new entrant into the sample. In-transfers are 0.7 percentage points more likely to be reported as having behavioral problems than other children.[19] Furthermore, new entrants in a given grade *g* are not only more likely to be poorly behaved in that grade, but they are also more likely to be persistently poorly behaved, and of being classified as poorly-behaved by our procedure. Column 2 of panel B of the table shows that in-transfers in grade *g* are 0.7 percentage points more likely to be reported as being persistently disruptive than other children, three years later.

We do not know why poorly-behaved children are more likely to move around schools. The decision could be voluntary, or a response to pressure from other parents, principals, and teachers. It is also not clear whether the effect of reshuffling disruptive children across schools is on aggregate positive (because children find a better school match) or negative (because disruptive children have a hard time

[18] Interestingly, these effect sizes are very similar to those we estimate for achievement—as shown in Table 3, the effects of having exactly 1, exactly 2, and 3 or more poorly-behaved students on 6th grade learning are .004, -.018, and -.059 SDs, respectively.

[19] Hanushek et al. (2004) also find that children who move schools perform worse than other schools, and that movers reduce achievement in receiving classrooms.

adjusting to a new environment, potentially causing even more disruption in their new classrooms).[20] Regardless, the reshuffling of poorly-behaved children means that we are likely to underestimate the effects of disruptive children on learning.[21]

4. Conclusion

In this paper, we show that children with persistent behavioral problems lower the achievement of their peers. When there are multiple children with behavioral difficulties in a classroom, they also lower their classmates' scores on a composite measure of non-cognitive outcomes that includes depression, self-esteem, growth mindset, and grit. The pattern of in- and out-transfers suggest that poorly-behaved students may be "passed around" schools. Finally, our results show that, while in principle both *poorly-behaved* and *low-achieving* children could disrupt learning, in practice, only those with persistent behavioral problems lower the achievement of their classmates in the setting we study.

The fact that children who have persistent behavioral problems have negative effects on their classmates raises important questions. How can policy-makers best ensure that any underlying medical conditions, like ADD, are diagnosed and treated? Should children with persistent behavioral problems be mainstreamed or placed in special needs classrooms? If children with persistently poor behavior are kept in regular classes, what are tools that teachers can use to effectively manage misbehavior?[22] Providing answers to these questions is difficult. Designing effective policies for children who have persistent behavioral problems is likely to be particularly challenging in developing countries, where resources are more limited.

[20] We also note that reshuffling of this sort is likely to occur in other settings—for example, moving poorly-performing employees around departments in a large company if these workers cannot be fired. Both cases involve situations of asymmetric information—the sending unit (school, department) is likely to know more about the poorly-performing individual (student, employee) than the receiving unit.

[21] This is because a child is classified as *poorly-behaved* in *g* if she was rated as being among the 5 worst-behaved students in the classroom at the end of grades *g-1*, *g-2* and *g-3*. Therefore, no matter how disruptive his behavior, a new arrival in 4th, 5th, and 6th grades cannot be classified as poorly-behaved.

[22] Developing countries spend considerable resources on in-service training for teachers, but most programs appear to be ineffective (see Popova et al. 2022). A coaching program for 1st grade teachers, implemented in a different sample of urban schools in Ecuador, did not raise achievement, and may have worsened outcomes as teachers struggled to change their in-class behaviors (see Carneiro et al. 2022).

References

Aizer, A. 2008. "Peer Effects and Human Capital Accumulation: The Externalities of ADD." National Bureau of Economic Research Working Paper 14354.

Araujo, M. C., P. Carneiro, Y. Cruz-Aguayo, and N. Schady. 2016. "Teacher Quality and Learning Outcomes in Kindergarten." *Quarterly Journal of Economics* 131(3): 1415-53.

Baker, M., J. Gruber, and K. Milligan. 2008. "Universal Child Care, Maternal Labor Supply, and Family Well-Being." *Journal of Political Economy* 116(4): 709-45.

----------. 2019. "The Long-Run Impacts of a Universal Child Care Program." *American Economic Journal: Economic Policy* 11(3): 1-26.

Bureau of Labor Statistics. 2024. "News Release: Usual Weekly Earnings of Wage and Salary Workers, First Quarter 2024". Available at https://www.bls.gov/news.release/pdf/wkyeng.pdf. Accessed on April 24, 2024.

Carneiro, P., Y. Cruz-Aguayo, R. Intriago, J. Ponce, N. Schady, and S. Schodt. 2022. "When Promising Interventions Fail: Personalized Coaching for Teachers in a Middle-Income Country." *Journal of Public Economics Plus* 3.

Carneiro, P., Y. Cruz-Aguayo, F. Salvati, and N. Schady. 2024. "The Effect of Classroom Rank on Learning Throughout Elementary School: Experimental Evidence from Ecuador." Forthcoming, *Journal of Labor Economics*.

Carrell, S. E., and M. L. Hoekstra. 2010. "Externalities in the Classroom: How Children Exposed to Domestic Violence Affect Everyone's Kids." *American Economic Journal: Applied Economics* 2(1): 211-28.

Carrell, S. E., M. L. Hoekstra, and Elira Kuka. 2018. "The Long-Run effects of Disruptive Peers." *American Economic Review* 108(11): 3377-3415.

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

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

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

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

Figlio, D. N. 2007. "Boys Named Sue: Disruptive Children and Their Peers." *Education Finance and Policy* 2(4): 376-94.

Hanushek, E. A., J. F. Kain, and S. G. Rivkin. 2004. "Disruption versus Tiebout Improvement: The Costs and Benefits of Switching Schools." *Journal of Public Economics* 88: 1721-46.

Hanushek, E., and L. Woessmann. 2020. The Economic Impacts of Learning Losses. Available at https://www.ams-forschungsnetzwerk.at/downloadpub/2020_OECD_The-economic-impacts-of-coronavirus-covid-19-learning-losses.pdf. Accessed on April 24, 2024.

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

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

Kristoffersen, J. H. G., M. V. Krægpøt, H. S. Nielsen, and M. Simonsen. 2015. "Disruptive School Peers and Student Outcomes." *Economics of Education Review* 45(4): 1-13.

Lazear, E. P. 2001 "Educational Production." *Quarterly Journal of Economics* 141(3): 777-803.

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

OECD 2019. "TALOS 2018 Results: Teachers and School Leaders as Lifelong Learners", Volume I, OECD.

Popova, A., D.K. Evans, M.E. Breeding, and V. Arancibia. 2022. "Teacher Professional Development around the World: The Gap between Evidence and Practice." *World Bank Research Observer* 37(1): 107-36.

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

Seguin, J. and P. Zelazo, "Executive Function in Early Physical Aggression," in Developmental Origins of Aggression, R. Tremblay, V. Hartup, and J. Archer, eds. (New York: Guilford Press, 2005).

Table 1: Distribution of poorly-behaved students across classrooms and persistence

	Nr poorly-behaved students	Nr classrooms with poorly-behaved student	Total nr classrooms	Proportion classrooms with poorly-behaved student	Proportion classrooms with 1 poorly-behaved student	Proportion classrooms with 2 poorly-behaved students	Proportion classrooms with 3+ poorly-behaved students	Proportion poorly-behaved students among bottom 5 in g	Proportion poorly-behaved in g+1	Proportion poorly-behaved in g+2	Proportion poorly-behaved in g+3
3rd grade	299	206	470	.44	.33	.08	.03	.64	.65	.49	.40
4th grade	338	224	479	.47	.32	.11	.04	.73	.77	.60	
5th grade	439	276	485	.57	.36	.14	.07	.68	.71		
6th grade	490	300	485	.62	.37	.17	.08	.66			

Notes: The table shows descriptive statistics about the distribution of poorly-behaved students across classrooms in every grade, as well as persistence in poor behavior. In any grade g between grades 3 and 6, a poorly-behaved student is a student who was ranked among the bottom 5 worst behaved students in the classroom according to the teacher in all grades between g-1, g-2 and g-3. In columns 1-2 we show how many students comply with this definition in each grade between 3rd and 6th grade, and the number of classrooms in which there is a poorly-behaved student according to this definition. In columns 3-7 we show the total number of classrooms in the sample, as well as the proportion of classrooms with a poorly-behaved student, and the proportion of classrooms with one, two, or three or more poorly-behaved students. In column 8, we show the proportion of poorly-behaved students in any given grade who are also listed among the bottom 5 worst behaved students in grade g, conditional on attrition. In columns 9-11 we show the proportion of poorly-behaved students in any given grade g who are also poorly-behaved in grades g+1, g+2 and g+3 using our definition, conditional on attrition.

Table 2: Characteristics of poorly-behaved students

	Poorly-behaved		Not poorly-behaved			
	Mean	N	Mean	N	Diff.	p-value
Female	.049	879	.523	9,408	-.474	.000
Age (months)	123	851	121	9,408	1.46	.000
Lagged math	-.321	878	.084	9,408	-.405	.000
Lagged language	-.464	878	.111	9,408	-.575	.000
Lagged math+language index	-.392	878	.098	9,408	-.490	.000
Lagged EF	-.277	878	.069	9,407	-.346	.000
Lagged EF inhibitory control	-.141	878	.028	9,407	-.169	.000
Lagged EF memory and attention	-.263	878	.068	9,407	-.332	.000
Lagged EF cognitive flexibility	-.116	878	.030	9,407	-.146	.000
Aggregate non-cognitive	-.244	356	.012	7,433	-.256	.000
Depression	-.193	356	.009	7,433	-.202	.000
Self-esteem	-.160	356	.008	7,433	-.167	.002
Growth mindset	-.224	356	.011	7,433	-.235	.000
Grit	-.181	356	.009	7,433	-.190	.000
Mother education less than secondary	.689	544	.680	8,876	.009	.659
Father education less than secondary	.756	389	.701	7,157	.055	.022
Mother age	29.6	541	30.4	8,829	-.811	.004
Father age	34.5	373	34.7	7,004	-.225	.588
Wealth	-.070	582	.014	9,277	-.084	.040
TVIP	-.019	565	.043	9,115	-.062	.141
Preschool	.707	590	.602	9,401	.105	.000

Notes: This table shows characteristics of poorly-behaved students according to our main definition, and compares them to non-poorly-behaved students. In any grade g between grades 3 and 6, a poorly-behaved student is a student who was ranked among the bottom 5 worst behaved students in the classroom according to the teacher in all grades between g-1, g-2 and g-3. The table reports the mean of each variable for poorly-behaved and non-poorly-behaved students, as well as the difference in means between poorly-behaved students and non-poorly-behaved students, and the p-values testing whether the differences in means are equal to zero, pooling across grades 3 to 6. Data on executive function are only available up to grade 4.

Table 3: Effect of poorly-behaved students on classmates' achievement

Panel A	Pooled	3rd grade	4th grade	5thgrade	6thgrade	F-test 1	F-test 2
Has poorly-behaved student - Lag 0	-.019	-.034	-.027	-.011	-.007	.475	.126
	(.006)	(.016)	(.014)	(.013)	(.011)		
Has poorly-behaved student - Lag 1	-.015	-.017	-.021	-.010			
	(.007)	(.015)	(.011)	(.011)			
Has poorly-behaved student - Lag 2	-.016	-.011	-.022				
	(.010)	(.016)	(.012)				
Has poorly-behaved student - Lag 3	.012	.012					
	(.015)	(.016)					
F-test 3	.128	.032	.859	.906			
Panel B	Pooled	3rd grade	4th grade	5th grade	6th grade	F-test 1	
Has poorly-behaved student	-.019	-.034	-.028	-.011	-.006	.441	.116
	(.006)	(.016)	(.014)	(.013)	(.011)		
Has low achieving student	.003	-.006	.013	.001	.001	.889	.880
	(.008)	(.018)	(.017)	(.012)	(.012)		
F-test 4	.037	.293	.065	.472	.665		
Panel C	Pooled	3rd grade	4th grade	5thgrade	6thgrade	F-test 1	F-test 2
1 student	-.011	-.027	-.027	.002	.004	.232	.040
	(.006)	(.017)	(.016)	(.013)	(.011)		
2 students	-.030	-.049	-.022	-.035	-.018	.807	.710
	(.010)	(.031)	(.022)	(.018)	(.018)		
3+students	-.051	-.071	-.076	-.027	-.059	.809	.441
	(.016)	(.052)	(.043)	(.031)	(.027)		
F-test 5	.010	.575	.477	.086	.041		

Notes: Panel A reports estimates from regressions of an index of math and language scores on an indicator for being randomly assigned to a classroom with a poorly-behaved student for various lags of year of assignment to a classroom with a poorly-behaved student (where lag 0 captures the contemporaneous effect), and for various grades. Column 1 pools information across grades 3-6. Columns 2-5 report estimates from regressions by grade. In each regression, we regress the math and language scores index on an indicator variable for being assigned to a classroom with a poorly-behaved student, controlling for a fourth-order polynomial in lagged achievement, an indicator for a poorly-behaved student, child age and gender, and school (by grade, when pooling data across grades) fixed effects. F-test 1 reports the p-value for a test of equality of contemporaneous effects across grades 3-6. F-test 2 reports the p-value for a test of equality of average effects in grades 3 and 4 vs grades 5 and 6. F-test 3 reports p-values of tests for equality of contemporaneous and lagged effects. Panel B reports estimates from regressions of an index of math and language scores on an indicator for being randomly assigned to a classroom with a poorly-behaved student and an indicator for being randomly assigned to a classroom with a low achieving student, for various grades. In any grade t between grades 3 and 6, a low achieving student is a student who was ranked among the bottom 5 worst achieving students in the classroom according to the teacher in all grades between t-1, t-2 and t-3. Column 1 pools information across grades 3-6. Columns 2-5 report estimates from regressions by grade. Each regression controls for a 4th-order polynomial in lagged achievement, an indicator for being a poorly-behaved or low achieving student as well as child gender and age, and school (by grade when pooling information across grades) fixed effects. F-test 1 reports the p-value for a test of equality of effects across grades 3-6. F-test 4 reports p-values of tests for equality of impacts of poorly-behaved and low achieving students. Panel C reports estimates from regressions of an index of math and language scores on indicators for being randomly assigned to a classroom with a varying number of poorly-behaved students, for various grades. Students can be assigned to classrooms with one, two, or three or more poorly-behaved students. Column 1 shows results for a specification in which we pool information across grades 3-6. Columns 2-5 report estimates from regressions by grade. In each regression, we regress the math and language scores index on indicators for the number of poorly-behaved students in the classroom (omitted category is 0), controlling for a fourth-order polynomial in lagged achievement, an indicator for being a poorly-behaved student, child age and gender, and school (by grade when pooling information across grades) fixed effects. F-test 1 reports the p-value for a test of equality of effects across grades 3-6. F-test 2 reports the p-value for a test of equality of average effects in grades 3 and 4 vs grades 5 and 6. F-test 5 reports p-values of tests for equality of coefficients across rows of a given column. Standard errors are clustered at the school level throughout.

Table 4: Effect of poorly-behaved students on student executive function and non-cognitive outcomes

Panel A	EF composite	Inhibitory control	Memory and attention	Cognitive flexibility	Aggregate non-cognitive	Depression	Self-esteem	Growth mindset	Grit
Has poorly-behaved student	.005	-.019	.011	-.013	-.015	-.015	-.008	-.014	-.011
	(.016)	(.017)	(.017)	(.018)	(.014)	(.016)	(.015)	(.014)	(.015)
Panel B	EF composite	Inhibitory control	Memory and attention	Cognitive flexibility	Aggregate non-cognitive	Depression	Self-esteem	Growth mindset	Grit
1 student	.011	-.021	.023	-.013	-.009	-.007	-.002	-.011	-.011
	(.018)	(.018)	(.019)	(.019)	(.015)	(.018)	(.016)	(.016)	(.017)
2 students	.012	-.026	-.010	-.011	-.013	-.026	-.011	-.005	-.008
	(.025)	(.033)	(.027)	(.031)	(.022)	(.023)	(.025)	(.020)	(.023)
3+students	-.005	.041	-.076	-.018	-.059	-.040	-.041	-.065	-.028
	(.027)	(.048)	(.048)	(.042)	(.029)	(.031)	(.034)	(.028)	(.031)
F-test	.546	.368	.097	.990	.215	.461	.506	.113	.807

Notes: Panel A reports estimates from regressions of composite EF scores, EF components and non-cognitive outcomes on an indicator for being randomly assigned to a classroom with a poorly-behaved student for various lags of year of assignment to a classroom with a poorly-behaved student. Data on EF is only available up to grade 4, thus the EF regressions pool information across grades 3-4. Data on non-cognitive outcomes are only available at the end of grade 6. In each regression, we regress the outcome on an indicator variable for being assigned to a classroom with a poorly-behaved student, controlling for a fourth-order polynomial in lagged achievement, an indicator for a poorly-behaved student, child age and gender, and school (by grade when pooling) fixed effects. Panel B reports estimates from regressions of composite EF scores, EF components and non-cognitive outcomes on indicators for the number of poorly-behaved students in the classroom. Data on EF is only available up to grade 4, thus the EF regressions pool information across grades 3-4. Data on non-cognitive outcomes are only available at the end of grade 6. In each regression, we regress the outcome on an indicator variable for being assigned to a classroom with different numbers of poorly-behaved students (omitted category is 0, controlling for a fourth-order polynomial in lagged achievement, an indicator for a poorly-behaved student, child age and gender, and school (by grade when pooling) fixed effects. F-test reports p-values of tests for equality of having different numbers of poorly-behaved students. Standard errors are clustered at the school level throughout.

Table 5: Poorly-behaved students and attrition

Panel A		
	Attritor from t to t+1	
	(1)	(2)
Has poorly-behaved student	-.001	
	(.005)	
Is poorly-behaved student		.024
		(.010)
Panel B		
	Bottom 5 worst behaved	Poorly-behaved
	(1)	(2)
New entrant in t	.007	
	(.004)	
New entrant in t-3		.007
		(.003)

Notes: Panel A, column 1 shows results from a regression of an indicator variable for being an attritor between any grades t and t+1 on an indicator for having a poorly-behaved student in the classroom in t, pooling information across grades. The regression controls for a fourth-order polynomial in lagged ability, an indicator for being a poorly-behaved student, child age and gender, as well as school-by-grade fixed effects. Column 2 shows results from a regression of an indicator variable for being an attritor between any grades t and t+1 on an indicator for being a poorly-behaved student in t, pooling information across grades. The regression controls for a fourth-order polynomial in lagged ability, child age and gender, as well as school-by-grade fixed effects. Panel B, column 1 reports estimates from a regression of an indicator for being among the 5 worst behaved students in the classroom on an indicator for being a new entrant in any given grade, pooling information across grades 1-6. We regress the outcome on an indicator for being a new entrant in that grade, child age and gender, and school-by-grade fixed effects. Column 2 reports estimates from a regression of an indicator for being poorly-behaved in a given grade t on an indicator for being a new entrant in grade t-3, pooling information across grades 4-6. In each regression, we regress the outcome on an indicator for being a new entrant 3 years before, child age and gender, and school-by-grade fixed effects. Standard errors are clustered at the school level throughout.

Appendix A

An important assumption underlying our empirical strategy is that poorly-behaved students are not purposefully matched to classrooms, due to random assignment of children to classrooms within schools in

every year.[23] Random assignment is closely monitored, and compliance was very high, 98.9 percent on average. In this appendix, we present tests of random assignment using a methodology developed in Jochmans (2023).

First, we briefly discuss the procedure outlined in Jochmans (2023). Consider our setting, in which we observe data on S schools, and each school has $n_1, \dots, n_s$ students. Within each school, children are assigned to a classroom—and therefore their peer group—every year. Let $x_{s,i}$ be an observable characteristic of child i in school s. If assignment to peer groups is random, $x_{s,i}$ will be uncorrelated with $x_{s,j}$, for all j belonging to the set of $i's$ classroom peers. Let $\bar{x}_{s,j}$ be the average value of characteristic x among student $i's$ peers. The procedure tests whether the correlation in a within-school regression of $x_{s,i}$ on $\bar{x}_{s,i}$ is statistically significantly different from zero (a methodology first proposed in Sacerdote (2001)), introducing a bias correction for the inclusion of group fixed effects (in our case, schools). It is important to control for school fixed effects, as randomization happens within schools, but there may be selection into a school based on individual characteristics. Jochmans (2023) shows that a fixed-effects regression of $x_{s,i}$ on $\bar{x}_{s,i}$ will yield biased estimates due to inconsistency of the within-group estimator. The proposed corrected estimator is given by

$$ts = \frac{\sum_{s=1}^{S} \sum_{i=1}^{n_s} \tilde{x}_{s,i} \left(\bar{x}_{s,j} + \frac{x_{s,i}}{n_s - 1} \right)}{\sqrt{\sum_{s=1}^{S} \left(\sum_{i=1}^{n_s} \tilde{x}_{s,i} \left(\bar{x}_{s,j} + \frac{x_{s,i}}{n_s - 1} \right) \right)^2}} \qquad (A.1)$$

where $\tilde{x}_{s,i}$ is the deviation of $x_{s,i}$ from its within-school mean. The null hypothesis is thus absence of correlation between $i's$ characteristics and those of her peers. To test the random assignment in our setting, we implement this procedure by testing for the presence of correlation between child $i's$ scores measured at the end of grade $t-1$ and the average end-of-grade scores in $t-1$ of the classroom peers assigned to her in a given grade t. We do so for each grade, for math and language achievement as well as executive function. The results are shown in tables A1, A2 and A3 . Note that, to check random assignment in kindergarten, we use TVIP scores collected at baseline. Our results show that we cannot reject the null hypothesis that there is no correlation between child $i's$ achievement and that of her classroom peers. Hence, we conclude that random assignment was successful in our setting.

[23] We use the word "random" as shorthand but, as discussed at length in Araujo et al. (2016), strictly speaking random assignment only occurred in 3rd through 6th grade. In the other grades, the assignment rules were as-good-as-random. Specifically, the assignment rules we implemented were as follows: In kindergarten, all children in each school were ordered by their last name and first name, and were then assigned to teachers in alternating order; in 1st grade, they were ordered by their date of birth, from oldest to youngest, and were then assigned to teachers in alternating order; in 2nd grade, they were divided by gender, ordered by their first name and last name, and then assigned in alternating order; in 3rd through 6th grades, they were divided by gender and then randomly assigned to one or another classroom.

Table A1: Testing for random assignment of children to classrooms, math

	Kindergarten	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6
Test statistic	1.36	-.550	1.04	.104	-.749	.304	.720
P-value	.174	.583	.299	.917	.454	.761	.471

Notes: In this table, we report results for tests of random assignment of children to classrooms within schools using a methodology proposed by Jochmans (2023). The null hypothesis is absence of correlation between a child's math ability measured at the end of the previous grade and the average math ability of classroom peers assigned to her at the beginning of a given grade, conditional on school. To check random assignment in kindergarten, we use TVIP scores collected at baseline.

Table A2: Testing for random assignment of children to classrooms, language

	Kindergarten	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6
Test statistic	1.36	-2.89	-.674	.231	-.383	-.780	-.084
P-value	.174	.004	.501	.818	.702	.435	.933

Notes: In this table, we report results for tests of random assignment of children to classrooms within schools using a methodology proposed by Jochmans (2023). The null hypothesis is absence of correlation between a child's language ability measured at the end of the previous grade and the average language ability of classroom peers assigned to her at the beginning of a given grade, conditional on school. To check random assignment in kindergarten, we use TVIP scores collected at baseline.

Table A1: Testing for random assignment of children to classrooms, EF

	Kindergarten	Grade 1	Grade 2	Grade 3	Grade 4
Test statistic	1.36	.161	-.083	-.988	-1.04
P-value	.174	.872	.934	.323	.299

Notes: In this table, we report results for tests of random assignment of children to classrooms within schools using a methodology proposed by Jochmans (2023). The null hypothesis is absence of correlation between a child's executive function score measured at the end of the previous grade and the average executive function score of classroom peers assigned to her at the beginning of a given grade, conditional on school. To check random assignment in kindergarten, we use TVIP scores collected at baseline. We only collected data on executive function up to fourth grade.

Appendix B

This appendix presents additional information on test scores, executive function, and non-cognitive skills. Figure B1 presents the univariate densities of our achievement measures, separately by grade. The figure shows that most of the distributions appear to have a reasonable spread and are generally symmetric. One clear exception is math achievement in kindergarten, which is left-censored.

Figure B2 presents comparable densities for executive function. It shows that the distributions of inhibitory control and cognitive flexibility are often highly skewed. This is not surprising given the nature of the tests. As an example, we describe the executive function tests we applied in kindergarten.

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

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

The working memory test had two parts. In the first part, children were given 2 minutes to find as many sequences of dog, house, and ball, in that order, on a sheet that has rows of dogs, houses, and balls in various possible sequences. The score on this part of the test is the number of correct sequences found by the child. In the second part of the test, the enumerator recited strings of numbers, and asked the child to repeat them, in the same order or backwards. Figure B2 shows that the aggregate working memory score is distributed smoothly, with little evidence of a concentration of mass at particular values.

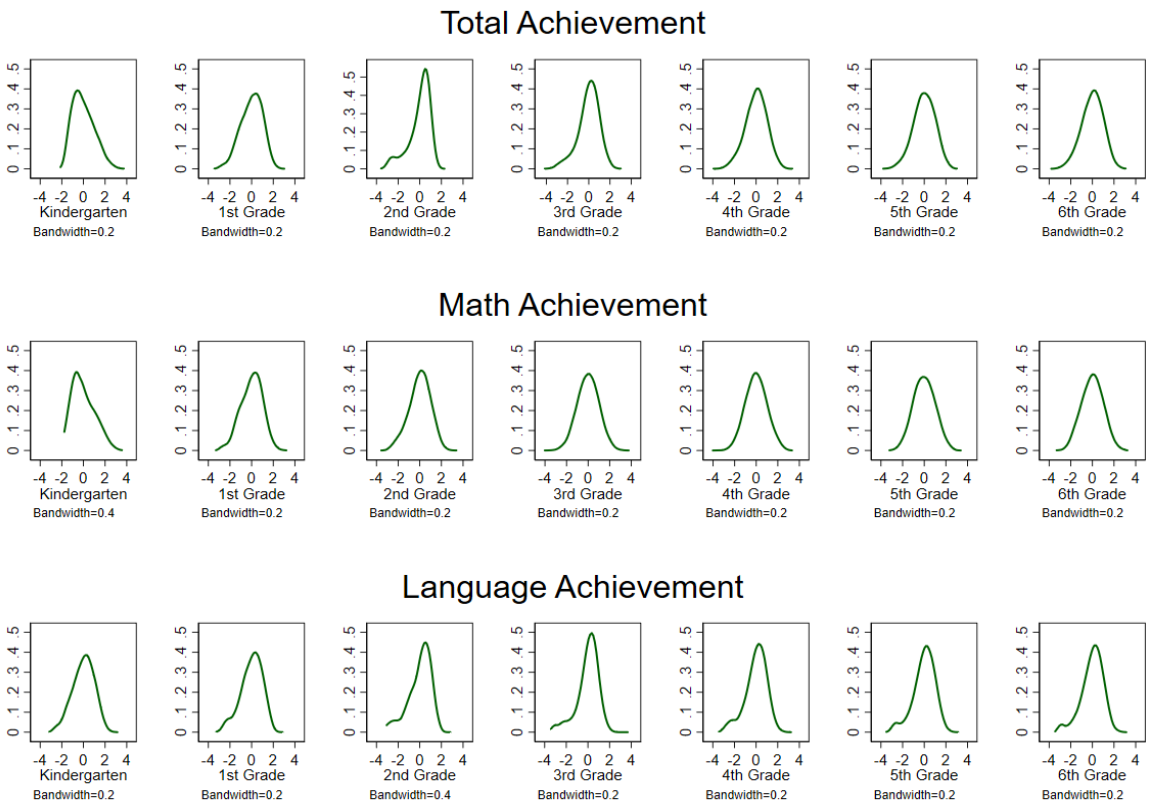
In practice the correlations of the scores across the three dimensions in our sample are low—in the range of 0.21 to 0.32 between cognitive flexibility and working memory, between 0.17 and 0.33 between

working memory and inhibitory control, and in the range of 0.12 to 0.15 between cognitive flexibility and inhibitory control—see Appendix Table B1.[24] When the scores across the three dimensions are averaged, the distributions of the total executive function score are generally smooth and symmetric.

Figure B3, finally, shows univariate densities of the four non-cognitive measures we applied in 6th grade. The figure shows that the distribution of the depression and grit scores appear to be right-censored. The distribution for the aggregate measure of non-cognitive outcomes, on the other hand, is smooth and symmetric. Table B2 shows that the different non-cognitive outcomes are positively correlated, although the correlations are far from unity—they range from 0.20 (between depression and grit) to 0.49 (between growth mindset and self-esteem).

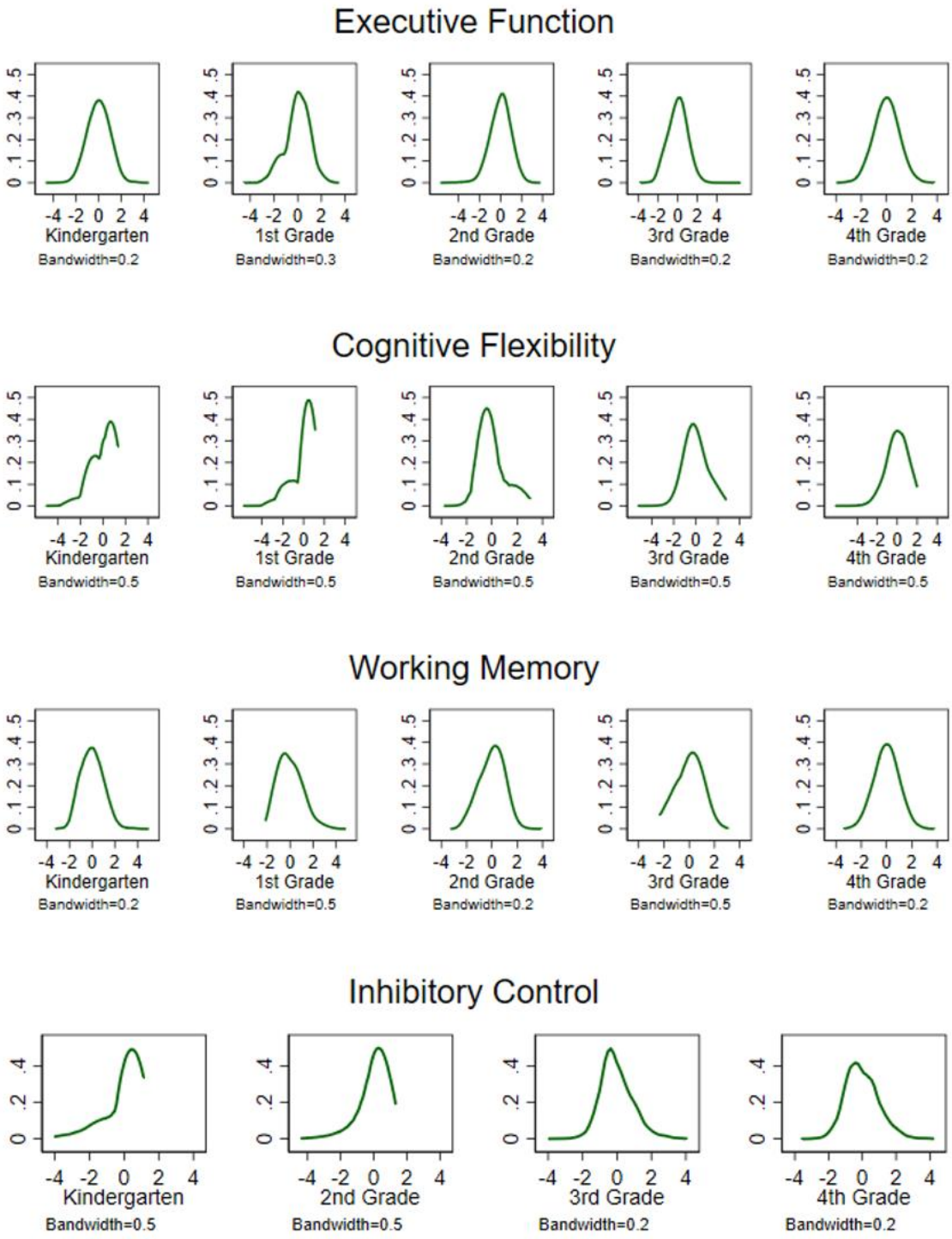
[24] The fact that these correlations are very low is likely to be a result of both measurement error and differences across the constructs that each domain measures.

Figure B1: Distributions of achievement, by grade



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

Figure B2: Distributions of executive function, by grade



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

Figure B3: Distributions of non-cognitive outcomes

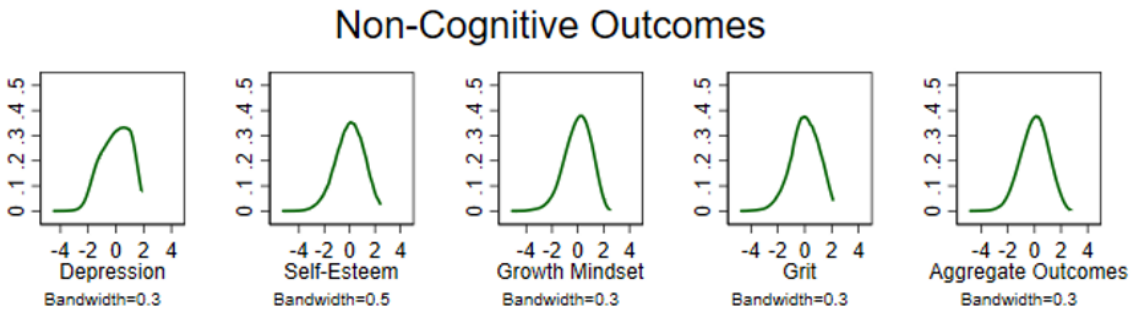


Table B1: Correlations across dimensions in executive function

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

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

Table B2: Correlations across non-cognitive outcomes

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

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

Appendix C

Table C1: Attrition

Panel A		
	Attritor from t to t+1	
	(1)	(2)
Has low-achieving student	-.003	
	(.003)	
Is low-achieving student		.021
		(.010)
Panel B		
	Bottom 5 achieving	Low-achieving
	(1)	(2)
New entrant in t	.011	
	(.004)	
New entrant in t-3		.005
		(.003)

Notes: Panel A, column 1 shows results from a regression of an indicator variable for being an attritor between any grades t and t+1 on an indicator for having a low-achieving student in the classroom in t, pooling information across grades. The regression controls for a fourth-order polynomial in lagged ability, an indicator for being a low-achieving student, child age and gender, as well as school-by-grade fixed effects. Standard errors are clustered at the student and classroom level. Column 2 shows results from a regression of an indicator variable for being an attritor between any grades t and t+1 on an indicator for being a low-achieving student in t, pooling information across grades. The regression controls for a fourth-order polynomial in lagged ability, child age and gender, as well as school-by-grade fixed effects. Panel B, column 1 reports estimates from a regression of an indicator for being among the 5 lowest achieving students in the classroom on an indicator for being a new entrant in any given grade, pooling information across grades 1-6. Standard errors are clustered at the student and classroom level. We regress the outcome on an indicator for being a new entrant in that grade, child age and gender, and school-by-grade fixed effects. Column 2 reports estimates from a regression of an indicator for being low-achieving in a given grade t on an indicator for being a new entrant in grade t-3, pooling information across grades 4-6. Standard errors are clustered at the student and classroom level. We regress the outcome on an indicator for being a new entrant 3 years before, child age and gender, and school-by-grade fixed effects.