Peer Effects and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya

Esther Duflo[1], Pascaline Dupas[2], and Michael Kremer[3, 4]

April 23, 2008

Abstract

This paper provides experimental evidence on the impact of tracking primary school students by initial achievement. In the presence of positive spillover effects from academically proficient peers, tracking may be beneficial for strong students but hurt weaker ones. However, tracking may help everybody if heterogeneous classes make it difficult to teach at a level appropriate to most students. We set up a randomized evaluation in Kenya to evaluate these competing claims. A total of 140 primary schools received funds to hire an extra teacher and create an additional section in first grade. In 70 (randomly selected) schools, students were randomly assigned to a section. In the remaining 70 schools, students were ranked by prior achievement (measured by their first term grades), and the top and bottom halves of the class were assigned to different sections. In all 140 schools, the teachers were randomly assigned to a section. After 18 months, students in tracking schools performed on average 0.13 standard deviations higher than students in the non-tracking schools. Furthermore, students benefited from tracking at all levels of the distribution. A regression discontinuity design analysis shows that, in tracking schools, the endline test score of the initially-median student is as high when she was assigned to the "bottom" section as when she was assigned to the "top" section. In non-tracking schools, where peers were randomly assigned to each student, we find no impact of the average peer quality either, but we find evidence there as well that heterogeneity hurts test scores.

[1] Abdul Latif Jameel Poverty Action Lab (JPAL), Massachusetts Institute of Technology, Paris School of Economics, NBER, and BREAD (eduflo@mit.edu).
[2] Dartmouth College, JPAL and BREAD (pascaline.dupas@dartmouth.edu).
[3] Harvard University, Brookings, CGD, JPAL, NBER, and BREAD (mkremer@fas.harvard.edu).
[4] We thank the Kenya Ministry of Education, Science and Technology, International Child Support Africa, and Matthew Jukes for their collaboration. We thank Jessica Morgan, Ian Tomb, Paul Wang, and especially Willa Friedman for excellent research assistance and seminar participants at Harvard/MIT, Dartmouth, Duke Development Workshop, and NEUDC 07 for comments. We are grateful to Grace Makana and her field team for collecting all the data. We thank the World Bank and the Government of the Netherlands for the grant that made this study possible.

1. Introduction

The desirability of tracking students by prior achievement is the subject of much controversy among academics and policy makers. On the one hand, if teachers find it less difficult to teach a homogenous group of students, tracking could improve school effectiveness and test scores. Many argue, on the other hand, that if students learn in part from their peers and so benefit from having higher achieving peers, tracking could disadvantage the low achieving students while benefiting the high achieving students, thereby exacerbating inequality. While most of the debate on tracking has been focused on the United States or Europe, it is probably even more pressing in many developing countries, where the influx of new students brought to school by efforts to universalize primary education has led to very large, very heterogeneous classes. A key policy question these countries are facing is how to allocate students to maximize efficiency while ensuring that the first generation students are learning.

Direct evidence on the effect of tracking on the achievement of students of various ability levels is mixed. Betts and Shkolnik (1999) review the existing literature and conclude that, while the emerging consensus is that high ability students do better in tracking schools than in non-tracking schools but low ability student do worse, this consensus is largely based on the wrong comparison. Most of the papers they review had been comparing the top students or the bottom students in tracking schools to the *average* students in non-tracking schools. When they compare students of similar ability levels in both tracking and non-tracking high schools, Betts and Shkolnik (1999) find that low ability students are neither hurt nor helped by tracking; top students are helped; and there is some evidence that middle students may be hurt.

The central challenge in identifying the impact of tracking on performance is that schools that track students may be different in many respects from schools that do not. For example, they are likely to have different pedagogies and attract a different pool of students. The ideal experiment to measure the impact of tracking on test scores at various levels of the distribution is thus to randomly assign students to tracking or non-tracking schools, and compare the performance of students across school types at different level of baseline achievement level (as measured before the achievement tracking). We were able to design and implement exactly such an experiment in the context of a class size reduction program in Kenya. In 2005, 140 primary schools in Western Kenya received funds from the World Bank to hire an extra teacher, allowing them to split their first grade class into two sections (or in very few cases, to go from 2 to 3 sections).[5] In 70 of these schools (randomly selected), students were randomly assigned to a section. In the remaining 70 schools, students were divided by prior achievement on the basis of their first term grades (the program started in the second term of the school year). The new and the existing teachers were then randomly assigned the "high prior achievement" or the "low prior achievement" section. Unless they repeated a grade, students stayed in the same section, with the same teacher, for the last two terms of grade 1 and for all of grade 2. After 18 months, the research team administered a comprehensive achievement test in all the schools.

The results suggest that, in the Kenyan context, tracking is beneficial for all students. On average, test scores were 0.13 standard deviations higher in tracking schools than in non-tracking schools (0.16 after controlling for baseline scores). After controlling for the baseline score, students in the top half of the pre-assignment distribution gained 0.17 standard deviations, and those in the bottom half of the pre-assignment distribution

[5] The program is described in more detailed in Duflo, Dupas and Kremer (2007).

gained 0.15 standard deviations. Students in *all quantiles* benefited from tracking, and those in the middle of the pre-assignment distribution gained as much as those in the bottom or the top of the distribution.

Strikingly, students in the middle of the pre-assignment distribution did not seem affected by which section they were assigned to. Specifically, in tracking schools, we cannot reject that there is no difference in end line achievement between the worst student assigned to the high achievement section and the best student assigned to the low assignment section. More generally, a regression discontinuity design analysis reveals no impact of being assigned to the high achievement or to the low achievement section for the median student in the pre-assignment distribution. Since the difference in pre-assignment test score between the two groups was 1.6 standard deviations on average at baseline, we are able to reject even a very modest direct effect of the average achievement of peers in this context.

We complete this analysis by an analysis of peer effects exploiting the random assignment of students to a section in the non-tracking schools. While the standard errors are large, making it difficult to draw sharp conclusions, we cannot reject the hypothesis that the average quality of the peer group has no effect. In contrast, we find that students perform significantly less well when they are either more relatively strong students or more relatively weak students in their class, whatever their initial level. These results confirm that students are hurt by heterogeneity, and are thus consistent with the positive impact of tracking.

These results suggest that for the most part students affect their group indirectly, through their impact on teacher behavior. We present some corroborative evidence for this: teachers are more likely to be in class and teaching in tracking schools, especially in the top section. When we disaggregate the test scores into different components, there is

suggestive evidence that students in the bottom half of the class gain comparatively more from tracking in the most basic competencies, and students in the top half of the class gain more from tracking in somewhat more advanced competencies.

This paper is related to a large literature which investigates peer effects in the classroom (e.g., Hoxby, 2000; Zimmerman, 2003; Angrist and Lang, 2004; Hoxby and Weingarth, 2006, Zhang (2008)). While, mainly for data reasons, this literature has mostly focused on "linear in means" specification, there are a few exceptions which are related to our context: Hoxby and Weingarth (2006) use the frequent re-assignment of pupils to schools in Wake county to estimate models of peer effects, and also find that students seem to benefit mainly from having homogenous peers. Angrist and Lang (2004) study the effect of the Metco busing program in Boston on the achievement of students in "host" schools. They find that, overall, host students do not do significantly worse when their school hosts a larger fraction of Metco (i.e., bused in) students, even though the Metco students have markedly lower test scores. Clark (2007) finds no impact on test scores of attending selective schools for marginal students who just qualified for the elite school on the basis of their score. This is consistent with what we find within tracking schools. Likewise, Zhang (2008) finds that girls who won a lottery to attend an elite school in China benefited if they were themselves very strong students beforehand, and were hurt if they were comparatively weak.

The remainder of this paper proceeds as follows. Section 2 describes the study design and data available. Section 3 presents the main results on test scores. Section 4 presents additional evidence on possible channels. Section 5 concludes.

2. The Tracking Experiment: Background, Experimental Design and Data

2.1. Background: Primary Education in Kenya

The education system in Kenya consists of eight years of primary school and four years of secondary school. Like many other developing countries, Kenya has recently made rapid progress toward the Millennium Development Goal of universal primary education by 2015. In part due to the elimination of school fees in 2003, enrollment in primary schools rose from 5.9 million to 7.6 million between 2002 and 2005, an increase of nearly 30 percent (UNESCO, 2006). This is representative of what is happening more generally in sub-Saharan Africa, where the number of new entrants to primary school increased by more than 30 percent between 1999 and 2004 (UNESCO, 2007).

This progress creates its own new challenges, however. First, the influx of new students has raised pupil-teacher ratios. In Western Kenya, for example, the average class size we observed in first grade in 2005 (two years after the introduction of free primary education) was 83 students per class. The median class size was 74. And 28 percent of first grade classes had more than 100 students. These classes are also very heterogeneous: Many of the new students are first generation learners, and have not attended pre-schools (which is neither free nor compulsory in Kenya). Students differ vastly in their age, their level of preparedness to school, and the support they get at home. These challenges are not unique to Kenya. They confront most developing countries where educational attainment rose sharply in recent years.

2.2. Study Design

This study took advantage of a class reduction program intervention and evaluation that involved 210 primary schools from the districts of Bungoma and Butere-Mumias in

Western Province, Kenya. Of these, 140 schools were randomly selected to participate in an "Extra-Teacher" Program (ETP). With funding from the World Bank, a non-governmental organization (ICS Africa) provided each of the 140 selected schools with funds to hire an additional teacher on a contractual basis starting in May 2005, the beginning of the second term of that school year. (The school year in Kenya starts in January and ends in November. It is divided into three terms, with month-long breaks in April and August). Schools were instructed to create one additional section in first grade, which was to be taught by the contractual teacher. Most schools had only one first grade class, and split it in two sections. Schools which already had two classes or more of first graders added one class. The average class size was reduced to 46 students in the 140 schools that received funds for a new teacher (compared to 84 before the program). The program continued for 18 months, for the last two terms of 2005 and the entire 2006 school year, and the same cohort of students remained enrolled in the program (the new teacher was assigned to grade 2 in 2006). Duflo, Dupas and Kremer (2007) present the design of this experiment in more details as well as its basic results.

To learn about the impact of tracking and about the importance of peer effects in the classroom, we overlaid the following intervention on the class size reduction program. In 70 (randomly selected) schools out of the 140 schools which received an extra teacher, grade 1 pupils were randomly assigned (by the research team) to one of the sections. We call these schools the "non-tracking schools". In each of the other 70 treatment schools (the "tracking schools"), the children were divided into sections by achievement level, according to their score on an exam administered by the school at the end of the first term of the school year 2005. In schools that had originally only one grade 1 class, the bottom 50% of the class according to the exam score was assigned to a section (we call this section the "bottom section") and the top 50% of the class was assigned to the other

section (the "top section"). In principle, exceptions could be granted for siblings, or if parents objected. In practice, the assignment was very well respected.[6] The new teacher and the regular teacher were then randomly assigned to a section. In schools that had two sections in grade 1 before the program started (there are only 9 such schools), the entire first grade was reassigned into 3 sections according to the first term exam score (regardless of the student's original section), and teachers were randomly assigned to a section. In the second year of the program, all children not repeating the grade remained assigned to the same group of peers and the same teacher.[7]

Table 1 presents summary statistics for the 140 schools that participated in the class size reduction program.[8] As would be expected given the random assignment, tracking and non-tracking schools look very similar. In tracking schools, there is a very large difference in the average baseline scores in the two groups. (Since the tests are all different from school to school, they are normalized such that the mean is 0 and the standard deviation is 1 in each school.) The average student has a baseline grade of -0.81 in the bottom section, and 0.79 in the top section, a difference of 1.6 standard deviations. Figure 1 shows the average baseline score of a student's classmates as a function of his own baseline score in tracking and non-tracking schools. The average peer quality is not affected by the student's own test score in non-tracking schools but, consistent with the discontinuous assignment at 50% for most schools, there is sharp discontinuity at 50% in the tracking schools. Note that while the baseline exams were administered by each school, and thus are not a comparable competency exam across schools, they nevertheless

[6] We use the initial assignment regardless of which section the student eventually joined.
[7] Students enrolled in Grade 2 in 2005 and who repeated Grade 2 in 2006 were randomly assigned to either the contract teacher or the regular teacher in 2006. They are excluded from the study. Students who repeated grade 1 in 2006 remain in the data set, and we are using the initial assignment.
[8] New pupils who joined the school after the introduction of the program were assigned to a class on a random basis. However, since the decision for these children to enroll in a treatment or control school might be endogenous, they are excluded from the analysis. The number of newcomers was balanced across school types (tracking and non-tracking) and rather limited, at 6 per school on average.

seem to be a good measure of academic achievement. They are strongly correlated with the endline test we administered: the correlation is 0.47. The average endline score in the bottom section in tracking school was -0.40 standard deviations, and the average endline score in the top section was 0.39 standard deviations. In tracking schools, the top section has somewhat more girls, and the average age (measured at the end of the program, in Table 1) is higher by almost a year.

2.3 Data

The sample frame for this study includes about 7,000 students who were enrolled in grade 1 in March 2005 in one of 138 primary schools enrolled in the study.[9] Slightly less than half are girls (48.8%). Students were 7 years old on average at the onset of the program (with a standard deviation of 1.3), with ages ranging from 5 to 14.

The key outcome of interest is student academic achievement, as measured by their scores on a standardized math and language test administered in all schools 18 months after the start of the program. The test was administered by trained enumerators and graded blindly by data processors. In each school, 60 students (30 per section) were drawn from the initial sample to participate in the tests. If a section had more than 30 students, students near the middle of the initial distribution were sampled with probability 1, while other students were randomly sampled after stratifying by their position in the initial distribution. The test was designed by a cognitive psychologist to measure a range of competencies students may master at the end of grade 2. One part of the test was written and the other part was oral, administered one-to-one. Students were asked math and literacy questions ranging from counting and identifying letters to subtracting three-digit numbers and reading and understanding sentences.

[9] Of the 140 schools in program, 2 are dropped (1 tracking, one non-tracking) because we could not administer the endline test there.

To limit attrition, enumerators were instructed to go to the homes of students who had dropped out or were absent on the day of the test, and to bring them to school for the test. This was not always possible, however. Appendix Table 1 shows that attrition on the test was 18% on average, though it is not significantly different in tracking and non-tracking schools. Students in tracking schools were as likely to have been transferred to a new school as students in non-tracking schools. In total, we have endline test score data for 5,841 students.

In addition, we collected data on grade progression and dropout .Overall, the dropout rate among Grade 1 students in our sample was low (below 0.5%). Finally, each school received un-announced visits several times during the course of the study. During these visits, the enumerators checked, upon arrival, whether teachers were present in school and whether they were in class and teaching, and then took a roll call of the students.

Note that we do not have baseline achievement test scores that can be compared across schools, although we have collected data on the position in the grade just before the start of the program.

2.3 Empirical Strategy

2.3.1 Tracking

Given this set up, measuring the overall impact of tracking on test scores is straight-forward. We run regressions of the form:

(1) $$y_{2ij}=\alpha T_j+X_{ij}\beta+\varepsilon_{ij}$$

where y_{2ij} is the endline test score of student i in school j (expressed in standard deviation of the distribution of scores in the 70 schools that did not receive an extra teacher), T_j is a dummy equal to 1 if school j was tracking, and X_{ij} is a vector of child and school control variables, and a constant (we include a specification without control variables, and a

specification with baseline score, whether the child was in the bottom half of the distribution in the school gender, age, and whether the section is taught by the new teacher or the regular teacher). To identify whether children who were assigned to the bottom section have differential effects, we also run:

$$(2) \qquad y_{2ij}=\alpha T_j+\gamma T_j*B_{ij}+X_{ij}\beta+\varepsilon_{ij}$$

where B_{ij} is a dummy variable which indicate whether the child was in the bottom half the baseline score distribution in her school (recall that B_{ij} is included in the vector X_{ij}). Finally, to investigate flexibly whether the effects of tracking are different at different levels of the initial test score distribution, we run two separate non-parametric regressions of endline test scores on baseline test scores in tracking and non-tracking schools, and we plot the results.

To understand better how tracking works, we then run similar regressions using as dependent variable a more disaggregated version of the test scores: the test scores in math and language, and the scores on specific competencies. Finally, we also run regressions of a similar form, using as outcome variable teacher presence in school, whether the teacher is in class teaching, and student presence in school.

2.3.2 Peer effects

This set up provides two separate opportunities to identify peer effects in the classroom.

a) Regression Discontinuity Design

Tracking schools provide a natural set up for a regression discontinuity (RD) design estimate of the impact of peer average quality. As shown in Figure 1, the two students close to the median were assigned to classes where the average level of their classmates was very different: the one with the lowest score of the pair was assigned to the bottom section, and the one with the highest score of the pair was assigned to the top section

(when the class had an odd number of student, the median student was randomly assigned to one section or the next).

Thus, we first estimate the following reduced form regression in tracking schools:

$$y_{2ij}=\delta B_{ij}+\lambda_1 P_{ij}+ \lambda_2 P_{ij}^2+ \lambda_3 P_{ij}^3+X_{ij}\beta+\varepsilon_{ij} \quad (3)$$

where P_{ij} is the percentile of the child in the distribution of the baseline grade in his school.

Since the assignment was done within each school, we also run the same specification, including school fixed effects:

$$y_{2ij}= \delta B_{ij}+\lambda_1 P_{ij}+ \lambda_2 P_{ij}^2+ \lambda_3 P_{ij}^3+X_{ij}\beta +\upsilon_j+\varepsilon_{ij} \quad (4)$$

Finally, we run similar specifications as equations (3) and (4), but allowing the polynomial to be estimated separately for each side of the discontinuity.

Note that this is an unusually favorable set up for an RD design. There are 60 discontinuities in the data set (one per tracking school that went for 1 to 2 streams), rather than just one as in most RD applications, and the number of discontinuities in principle grows with the number of observations (since the way to add observations in the data set is to add schools). We can thus do what the RD framework suggests should be done asymptotically (but cannot happen in practice in finite datasets), that is, compare students in an extremely narrow band around each discontinuity. In fact, we run a specification where we include only the pair of students straddling the median (the better student of the pair was assigned to the top section, and the worse student was assigned to the bottom section).

$$y_{2ij}=\delta B_{ij} +X_{ij}\beta+\upsilon_j+\varepsilon_{ij} \quad (5)$$

These reduced form results are of independent interest, but they can also be combined with the impact of tracking on average peer test scores for instrumental variable estimation of the impact of average peer quality for the medium child in a tracking environment. Specifically, the first stage of this regression is:

$$A(y_{2ij})=\pi B_{ij}+\varphi_1 P_{ij}+\varphi_2 P_{ij}^2+\varphi_3 P_{ij}^3+ X_{ij}\beta +\varepsilon_{ij,}$$

where $A(y_{2ij})$ is the average endline test scores of the classmates of student i in school j. The structural equation:

$$y_{2ij}= \kappa A(y_{2ij})+ \lambda_1 P_{ij}+ \lambda_2 P_{ij}^2+ \lambda_3 P_{ij}^3+ X_{ij}\beta +\varepsilon_{ij,}$$

is estimated using B_{ij} as an instrument for $A(y_{2ij})$.

Note that this strategy will give an estimate of the effect of peer quality for the median children in a tracking environment, where teaching method may be affected by the tracking, and where having high achieving peers on average also means that the child is the lowest achieving child of his track (at least at baseline) and having low quality peers means that the child is the highest achieving child of his track. There may be an independent effect of being the best or the worst child in the section. Therefore, the results from this regression will not necessarily extend to set ups where there is no tracking.

b) Random Assignment to Peers

The second method to estimate peer effects in our setting is to take advantage of the random variation in peer groups in the non-tracking schools. Since children were randomly assigned to a section in these schools, their peer group is randomly assigned to them.

In the sample of non-tracking schools, we start by estimating the effect of a student's peer average test scores by OLS:

$$y_{2ij} = \kappa A(y_{2ij}) + X_{ij}\beta + \varepsilon_{ij,} + \upsilon_j + \varepsilon_{ij},$$

where the vector of control variables X_{ij} includes a measure of the student's own baseline score.

Given the importance of heterogeneity, we also use the sample of non-tracking schools to estimate the impact of other measures of heterogeneity. Specifically, we estimate:

(6) $$y_{2ij} = \mu_1 H_{ij} + \mu_2 L_{ij} + X_{ij}\beta + \varepsilon_{ij,} + \upsilon_j + \varepsilon_{ij},$$

where H_{ij} (resp. L_{ij}) is the share of students in the class of student *i* who were in the top third of the distribution of the pre-assignment school test scores in *school j*. The omitted category is the share of students in the middle of the distribution. λ_1 and λ_2 indicate whether students benefit or are hurt from having more of the strong students and more of the weak students in their section.[10]

3. Results

3.1 The impact of tracking by prior achievement

Table 2 presents OLS estimates of the effect of being in a tracking school on test scores. We find that tracking by initial achievement led to a significant increase in test scores. Students in tracking schools performed 0.138 standard deviations (with a standard error of 0.078) better than students in non-tracking schools overall (table 2, column 1). The effect becomes somewhat larger (0.172 standard deviations, with a standard error of 0.076) when we add individual level control variables (column 2). Note that the baseline position in the roster of test scores has strong predictive content.

[10] The standard errors in this regression are clustered at the class level (post assignment), since the regression includes a school fixed effect.

Table 2 also allows us to see whether the effect of tracking by initial achievement is different for initially strong students (who are paired with other strong students in tracking schools) and initially weak students (who are paired with other weak students). We find that both strong and weak students benefited from tracking (in Row 2, Column 3, the interaction between being in the bottom half and in a tracking school cannot be distinguished from zero).

Columns 4 through 6 and columns 7 through 9 look at the impact of the program on math and language, respectively. There is no significant difference between the two subjects, though the effects are more precisely estimated for math than for language.

Panel B and C look separately at boys and girls. Although the coefficients are not significantly different from each others, the point estimates suggest that the effects are much larger for girls in math. The coefficients are almost twice as large for girls as they are for boys (0.16 standard deviations for girls versus 0.89 standard deviations for boys). In language, the coefficients are similar for boys and for girls. For both boys and girls, initially weaker students benefit as much as initially stronger students.

Overall, the estimates in Table 2 suggest that all students, irrespective of achievement at baseline, benefited from the tracking. Figure 2 provides graphical evidence. It plots a child's endline test score as a function of the baseline test score using a local non-linear regression in both the tracking and non-tracking schools. This figure shows that, both in language and in math, tracking seems to have a beneficial impact regardless of the initial level of the child in the distribution of test scores. If anything, the students initially at the median (who are now at the top or the bottom of their section), seem to have benefited more from the tracking than the students initially at the 33% or 66% percentile, who became the median students in the tracking schools.

3.2 Peer effects: Regression Discontinuity Design

The main thrust of the results on peer effects are shown in Figure 3. In Panel A, following Lee (2007), we regress test scores on a third-order local polynomial in initial percentile separately for students below the 50th percentile cutoff and students above the 50th percentile. Each point is an average of the test scores for each percentile of the initial distribution. The vertical line represents the cutoff line for being assigned to the bottom section in tracking schools (being at the 50th percentile score).[11] As is apparent from the figure, there is no discontinuity at the vertical line in the schools tracking by initial achievement, despite the strong discontinuity in peer attainment observed in Figure 1 (a difference of 1.6 standard deviations in the baseline scores). The data exhibit a continuous and smooth relationship throughout the. When we use a linear fit, rather than a polynomial, we again do not see an effect of the group in which the students were placed for students in the middle of the distribution (figure not shown). In Panel B, we use Fan locally weighted regressions with a biweight kernel and a bandwidth of 2.0. Again, we see no discontinuity at the threshold for being assigned to the bottom stream.

We examine this result in a regression framework in Table 3, where we estimate equations (3) through (5) in the sample of tracking schools. Panel A reports the reduced form coefficient of being in the bottom half of the class (controlling for a cubic polynomial in initial percentile). Column 1 and 3 use all the students. Column 3 introduces a school fixed effect (since the assignment was decided within school). Columns 2 and 4 use only two students per school, the best students assigned to the bottom section, and the worst student assigned to the top section (we have 60 such pairs, since we have 60 schools that went from 1 to 2 sections and for which we have endline test data). The results confirm what the graphs showed: despite the big gap in average

[11] Schools which went from 2 to three sections, where the cutoff is not at the 50th percentile, are excluded from this graph.

peer achievement (1.6 standard deviations of the baseline school grade), there appears to be no penalty for the marginal student to be assigned to the bottom section.

Panel B shows the instrumental variable estimate of the impact of classmates' test average score on a child's test score. We use the average endline score of classmates (because the baseline scores are school specific), and instrument it using the dummy for being in the "bottom half" of the initial distribution as the instrument. The first stage is shown in panel C, and shows that the average of the endline test scores of a child's classmates is about 0.87 standard deviations lower if she was assigned to the bottom section in a tracking school. The IV estimates in panel B are all small and insignificant. Our preferred specification is column 3, which has school fixed effects and uses all the data. It suggests that an increase in one standard deviation in the classmates' test score *reduces* a child's test score by 0.007 standard deviations, a point estimate extremely close to zero. The 95% confidence interval in this specification is [-0.18; 0.17]. Thus, we are able to reject at 95% reasonably modest direct effects of the average of one's peers on a student's performance.

Table 4 presents the results of reduced form regressions similar to equation 1, but in a sample which includes both tracking and non-tracking schools. The regression includes a dummy for being in a tracking school, and the interaction between being in a tracking school and in the bottom half of the initial distribution. The polynomial in initial percentile in one's school is restricted to have the same form in tracking and non-tracking schools:

$$y_{2ij}=\alpha T_j+\gamma T_j * B_{ij} + \lambda_1 P_{ij}+ \lambda_2 P_{ij}^2+ \lambda_3 P_{ij}^3+ X_{ij}\beta +\varepsilon_{i,.}$$

The coefficient of T_j in this regression indicates whether the students close to the median do better in tracking than in non tracking schools. We find that they do significantly better (0.194 standard deviations better in the RD specification that uses all the data). The

coefficient of B_{ij} indicates whether there is a discontinuous jump in test scores at the 50^{th} percentile for student in non tracking schools (as we may expect, there is no such discontinuity), while the coefficient of $T_j * B_{ij}$ indicates whether there is a differential jump in tracking schools. Such a differential jump in tracking schools would suggest the presence of peer effects. We do not observe any.

These results are quite striking. They imply that being the best students in a class of relatively weak students or being the worst student in a class of relatively strong students does not matter, but that being the middle student in a heterogeneous class is not as good. This rejects a model of peer effects that is purely "linear in means", as well as a model where heterogeneity hurts because the teacher is teaching to the median student in the class. It suggests that students benefit from homogeneity because the teacher can better tailor her teaching to the entire class group (called the "boutique" or "focus" models of teaching by Hoxby and Weingarth, 2006). We will provide some additional evidence of what could be happening in the classroom in Section 4. We now turn to alternative estimate of peer effects, based on the random assignment of students to sections in the non-tracking schools.

3.3. Peer effects: Random Variation in Peer Composition

The regression discontinuity design approach we have discussed has the advantage of generating large differences in average peer initial achievement, but the drawback is that we can only look at its effect for children in the middle of the distribution. The effect of peer initial achievement could potentially be very different for children with different initial levels of achievement. The evaluation generated another source of random variation in the average achievement in the peer group, which we attempted to exploit. In non-tracking schools, children were randomly assigned to either class and very few re-

arrangements between classes took place. As shown in Figure 4, this generates a fair amount of random variation in the composition of the different classes. We can thus implement methods to evaluate the impact of the composition of a class similar to those introduced by Hoxby (2000), with the difference that we use actual random variation in peer group and have lower sample size. The results of various specifications are presented in Table 5. Unfortunately the standard errors are quite large. While we cannot reject that there are no peer effects (the point estimate in the regression with school fixed effect are 0.17 with a standard error of 0.14), we also cannot reject relatively large effects of peers (of the order of magnitude found in the previous literature).

Given the positive impact of tracking, it is likely, however, that this noisy estimates masks heterogeneity in the effect of peers. Table 6 explores this, by estimating equation (6), which regresses the test score of a child on the fraction of the students is her section who came from the top third and the bottom third (respectively) of the distribution of students in the entire school, at the pre-assignment test. Due to the random assignment, these students have ended up in slightly different number in the sections formed in a school after they received the extra teachers. The results in this table are much more clear cut: column (1), (5) and (9) shows that students are hurt when there are more weak students *and* when there are more strong students in their class, both overall, and looking at math and language separately. Furthermore, when we run this regression separately by prior achievement, we find that this is the case for *all* students, although the results are significant only for the top students. Once again, this suggests that what hurts the students is to study in a group where there are too many students at either extreme.

4. Why did tracking work? Exploring the Channels

The basic results suggest that tracking had a positive effect for all children, regardless of the group they were assigned to, and regardless of their place in the initial class distribution. The absence of direct effects of the average quality of the peers suggests that they may have benefited from more focused teaching. In this section, we explore two additional pieces of evidence which shed some light on this question. First, we look at teacher presence in school and effort (do they spend more time in class and teaching?). Second, we disaggregate the test scores gains in tracking schools by competencies.

4.1 Teacher effort

Table 7 shows the result of a regression of teacher presence and effort while in school on tracking (using a specification similar to equation 1, though the set of control variables now include teacher age and teacher experience teaching). We also present the results separately for regular government teachers, and new teachers, because they face very different incentives. The new teachers are on short term (one year) contract, and have incentives to work harder to be renewed, and to have a chance to be hired by the government as regular teachers.

The first three columns in Table 7 show that on average teachers in tracking schools are significantly more likely to both be in school and be in class teaching than teachers in non-tracking schools. Overall, teachers in tracking schools are 9.6 percentage points more likely to be found in school and teaching during a random spot check than their counterparts in non-tracking schools. This corresponds to an increase of almost 19%.

There are, however, large differences across teachers. The new contract teachers attend more than the regular (government-employed) teachers, are more likely to be found in class and teaching (74% versus 45% for the regular teacher), and are unaffected

by tracking. The regular teachers are 5.4 percentage points more likely to be in schools in tracking schools than in non-tracking schools when they were assigned to the top section, and the difference is significant (recall that teacher assignment to each section was random, so this is indeed the causal effect of being assigned to a group of strong students, rather than a non-tracked group). However, the difference disappears entirely for the bottom section: the interaction between tracking and bottom section is minus 7.7 percentage point, and is also significant. Conditional on being in school, regular teachers are also more likely to actually be in class and teaching in tracking schools than in non-tracking schools when they are assigned to the top section, and again, this difference largely disappears in the bottom section. Overall, these teachers are 11 percentage points more likely to be in class and teaching when they are assigned to the top section in tracking school than when they are in non-tracking schools, which represent a 25% increase in teaching time. When they are assigned to the bottom section, they are about as likely to be teaching as their counterparts in non-tracking schools. The students' attendance record is not affected by tracking, nor by the section they were assigned to.

These results suggest that teachers may be more motivated to teach a strong group than a weak or heterogeneous group. However, they also suggest that teacher effort is not the whole story, since teacher effort did not increase in the bottom section, but results improved nevertheless.

4.2 Focused Teaching

One hypothesis consistent with both the tracking results and the results from random peer assignment is that tracking by initial achievement improves student learning because it allows teachers to focus instruction. Teaching a group of more homogeneous students might allow teachers to adjust the pace of instruction to students' needs. For example, a

teacher might instruct at a slower pace, providing more repetition and reinforcement when students are initially less prepared, but with a group of initially higher achieving students, the teacher can increase the complexity of the tasks and pupils can learn at a faster pace. It is unfortunately difficult to obtain direct evidence on how teachers changed the materials they covered in tracking schools. We conducted classroom observations, which we are currently examining to see whether they reveal a change in general teaching methods in the classrooms (measured by the number of times students were asked to speak, the number of times teachers corrected the students' work, etc… in a 30 minutes period). But the classroom observations did not capture the level of instruction. It could be that the teaching style remains the same, but the material covered is more appropriate to the group in a tracking school.

A way to look at this is to see whether children at different levels of the distribution gained from tracking differentially at different skills. Table 8 reports specifications similar to equation (2), but where the test scores are disaggregated by specific competency for math and language. The equations are estimated jointly in a simultaneous equation framework (allowing for correlation between the error terms). There is no clear pattern for language, but for math, the estimates suggest that, while the total effect of tracking on children initially in the bottom half of the distribution (thus assigned to the bottom section in the tracking schools) is significantly positive for all levels of difficulty, these children gained from tracking more than other students on the easiest questions, and less on the more difficult questions: the interaction tracking * bottom half is positive for the easiest competencies, and negative for the hardest competencies. A chi-square test allows us to reject the equality of the coefficient of the interaction in the "easy competencies" regression and the "difficulty competencies" regression at the 6% level.

Conversely, students assigned to the top track benefited less on the easiest questions, and more on the difficult questions (in fact, they did not significantly benefit from tracking for the easiest questions, but they did significantly benefit from it for the hardest questions). Overall, this table suggests that tracking helped by giving teachers the opportunity to focus on the competencies that children were not mastering.[12]

5. Conclusion

This paper provides experimental evidence that students at all level of the initial achievement spectrum benefited from being tracked into classes by initial achievement. Despite the critical importance of this issue for the educational policy both in developed and developing countries, there is surprisingly little rigorous evidence addressing it, and to our knowledge this paper provides the first experimental evaluation of it.

After 18 months, the point estimates suggest that the average score of a child in a tracking school is 0.13 standard deviations higher than that of a child in a non-tracking school. Moreover, a regression discontinuity design approach reveals that children who are very close to the 50^{th} percentile in their school, and thus were assigned either to the low initial achievement track or the high initial achievement track, do as well (and significantly better in tracking schools) regardless of the section they were assigned to.

This suggests that peers matter in the classroom, though what matters most is not their average level of achievement, but their homogeneity. This is reinforced by exploiting the random assignment of peers in the non-tracking schools. In these regressions, we also find no impact of average peer quality, but strong negative impact of the share of either very strong or very weak students.

[12] We also estimated a version of equation (6) allowing the effect to vary by quarter of the distribution for each competencies, and the pattern are very similar, with progressively weaker students befitting the most from tracking for the bottom competencies, and progressively strongest students benefiting the most for the hardest competencies.

Greater homogeneity is presumably beneficial because it allows the teachers to adapt her material better to the students she has to teach. Consistent with this, we find that students assigned to the bottom section in the tracking school gained most in the easiest competencies, and less for the hardest competencies. It may also reduce disturbance (Lazear, 2001). Teacher effort also seems to have increased in tracking schools when the teacher was assigned to the top section, suggesting that teaching a better group of student is more motivating. And while teachers assigned to the bottom section did not work harder than those assigned to heterogeneous classrooms, they did not decrease their level of effort either, suggesting that teachers are at least as motivated when teaching a homogenous group, even if it is weaker on average.

These conclusions echo those reached by Banerjee et al. (2007), who study a remedial education and a computer assisted learning programs in India. They found that both programs were very effective, mainly because they allowed students to learn at their own level of achievement. A central challenge of educational systems in developing countries is that students are extremely diverse, and the curriculum is largely not adapted to new learners. These results show that grouping students by preparedness or prior achievement and focusing the teaching material at a level which is pertinent for them could potentially have large positive effects at no resources cost.[13]

Note that our design did not allow teacher quality to vary with tracking, since teachers were randomly assigned to each section. Class size was also constant. In principle, one could also target more resources to the weaker group, further helping them to catch up with their counterparts. It is often believed that there is a tradeoff between the value of targeting resources to weaker students, and the costs imposed on them by

[13] This is also the conclusion of Glewwe et al. (2007), who finds that textbooks only help students who were doing well to start with, because those textbooks are too hard for most students.

separating them from stronger students–see Piketty (2004) for a discussion of this issue in the French context. This trade-off seems absent in our context.

Also note that in practice the best teachers may request to be assigned to the best group of students, which could potentially hurt the low achievement students. In our experiment, some teachers did complain when they were assigned the weakest group. One issue is that explicit and implicit incentives for teachers in the Kenya system are based on the average performance of the group of students, rather than their progression. To make a tracking system work, the emphasis in evaluating teacher's performance would need to be placed on value added, rather than initial level.

References

Angrist, J. and K. Lang, 2004. “Does School Integration Generate Peer Effects? Evidence from Boston's Metco Program”. *American Economic Review*, Vol. 94, No. 5, 1613-1634.

Banerjee, A., Cole, S., Duflo, E. and Linden, L. (2007). “Remedying Education: Evidence from Two Randomized Experiments in India.” Forthcoming, *Quarterly Journal of Economics*.

Betts, J.R. and Shkolnik, J. 2000. Key difficulties in identifying the effects of ability grouping on student achievement. *Economics of Education Review,* 19 (1), 21-26.

Clark, Damon (2007). “Selective Schools and Academic Achievement”. Mimeo, IZA.

Duflo, E., P. Dupas, and M. Kremer (2007). “Peer Effects, Pupil-Teacher Ratios, and Teacher Incentives: Evidence from a Randomized Evaluation in Kenya”. Mimeo, MIT.

Glewwe, P., M. Kremer and S. Moulin (2007). Many Children Left Behind? Textbooks and Test Scores in Kenya. Mimeo, Harvard.

Hoxby, C. (2000). "Peer Effects in the Classroom: Learning from Gender and Race Variation," NBER Working Papers 7867, National Bureau of Economic Research, Inc.

Hoxby, C. and G. Weingarth (2006). “Taking Race Out of the Equation: School Reassignment and the Structure of Peer Effects.” Mimeo, Harvard University.

ILO/UNESCO (2007). *Joint ILO/UNESCO Committee of Experts on the Application of the Recommendations concerning Teaching Personnel*. Geneva: ILO Publications.

Lavy, V. and A. Schlosser (2007). “Mechanisms and Impacts of Gender Peer Effects at School.” Mimeo, Hebrew University and Princeton University.

Lazear (2001). “Educational Production," *The Quarterly Journal of Economics*, MIT Press, vol. 116(3), pages 777-803.

Lee (2007). “Randomized experiments from non-random selection in U.S. House elections”. *Journal of Econometrics*, 142 675-697.

Piketty, T. (2004). « L'Impact de la taille des classes et de la ségrégation sociale sur la réussite scolaire dans les écoles françaises : une estimation à partir du panel primaire 1997. » Mimeo, PSE.

Sacerdote, Bruce (2001). "Peer Effects With Random Assignment: Results for Dartmouth Roommates." *Quarterly Journal of Economics*, Vol. 116.

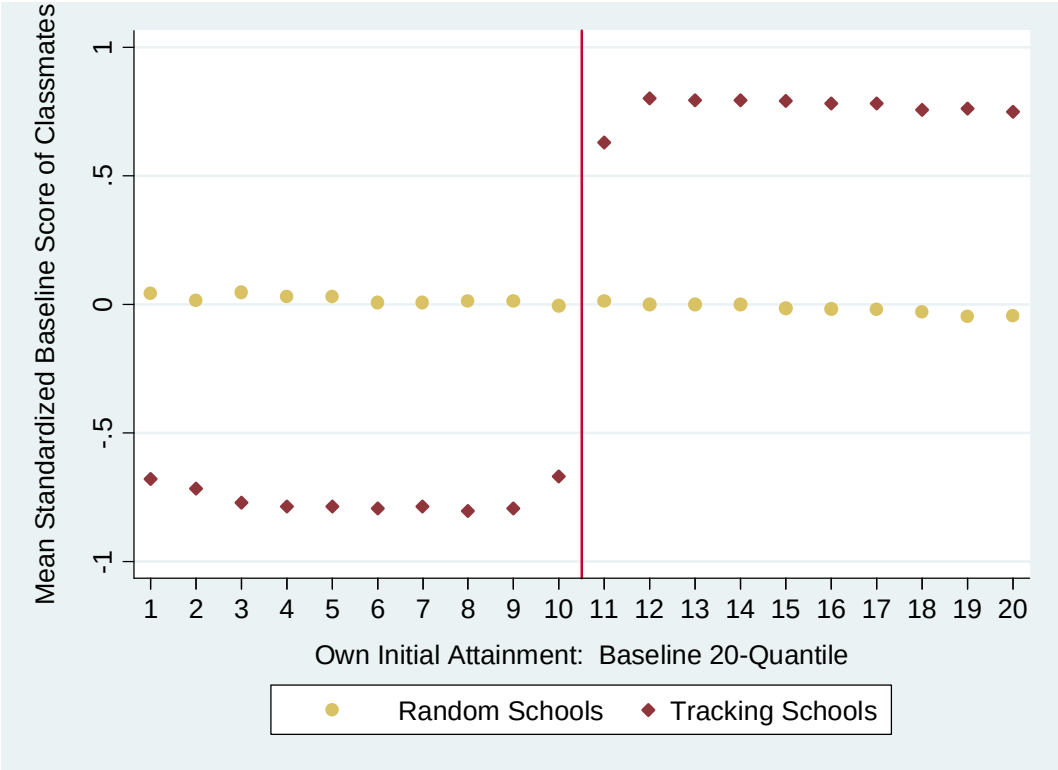
UNESCO (2006). *Fact Book on Education for All.* UNESCO Nairobi Office.

UNESCO (2007). *Strong Foundations: Early Childhood Care and Education.* Paris: UNESCO Publishing.

Zimmerman, D. (2003) "Peer Effects in Academic Outcomes: Evidence from a Natural Experiment," *The Review of Economics and Statistics*, MIT Press, vol. 85(1), pages 9-23.

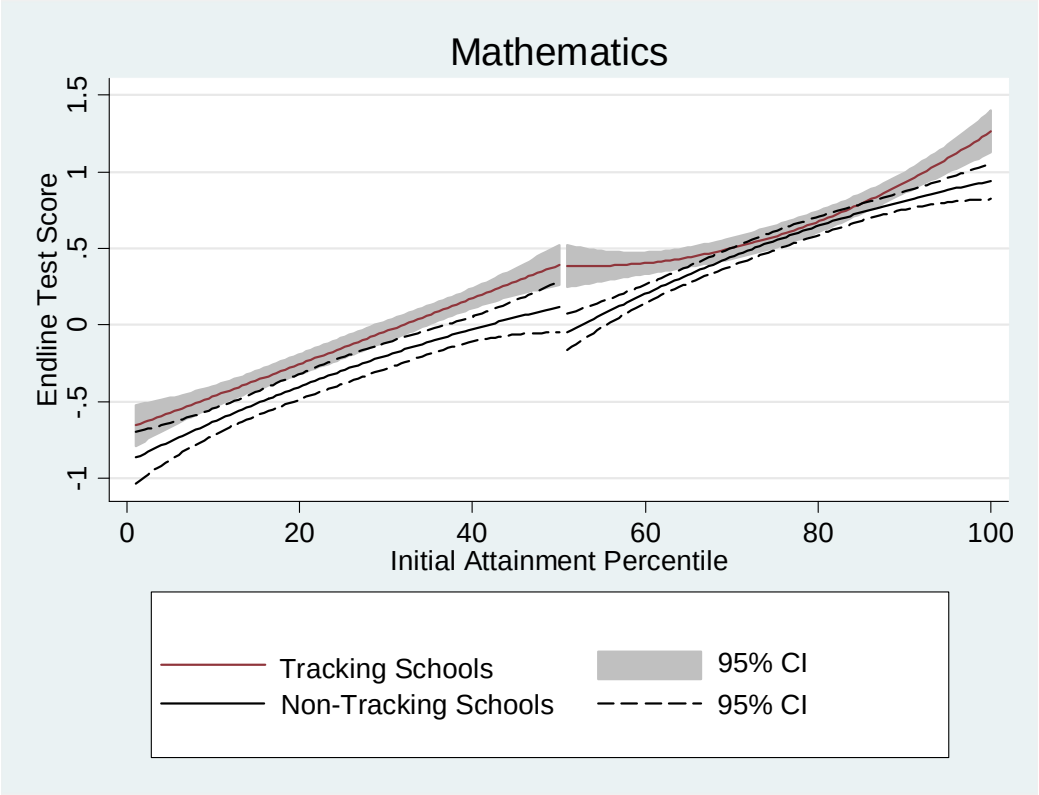
Zhang, H. (2008). “Elite Schools and Student Academic Achievement: Evidence from A Randomized Natural Experiment in China”

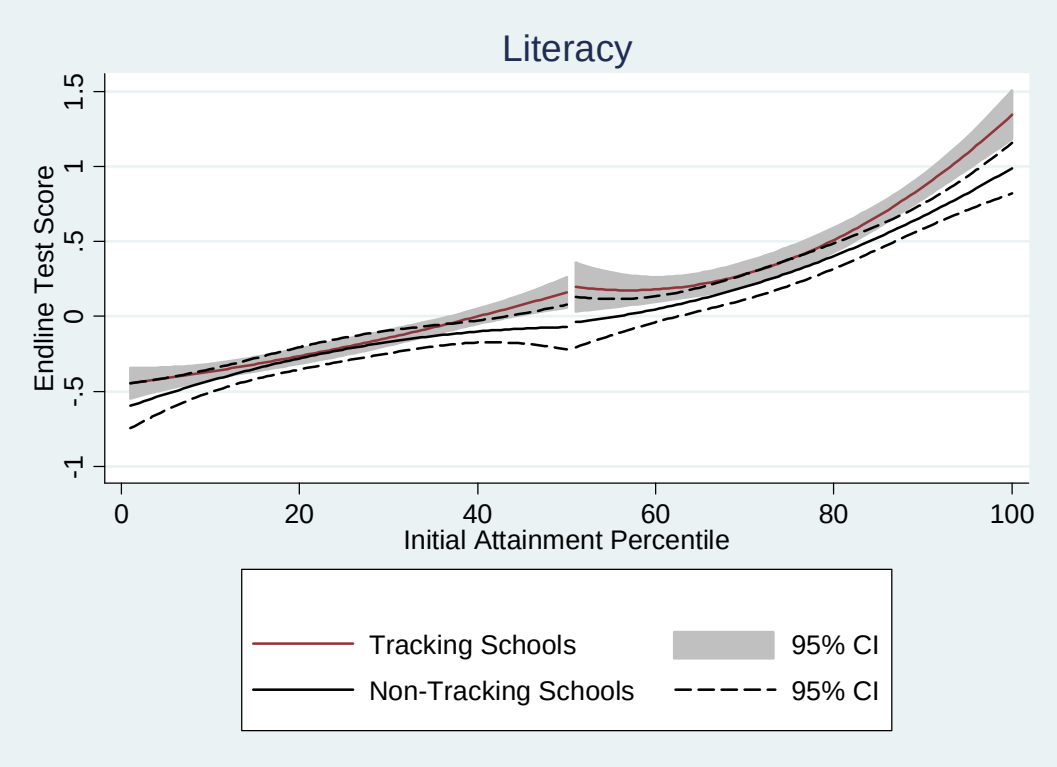
Figure 1
Experimental Variation in Peer Composition: Tracking vs. Non-Tracking Schools



Note: Each dot corresponds to the average peer quality across all students in a given 20-quantile, for a given treatment group.

Figure 2
Local Polynomial Fits of Endline Score by Initial Attainment

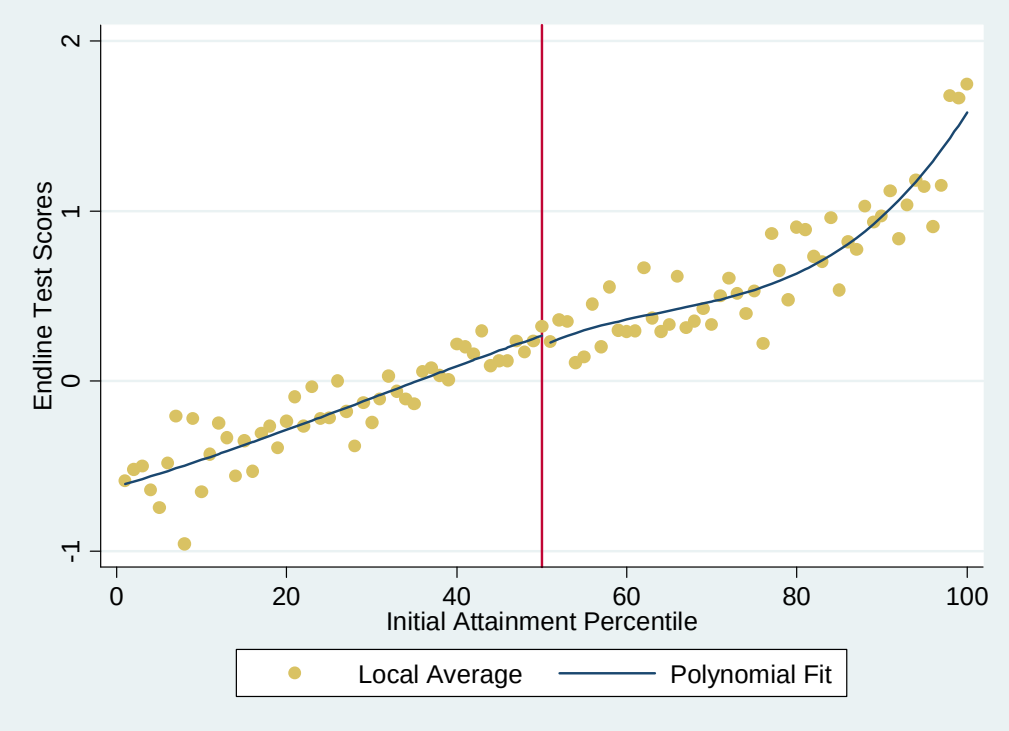




Notes: Fitted values from regressions that include a second order polynomial estimated separately on each side of the percentile=50 threshold.

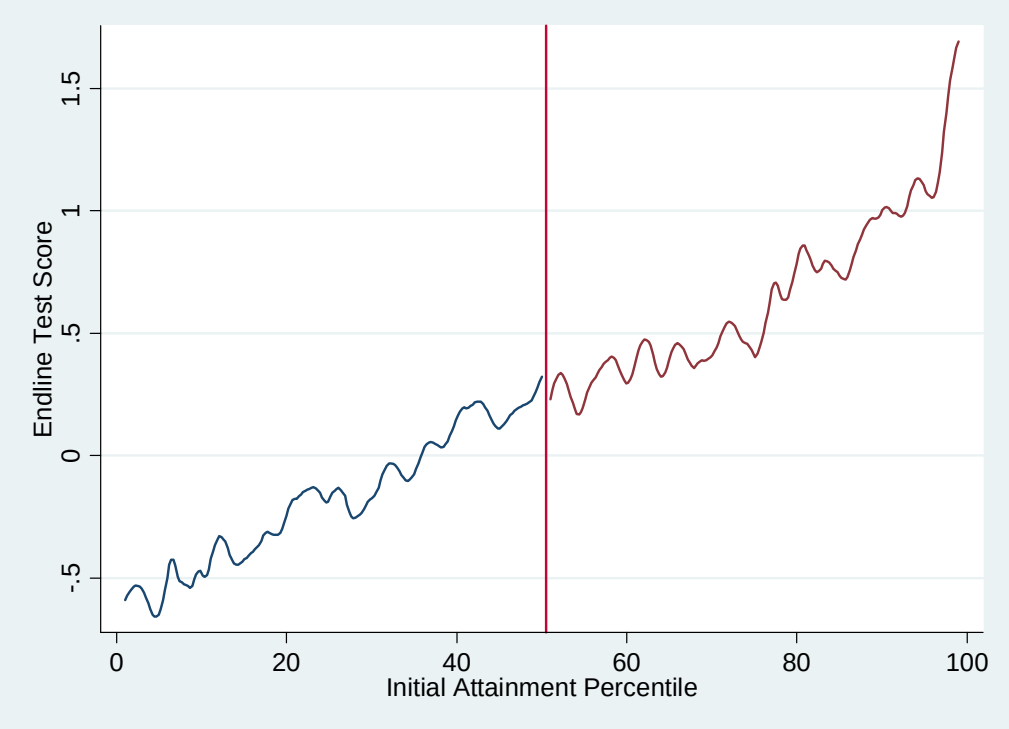
Figure 3

Panel A. Quadratic Fit



Notes: the points are the average score. The fitted values are from regressions that include a second order polynomial estimated separately on each side of the percentile=50 threshold.

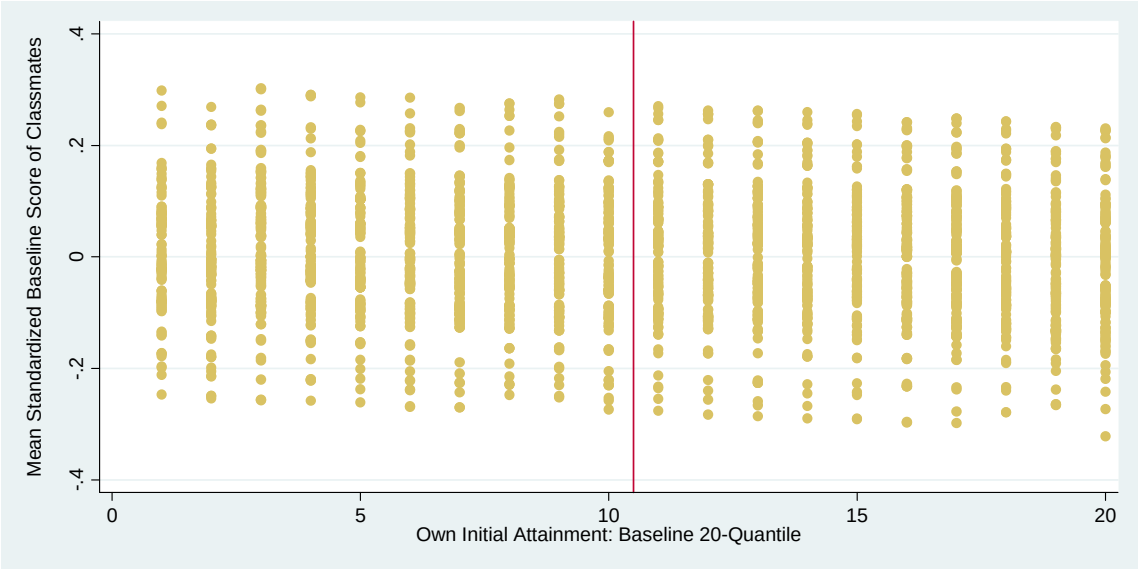
Panel B. Fan's Locally weighted regression



Notes: Fitted values from Fan's locally weighted regressions with quartic (biweight) kernels and a bandwidth of 2.0.

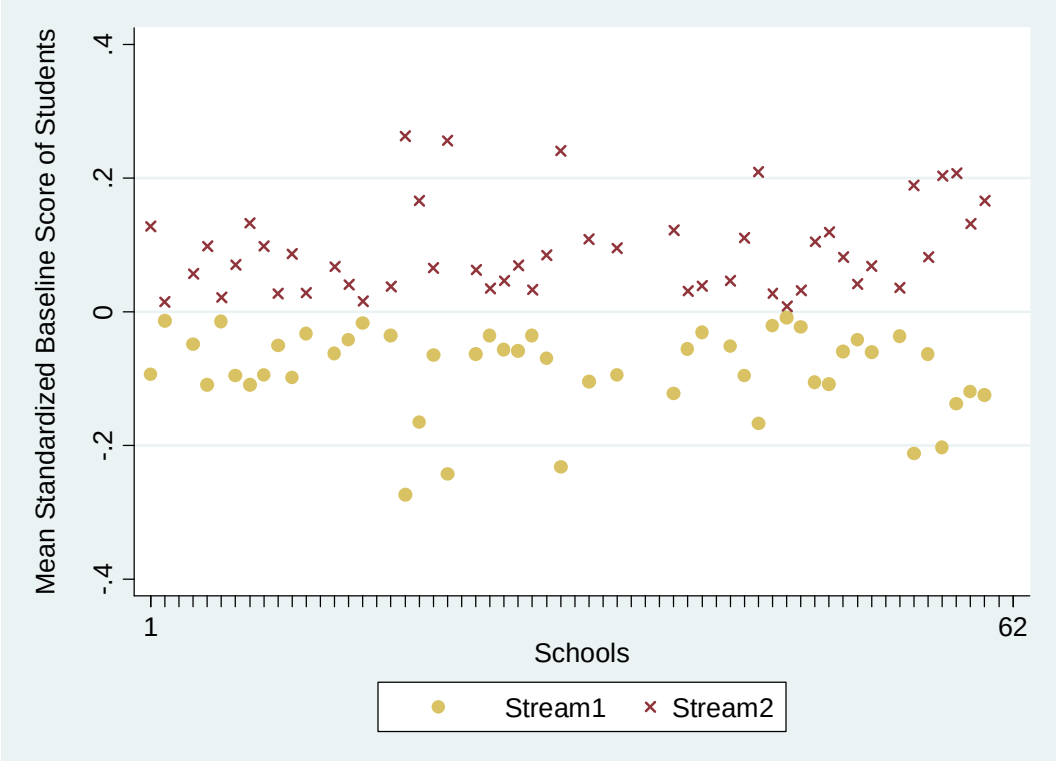
Figure 4
Exogenous Variation in Peer Composition Created by Class Size Reduction

2A. Variation in Peer Composition in Non-Tracking Schools



Note: Each dot corresponds to a student. This graph shows that within each initial attainment 20-quantile, there is variation in the peer quality.

2B. Within-School Variation in Peer Composition in Non-Tracking Schools



Note: Schools are ordered alphabetically long the x axis. The graph displays two data points per school (one for each section).

Table 1
School and Class Characteristics, by Treatment Group, Pre- and Post-Program Inception

	ETP Schools			Within Tracking Schools	
	All	Non-Tracking	Tracking by initial achievement	Bottom Section	Top Section
Panel A. Baseline School Characteristics					
Total enrollment in 2004	666 (249)	678 (263)	654 (235)		
Number of government teachers in 2004*	11.9 (3.4)	12.0 (3.8)	11.8 (2.9)		
Number of PTA teachers in 2004	1.0 (1.4)	1.2 (1.5)	0.8 (1.2)		
School pupil/teacher ratio	53 (28)	54 (36)	52 (14)		
Performance at national exam in 2004 (out of 400)	259 (24)	258 (23)	260 (25)		
Number of classes without a classroom (classes held outside)	0.55 (1.63)	0.55 (1.68)	0.55 (1.60)		
Panel B. Class Size Prior to Program Inception (March 2005)					
Average class size in first grade	84 (34)	85 (35)	84 (32)		
Proportion of female first grade students	0.49	0.49	0.48		
Proportion of schools with only one class in the first grade	0.84	0.84	0.84		
Proportion of schools with a pre-primary (ECD) class	0.92	0.89	0.96		
Average class size in second grade	84 (32)	84 (35)	85 (29)		
Panel C. Class Size 6 Months After Program Inception (October 2005)					
Average class size in first grade	46 (16)	47 (17)	45 (15)		
Range of class sizes in sample (first grade)	19-98	20-97	19-98		
Panel D. Class Size in Year 2 of Program (March 2006)					
Average class size in first grade	77 (30)	78 (32)	75 (28)		
Average class size in second grade	45 (15)	46 (15)	45 (15)		
Range of class sizes in sample (second grade)	18-95	18-93	21-95		
Panel E. Within Tracking Schools: Students Characteristics by Tracking Status					
Proportion Female				0.476	0.494
Average Age at Endline				8.76 (1.46)	9.22 (1.44)
Was in preschool in 2004				0.111	0.073
Was in grade 1 in 2004				0.011	0.024
Average Standardized Baseline Score (Mean 0, Std. Dev. 1 at school level)				-0.81	0.79
Average Standardized Endline Score (Mean 0, Std. Dev. 1 at school level)				-0.40	0.39
Number of Schools	**140**	**70**	**70**	**70**	

**PTA (Parents-Teachers Association) teachers are locally hired by school committees*

Table 2
Overall Effect of Tracking

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Dep. Var: Total Score (standardized)			Dep. Var: Math Score (standardized)			Dep. Var: Lit Score (standardized)		
Panel A: All									
Tracking School	0.134	0.165	0.174	0.127	0.160	0.137	0.112	0.136	0.169
	(0.077)*	(0.076)**	(0.092)*	(0.065)*	(0.064)**	(0.072)*	(0.08)	(0.08)	(0.11)
In Bottom Half of Initial Distribution			-0.018			0.046			-0.07
x Tracking School			(0.07)			(0.07)			(0.08)
Individual Controls	no	yes	yes	no	yes	yes	no	yes	yes
Observations	5841	5315	5315	5841	5315	5315	5842	5316	5316
Mean in Non-Tracking Schools	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Std Dev in Non-Tracking Schools	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Panel B: Boys Only									
Tracking School	0.11	0.141	0.15	0.089	0.115	0.083	0.106	0.134	0.177
	(0.08)	(0.077)*	(0.10)	(0.07)	(0.066)*	(0.08)	(0.08)	(0.09)	(0.12)
In Bottom Half of Initial Distribution			-0.018			0.064			-0.088
x Tracking School			(0.10)			(0.08)			(0.11)
Individual Controls	no	yes	yes	no	yes	yes	no	yes	yes
Observations	2985	2709	2709	2985	2709	2709	2985	2709	2709
Mean in Non-Tracking Schools	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Std Dev in Non-Tracking Schools	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Panel C: Girls Only									
Tracking School	0.147	0.188	0.195	0.160	0.206	0.191	0.108	0.137	0.160
	(0.087)*	(0.085)**	(0.100)*	(0.073)**	(0.072)***	(0.078)**	(0.09)	(0.09)	(0.12)
In Bottom Half of Initial Distribution			-0.015			0.03			-0.049
x Tracking School			(0.08)			(0.08)			(0.08)
Individual Controls	no	yes	yes	no	yes	yes	no	yes	yes
Observations	2876	2606	2606	2876	2606	2606	2877	2607	2607
Mean in Non-Tracking Schools	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Std Dev in Non-Tracking Schools	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Notes: The sample includes 60 tracking and 62 non-tracking schools. Robust standard errors clustered at the school level are presented in parentheses. Individual controls included: age, gender, and a cubic polynomial in initial attainment quantile.

Table 3
Peer Quality: Regression Discontinuity Approach (Tracking Schools Only)

	Dep. Var: Total Score (standardized)				Dep. Var: Math Score (standardized)				Dep. Var: Literacy Score (standardized)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	ALL	Pair around median	ALL	Pair around median	ALL	Pair around median	ALL	Pair around median	ALL	Pair around median	ALL	Pair around median
Panel A: Reduced Form												
In Bottom Half of Initial Distribution	0.011	0.026	0.006	-0.017	0.014	0.016	0.007	-0.008	0.006	0.03	0.004	-0.022
	(0.096)	(0.141)	(0.079)	(0.148)	(0.096)	(0.152)	(0.081)	(0.146)	(0.100)	(0.138)	(0.085)	(0.163)
Observations (Students)	2955	148	2955	148	2956	148	2956	148	2955	148	2955	148
Mean in Bottom Half	-0.16	0.27	-0.16	0.27	-0.12	0.29	-0.12	0.29	-0.16	0.20	-0.16	0.20
Std Dev in Bottom Half	0.88	0.93	0.88	0.93	0.97	0.85	0.97	0.85	0.81	0.98	0.81	0.98
School Fixed Effects	no	no	yes	yes	no	no	yes	yes	no	no	yes	yes
Panel B: IV												
Mean Total score of Peers	0.010	0.022	-0.004	0.036								
	(0.102)	(0.168)	(0.090)	(0.189)								
Mean Math score of Peers					0.002	0.081	-0.001	0.190				
					(0.117)	(0.194)	(0.098)	(0.179)				
Mean Literacy score of Peers									0.018	-0.038	-0.007	-0.125
									(0.127)	(0.223)	(0.115)	(0.285)
Observations (Students)	2956	121	2956	121	2956	121	2956	121	2956	121	2956	121
School Fixed Effects	no	no	yes	yes	no	no	yes	yes	no	no	yes	yes
Panel C: First Stage for IV	*Dep. Var: Average Total Score of Peers*				*Dep. Var: Average Math Score of Peers*				*Dep. Var: Average Literacy Score of Peers*			
In Bottom Half of Initial Distribution	-0.875	-0.769	-0.868	-0.880	-0.820	-0.730	-0.822	-0.852	-0.748	-0.648	-0.735	-0.728
	(0.048)***	(0.088)***	(0.020)***	(0.051)***	(0.040)***	(0.074)***	(0.020)***	(0.057)***	(0.051)***	(0.094)***	(0.022)***	(0.054)***
Observations (Students)	2956	121	2956	121	2956	121	2956	121	2956	121	2956	121
R-squared	0.44	0.43	0.82	0.85	0.50	0.49	0.81	0.81	0.33	0.33	0.72	0.78
School Fixed Effects	no	no	yes	yes	no	no	yes	yes	no	no	yes	yes

Notes: Sample restricted to 60 schools where students were tracked by initial attainmnent into two streams. (Nine schools are dropped from the analysis because they had more than two sections in grade 1 at baseline.) Students in the bottom half of the initial distribution were assigned to the "bottom stream" where the average peer quality was much lower than in the top stream (see Figure 1). All regressions include individual controls (age, gender, and a third order polynomial in initial percentile fully interacted with a dummy for in the bottom half of the initial distribution).

Regressions in columns (3) and (6) include 1 pair of student per school: The top student in the low stream and the bottom student in the high stream. The number of observations is greater than 120 due to ties in some schools. In Panel B, the mean score of class peers is instrumented by the dummy "In bottom half of initial distribution" and controls.

Table 4

Interactions between Tracking and Peer Quality: Regression Discontinuity Approach

All ETP schools (Tracking and Non-Tracking)

	Dep. Var: Total Score (standardized)				Dep. Var: Math Score (standardized)				Dep. Var: Literacy Score (standardized)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	ALL	Pair around 50th percentile	ALL	Pair around 50th percentile	ALL	Pair around 50th percentile	ALL	Pair around 50th percentile	ALL	Pair around 50th percentile	ALL	Pair around 50th percentile
Tracking School	0.194	0.422			0.142	0.330			0.201	0.419		
	(0.096)**	(0.141)***			(0.073)*	(0.138)**			(0.113)*	(0.161)**		
In Bottom Half of Initial Distribution	0.052	0.047	0.024	0.029	0.035	0.021	0.010	-0.031	0.055	0.060	0.029	0.076
	(0.071)	(0.094)	(0.064)	(0.125)	(0.069)	(0.118)	(0.065)	(0.134)	(0.076)	(0.091)	(0.068)	(0.130)
In Bottom Half of Initial Distribution	-0.034	-0.027	-0.015	-0.042	0.043	-0.012	0.058	0.021	-0.093	-0.036	-0.075	-0.089
x Tracking School	(0.072)	(0.168)	(0.044)	(0.180)	(0.069)	(0.192)	(0.045)	(0.193)	(0.080)	(0.166)	(0.048)	(0.188)
Constant	-0.664	0.297	-0.563	0.440	-1.048	0.305	-0.942	0.495	-0.191	0.232	-0.116	0.304
	(0.149)***	(0.337)	(0.103)***	(0.403)	(0.138)***	(0.353)	(0.104)***	(0.432)	(0.159)	(0.330)	(0.110)	(0.420)
Observations	5314	306	5314	306	5315	306	5315	306	5315	306	5315	306
Mean in Bottom Half	-0.23	0.09	-0.23	0.09	-0.20	0.15	-0.20	0.15	-0.20	0.02	-0.20	0.02
Std Dev in Bottom Half	0.87	0.91	0.87	0.91	0.95	0.88	0.95	0.88	0.82	0.95	0.82	0.95
School Fixed Effects			x	x			x	x			x	x

Notes: Sample includes 62 non-tracking schools and 60 tracking Schools. (9 tracking schools and 7 non-tracking schools are dropped from the analysis because they had more than two sections in grade 1 at baseline.) In tracking schools, students in the bottom half of the initial distribution were assigned to the "bottom stream" where the average peer quality was much lower than in the top stream (see Figure 1). All regressions include individual controls (age, gender, and a third order polynomial in initial percentile).

Regressions in columns (3) and (6) include 1 pair of students per school. The number of observation is above 244 because of ties.

Table 5
Peer Quality: Exogenous Variation in Peer Quality
(Non-Tracking Schools Only)

	Dep Var: Total Score					
	(1)	(2)	(4)	(5)	(6)	(7)
Average (Standardized) Baseline Score of Classmates	0.24		0.23	0.17		0.16
	(0.17)		(0.17)	(0.16)		(0.16)
Std. Dev. in Ave. Baseline Score of Classmates		-0.12	-0.11		-0.09	-0.08
		(0.21)	(0.21)		(0.20)	(0.20)
Own Baseline Score (Standardized)	0.53	0.52	0.53	0.53	0.53	0.53
	(0.02)***	(0.02)***	(0.02)***	(0.02)***	(0.02)***	(0.02)***
Girl	0.05	0.05	0.05	0.04	0.04	0.04
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Age	-0.03	-0.03	-0.03	-0.03	-0.04	-0.03
	(0.01)**	(0.01)**	(0.01)**	(0.01)**	(0.01)***	(0.01)**
Taught by ETP Teacher	0.15	0.15	0.15	0.14	0.14	0.14
	(0.04)***	(0.04)***	(0.04)***	(0.04)***	(0.04)***	(0.04)***
Constant	0.28	0.42	0.39	0.33	0.47	0.41
	(0.13)**	(0.25)*	(0.25)	(0.13)***	(0.23)**	(0.24)*
Observations	2225	2225	2225	2225	2225	2225
School Fixed Effects				x	x	x

Notes: Sample restricted to schools where students were randomly assigned to a section. Robust standard errors clustered at the section level in parentheses.

Table 6
Peer Quality: Exogenous Variation in Peer Composition
Non-Tracking Schools Only

	Dep. Var: Total Score (standardized)				Dep. Var: Math Score (standardized)				Dep. Var: Literacy Score (standardized)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	ALL	Bottom Tier	Middle Tier	Top Tier	ALL	Bottom Tier	Middle Tier	Top Tier	ALL	Bottom Tier	Middle Tier	Top Tier
Share of Classmates in Bottom Tier†	-0.95	-0.50	-0.41	-1.75	-0.87	-0.88	-0.20	-1.40	-0.83	-0.07	-0.53	-1.72
	(0.27)***	(0.50)	(0.40)	(0.49)***	(0.32)***	(0.42)**	(0.38)	(0.50)***	(0.28)***	(0.58)	(0.44)	(0.50)***
Share of Classmates in Top Tier†	-0.92	-0.40	-0.80	-1.73	-0.85	-0.67	-0.75	-1.67	-0.80	-0.08	-0.72	-1.44
	(0.31)***	(0.56)	(0.49)	(0.70)**	(0.36)**	(0.48)	(0.55)	(0.63)***	(0.33)**	(0.65)	(0.46)	(0.78)*
Own Baseline Score (Standardized)	0.52	0.50	0.55	0.47	0.51	0.55	0.52	0.32	0.43	0.36	0.47	0.52
	(0.03)***	(0.09)***	(0.09)***	(0.07)***	(0.02)***	(0.09)***	(0.09)***	(0.05)***	(0.03)***	(0.09)***	(0.09)***	(0.08)***
Girl	0.04	0.04	0.10	-0.01	-0.03	0.02	-0.02	-0.08	0.09	0.05	0.19	0.06
	(0.04)	(0.06)	(0.06)*	(0.07)	(0.04)	(0.07)	(0.06)	(0.06)	(0.04)**	(0.06)	(0.06)***	(0.09)
Age	-0.04	0.01	-0.05	-0.07	0.00	0.01	0.00	-0.02	-0.06	0.01	-0.07	-0.11
	(0.02)*	(0.03)	(0.02)**	(0.03)**	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)***	(0.03)	(0.02)***	(0.03)***
Taught by ETP Teacher	0.11	0.05	0.13	0.18	0.06	-0.02	0.10	0.08	0.14	0.09	0.13	0.24
	(0.04)***	(0.05)	(0.05)**	(0.06)***	(0.04)*	(0.05)	(0.05)*	(0.06)	(0.03)***	(0.05)*	(0.05)**	(0.06)***
Constant	0.98	0.31	0.76	1.95	0.64	0.57	0.40	1.52	1.09	0.02	0.95	1.94
	(0.27)***	(0.50)	(0.33)**	(0.46)***	(0.27)**	(0.40)	(0.36)	(0.44)***	(0.29)***	(0.55)	(0.33)***	(0.48)***
Observations	2223	697	775	751	2223	697	775	751	2224	697	776	751
School Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Ave. Share of Classmates in Bottom Tier	0.33	0.33	0.32	0.32	0.33	0.33	0.32	0.32	0.33	0.33	0.32	0.32
Ave. Share of Classmates in Top Tier	0.33	0.32	0.33	0.33	0.33	0.32	0.33	0.33	0.33	0.32	0.33	0.33

Data from 49 non-tracking schools.

† *Tiers based on the initial test score distribution.*

Table 7
Teacher Effort: Presence

	All Teachers			Government Teachers			ETP Teachers			Students	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Dep Var: Teacher Found in school on random school day	*Dep Var:* If in school, found in class teaching	Teacher found in class teaching (unconditional on presence)	*Dep Var:* Teacher Found in school on random school day	*Dep Var:* If in school, found in class teaching	Teacher found in class teaching (unconditional on presence)	*Dep Var:* Teacher Found in school on random school day	*Dep Var:* If in school, found in class teaching	Teacher found in class teaching (unconditional on presence)	*Dep Var:* Student found in school on random school day	
Tracking School	0.041	0.079	0.096	0.054	0.094	0.112	-0.009	0.015	0.007	-0.012	-0.012
	(0.021)**	(0.039)**	(0.038)**	(0.025)**	(0.047)**	(0.044)**	(0.034)	(0.036)	(0.045)	(0.013)	(0.013)
Bottom Half x Tracking School	-0.049	-0.034	-0.062	-0.073	-0.036	-0.076	0.036	-0.034	-0.004	-0.014	0.010
	(0.029)*	(0.042)	(0.040)	(0.034)**	(0.059)	(0.053)	(0.046)	(0.050)	(0.057)	(0.005)***	(0.007)
Years of Experience Teaching	0.000	-0.006	-0.005	0.002	0.001	0.002	-0.002	-0.007	-0.008		
	(0.001)	(0.001)***	(0.001)***	(0.001)*	(0.002)	(0.001)	(0.003)	(0.007)	(0.008)		
Female	-0.023	0.033	0.012	-0.004	0.121	0.101	-0.034	-0.034	-0.061		-0.004
	(0.018)	(0.027)	(0.026)	(0.020)	(0.035)***	(0.031)***	(0.032)	(0.037)	(0.043)		(0.004)
Assigned to ETP Teacher										0.013	0.013
										(0.004)***	(0.004)***
Constant	8.903	51.006	46.459	11.544	45.866	42.615	-15.785	45.127	26.992	0.895	0.877
	(8.353)	(15.368)***	(13.329)***	(9.611)	(17.047)***	(14.294)***	(17.723)	(21.770)**	(23.863)	(0.009)***	(0.022)***
Observations	2098	1782	2098	1633	1367	1633	465	415	465	44225	44050
Mean in Non-Tracking Schools	0.837	0.609	0.510	0.825	0.545	0.450	0.888	0.842	0.748	0.865	0.865
F test	2.718	7.693	9.408	2.079	4.414	5.470	2.426	2.570	3.674	8.901	8.235
Prob >F	0.011	0.000	0.000	0.050	0.000	0.000	0.023	0.016	0.001	0.000	0.000

Notes: Linear probability model regressions. Multiple observations per teacher and per student. Standard errors clustered at school level. Region and date of test dummies were included in all regressions but are not shown.

Table 8
Effect of Tracking by Level of Complexity and Initial Attainment

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Mathematics			Test	Literacy			
	Difficulty Level 1	Difficulty Level 2	Difficulty Level 3	Coeff (Col 3) = Coeff (Col 1)	Reading letters	Spelling Words	Reading Words	Reading Sentences
(1) In Bottom Half of Initial Distribution	-1.43	-1.21	-0.50		-3.87	-4.09	-4.16	-1.14
	(0.10)***	(0.07)***	(0.05)***		(0.32)***	(0.42)***	(0.39)***	(0.21)***
(2) Tracking School	0.15	0.17	0.21		1.59	0.95	1.03	0.39
	(0.09)	(0.12)	(0.10)**		(0.65)**	(0.78)	(0.74)	(0.34)
(3) In Bottom Half of Initial Distribution	0.18	0.08	-0.09	$X^2 = 3.69$	-0.43	-0.57	-0.39	-0.44
x Tracking School	(0.14)	(0.12)	(0.08)	(Prob > X^2) = 0.055	(0.46)	(0.61)	(0.55)	(0.29)
Constant	4.95	1.84	0.58		11.70	10.14	10.26	3.92
	(0.22)***	(0.22)***	(0.16)***		(0.99)***	(1.19)***	(1.12)***	(0.56)***
Observations	5316	5316	5316		5315	5311	5316	5316
Maxiumum possible score	6	6	6		24	24	24	24
Mean in Non-Tracking Schools	4.15	1.60	0.67		7.01	5.54	5.03	2.52
Std Dev in Non-Tracking Schools	2.02	1.62	0.94		6.55	7.61	7.30	3.92
Total effect of tracking on bottom half:								
Coeff (Row 2)+Coeff (Row 3)	0.33	0.25	0.12	$X^2 = 2.33$ (Prob > X^2) = 0.127	1.16	0.38	0.64	-0.05
F Test: Coeff (Row 2)+Coeff (Row 3) = 0	3.69	6.84	4.74		4.42	0.68	1.60	0.09
Prob > F	0.06	0.01	0.03		0.04	0.41	0.21	0.77

Notes: The sample includes 60 Tracked Schools and 62 Untracked Schools. Robust standard errors clustered at the school level are presented in parentheses.

Difficulty level 1: addition or substration of 1 digit numbers

Difficulty level 2: addition or substration of 2 digit numbers, and multiplication of 1 digit numbers

Difficulty level 3: addition or substration of 3 digit numbers

Appendix Table 1
Does Attrition Vary Across Tracking and Non-Tracking Schools?

	(1)	(2)	(3)
	Transferred to other school	If not transferred: missed test	Total Attrition
Tracking School	0.02	0.01	0.03
	(0.02)	(0.03)	(0.04)
(decile=2) x Tracking School	-0.02	-0.01	-0.02
	(0.03)	(0.04)	(0.05)
(decile=3) x Tracking School	-0.04	-0.04	-0.06
	(0.03)	(0.04)	(0.05)
(decile=4) x Tracking School	0.00	-0.01	0.00
	(0.03)	(0.04)	(0.04)
(decile=5) x Tracking School	-0.02	-0.01	-0.03
	(0.03)	(0.04)	(0.05)
(decile=6) x Tracking School	0.00	-0.03	-0.02
	(0.03)	(0.04)	(0.04)
(decile=7) x Tracking School	0.01	-0.03	-0.01
	(0.03)	(0.04)	(0.04)
(decile=8) x Tracking School	-0.01	-0.01	-0.03
	(0.03)	(0.04)	(0.04)
(decile=9) x Tracking School	0.00	-0.02	-0.02
	(0.03)	(0.04)	(0.05)
(decile=10) x Tracking School	-0.02	-0.01	-0.02
	(0.03)	(0.04)	(0.04)
Bungoma District	-0.01	-0.01	-0.01
	(0.01)	(0.01)	(0.01)
Constant	0.07	0.18	0.22
	(0.01)***	(0.03)***	(0.03)***
Observations	7381	6855	7381
Mean	0.08	0.13	0.18
Std Dev.	0.28	0.33	0.38
F test	0.81	0.23	0.93
Prob >F	0.45	0.80	0.40

Notes: OLS Regressions; standard errors clustered at school level. Decile dummies included but not shown.