NBER WORKING PAPER SERIES

PEER EFFECTS AND THE IMPACT OF TRACKING: EVIDENCE FROM A RANDOMIZED EVALUATION IN KENYA

Esther Duflo
Pascaline Dupas
Michael Kremer

Working Paper 14475
http://www.nber.org/papers/w14475

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2008

We thank Josh Angrist, Abhijit Banerjee, Michael Greenstone, Caroline Hoxby, Guido Imbens, Brian Jacob, and many seminar participants for helpful comments and discussions. 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, Nicolas Studer, and especially Willa Friedman for excellent research assistance. We are grateful to Grace Makana and her field team for collecting all the data. We thank, without implicating, the World Bank and the Government of the Netherlands for the grant that made this study possible. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

Peer Effects and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya
Esther Duflo, Pascaline Dupas, and Michael Kremer
NBER Working Paper No. 14475
November 2008
JEL No. I20,O1

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 test these competing claims using a randomized evaluation in Kenya. One hundred and twenty one primary schools which all had a single grade one class received funds to hire an extra teacher to split that class into two sections. In 60 randomly selected schools, students were randomly assigned to sections. In the remaining 61 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. After 18 months, students in tracking schools scored 0.14 standard deviations higher than students in non-tracking schools, and this effect persisted one year after the program ended. Furthermore, students at all levels of the distribution benefited from tracking. A regression discontinuity analysis shows that in tracking schools scores of students near the median of the pre-test distribution score are independent of whether they were assigned to the top or bottom section. In contrast, in non-tracking schools we find that on average, students benefit from having academically stronger peers. This suggests that tracking was beneficial because it helped teachers focus their teaching to a level appropriate to most students in the class.

Esther Duflo
Department of Economics
MIT, E52-252G
50 Memorial Drive
Cambridge, MA 02142
and NBER
eduflo@mit.edu

Michael Kremer
Harvard University
Department of Economics
Littauer Center M20
Cambridge, MA 02138
and NBER
mkremer@fas.harvard.edu

Pascaline Dupas
Department of Economics
UCLA
8283 Bunche Hall
Los Angeles, CA 90095
and NBER
pdupas@econ.ucla.edu

1. Introduction

Tracking students by prior achievement is controversial among academics and policymakers. On one hand, if teachers find it easier to teach a homogeneous 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 low achieving students while benefiting high achieving students, thereby exacerbating inequality.

Direct evidence on the effect of tracking on the achievement of students of various ability levels is mixed. Betts and Shkolnik (1999) review the 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 invalid comparisons. Most of the papers they review compare 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 tracking and non-tracking high schools, Betts and Shkolnik (1999) conclude that low ability students are neither hurt nor helped by tracking; top students are helped; and there is some evidence that middle scoring 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 may attract a different pool of students and possibly a different pool of teachers. Manning and Pischke (2006) show that controlling for baseline scores is not sufficient to eliminate the selection bias when comparing students attending comprehensive versus selective schools in the United Kingdom. The ideal experiment to measure the impact of tracking on test scores at various levels of the distribution would

be to randomly assign students to tracking or non-tracking schools, and compare the performance of students of different prior achievement across school types.

We shed light on these issues using experimental data from Kenya. In 2005, 140 primary schools in western Kenya received funds from the non-governmental organization ICS Africa to hire an extra teacher. One hundred and twenty one (121) of these schools had a single first-grade class, and split their first-grade class into two sections. In 60 schools (randomly selected out of the 121), students were assigned to sections based on prior achievement as measured by first term grades (the program started in the second term of the school year). In the remaining 61 schools, students were randomly assigned to one of the two sections. 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, students in all schools took a comprehensive achievement test and the program ended. One year later, a long-term follow-up test was administered.

The results suggest that in the Kenyan context all students benefited from tracking. On average, test scores were 0.14 standard deviations higher in tracking schools than in non-tracking schools (0.18 standard deviations after controlling for baseline scores and other control variables). After controlling for the baseline score, students in the top half of the pre-assignment distribution gained 0.19 standard deviations, and those in the bottom half of the pre-assignment distribution gained 0.16 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. Furthermore, these results are persistent: students in tracking schools scored 0.16 standard deviations higher (0.18 standard deviations with control variables) one year after tracking ended.

Strikingly, test scores of students in the middle of the pre-assignment distribution did not seem affected by which section they were assigned to. Specifically, in tracking schools, using a regression discontinuity design strategy, we cannot reject the hypothesis that there is no difference in end line achievement between the lowest scoring student assigned to the high achievement section and the highest scoring student assigned to the low achievement section, despite large differences in pre-assignment test scores between the two groups.

We complete this analysis by exploiting the random assignment of students to sections in the non-tracking schools to estimate the impact of peer achievement in the absence of tracking. We find positive impacts of average peer achievement in this context, and no effect of the variance of test scores.

A natural interpretation of these contrasting results is that while there may be a direct effect of peer scores on performance, students also affect their group indirectly, through their impact on teacher behavior. Small variations in test scores do not allow teachers to adjust their teaching, but tracking provides them with a much more homogeneous group of students, allowing them to change their teaching practices. This implies that estimating peer effects through small variation in peer achievement would not be a good guide to what would happen to low and high achieving students in a tracking system. We present some corroborative evidence that teacher behavior was affected by tracking: First, teachers are more likely to be in class and teaching in tracking schools, although this effect seems to be entirely due to greater presence of teachers assigned to the top section. Second, students in the bottom half of the initial distribution, seem to gain comparatively more from tracking in the most basic competencies, while students in the top half of the initial distribution seem to gain more from tracking in somewhat more advanced

competencies, consistent with the hypothesis that teachers are tailoring instruction to class composition.

This paper is related to a large literature that investigates peer effects in the classroom (e.g., Hoxby, 2000; Zimmerman, 2003; Angrist and Lang, 2004). While, mainly for data reasons, this literature has mostly focused on "linear in means" specification, there are a few exceptions with results generally consistent with ours. 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 homogeneous peers. Lavy, Paserman and Schlosser (2008) find that the fraction of repeaters in a class has a negative effect on the scores of the other students, in part due to deterioration of the teacher's pedagogical practices. 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, consistent with our regression discontinuity design results. Finally, Lyle (2007) and Ding and Lehrer (2007) both find that variance of peer test scores matters more than their average level among West Point cadets and Chinese secondary school students respectively.

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 Kenya education system includes 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. Largely due to the elimination of school fees in 2003, primary school enrollment rose nearly 30 percent, from 5.9 million to 7.6 million between 2002 and 2005 (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, pupil-teacher ratios have grown dramatically, particularly in lower grades. In our sample of schools in western Kenya, for example, the median first grade class in 2005 (two years after the introduction of free primary education, and prior to the class size reduction program we exploit here) had 74 students; average class size was 83; 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 preschools (which are neither free nor compulsory in Kenya). Students differ vastly in age, school preparedness, and support at home. These challenges are not unique to Kenya. They confront many developing countries where school enrollment has risen sharply in recent years: understanding the roles of tracking and peer effects in this context is thus particularly important.

2.2. Study Design

This study took advantage of a class-size reduction program 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 the Extra-Teacher Program (ETP). Under ETP, with funding from the World Bank, 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.[5] The program provided funds for schools to create one additional section in first grade, taught by the contractual teacher. Most schools (121) had only one first grade section, and split it into two sections. Schools that already had two or more first grade sections added one section. The average section 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, which included 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).

To learn about the impacts of tracking and peer effects, we overlaid the following intervention on the class size reduction program.[6] In 60 schools randomly selected out of the 121 schools that had originally only one grade 1 class and that received an extra teacher, grade 1 pupils were randomly assigned (by the research team) to one of the two sections. We call these schools the "non-tracking schools." In the remaining 61 schools (the "tracking schools"), the children were divided into sections by achievement level, according to their scores on exams administered by the school during the first term of the 2005 school year. In the tracking schools, the 50 percent of the class with the lowest exam scores were assigned to a section (we call this section the "bottom section") and the rest of the class was assigned to the other section (the "top section"). The 19 schools that originally had two or more grade one classes were also randomly divided into tracking and non-tracking schools, but it proved difficult to organize the tracking consistently in these schools.[7] Thus, in the analysis that follows, we focus on the 121 schools that

[5] 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.
[6] In a companion paper, we study the impact of class size and contractual arrangements.
[7] In these schools, the sections that were taught by civil service teachers rather than contract teachers sometimes recombined or exchanged students.

initially had a single grade 1 class and exclude 19 schools (10 tracking, 9 non-tracking) that initially had two or more.[8]

In principle, exceptions to the initial assignment could be granted for siblings, or if parents objected. However, there were few cases in which students or schools made transfers. The analysis below is based on the initial assignment regardless of which section the student eventually joined.

After students were assigned to sections, the contract teacher and the civil-service teacher were then 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.[9]

Table 1 presents summary statistics for the 121 schools in our sample[10]. 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 differ from school to school, they are normalized such that the mean is 0 and the standard deviation is 1 in each school.) The average baseline score was -0.81 in the bottom section, and 0.80 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 the student's own baseline score in tracking and non-tracking schools. Average peer test score is not affected by the student's own test score in non-tracking schools but, consistent with the discontinuous assignment at the 50th percentile for most

[8] Note that the randomization of the schools into the tracking and non tracking was stratified according to whether the school originally had one or more grade 1 classes. We obtain similar results when using the entire sample of schools, however.

[9] Students enrolled in grade 2 in 2005 and who repeated grade 2 in 2006 were randomly assigned to either the contract teacher or the civil-service teacher in 2006. They are excluded from the study. Students who repeated grade 1 in 2006 remain in the data set, and the analysis is based on the initial assignment.

[10] 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) at six per school on average.

schools, there is sharp discontinuity at the 50th percentile in tracking schools.[11] Note that while the baseline exams were administered by each school, and thus are not a comparable competency exam across schools, they nevertheless seem to be a good measure of academic achievement. They are strongly correlated with the end line test we administered: the correlation is 0.47. The average end line score in the bottom section in tracking school was -0.44 standard deviations, and the average end line score in the top section was 0.42 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 consists of approximately 10,000 students enrolled in first grade in March 2005 in one of 121 primary schools enrolled in the study. Slightly less than half are girls (49 percent).

The key outcome of interest is student academic achievement, as measured by their scores on a standardized math and language test first administered in all schools 18 months after the start of the program. Trained enumerators administered the test, which was then 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 were randomly sampled (using a random number generator, and before reaching the school) after stratifying by their position in the initial distribution.

[11] Peer quality is slightly more similar for children on the left and the right of the 50th percentile than for students at other percentiles because the assignment procedure used a manually computed ranking variable that was very strongly correlated with the ranking based on the actual school grades but had a few discrepancies (due to clerical errors). Thus, some children close to the median who should have been assigned to one section wound up in the other one. We are using the rank based on the actual school grade as our control variable in what follows, in case the ranking variable that was used for assignment was in fact manipulated.

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 answered math and literacy questions ranging from identifying letters and counting to subtracting three-digit numbers and reading and understanding sentences.

To limit attrition, enumerators were instructed to go to the homes of sampled students who had dropped out or were absent on the day of the test, and to bring them to school for the test. It was not always possible to find the child, however, and the resulting attrition rate on the test was 18 percent. However, there was no difference between tracking and non-tracking schools in overall attrition rates. The characteristics of those who attrited are similar across groups, except that girls in tracking schools were less likely to attrit in the endline test (see appendix table 1). 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 end line test score data for 5,796 students.

To measure whether the effects of the program persisted, the children that had been sampled for the end line were tested again following the same procedures in November 2007, one year after the program ended. During the 2007 school year, students were overwhelmingly enrolled in grades for which their school had a single section, so tracking was no longer an option. Most of these students had reached grade 3 at the time, but repeaters were also tested. The attrition for this long-term follow-up was 22 percent only 4 points higher than attrition at the endline test. Neither the proportion nor the characteristics of children who could not be tested differed across treatment arms (appendix table 1).

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 percent). Finally, each

school received unannounced 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 grades awarded by the school (and therefore the rank of the child in the class) before the program, and we show below that this initial ranking is strongly predictive of the end line test scores, which indicate that it contains some valuable information.

2.3 Empirical Strategy

2.3.1 Conceptual Framework

The following framework may be useful in thinking about the impact of the program. Suppose that educational outcomes for student i in school j, y_{ij}, are given by:

$$y_{ij} = f(x_{ij}, \overline{x}_j, e_{ij}, u_{ij})$$

where x_{ij} is the student's pretest score, $\overline{x}_j$ is the average scores of other student in the class, e_{ij} is an index of teacher effort and attention directed to student i, and u_{ij} represents other student and class-specific factors, possibly stochastic (such as student effort, class size, ability, etc.). In this equation, $\overline{x}_j$ reflects the direct effect of a student's peers on his learning, e.g. through peer-to-peer interactions.

Teacher effort and attention directed to student i depends on the overall level of effort expended by the teacher (e_j) as well as the extent to which student i benefits from this effort. The share of teacher attention directly relevant to student i presumably depends on the position of student i in the distribution of student achievement. For

example, if teachers and schools orient their teaching to the strongest students in the class, either because they prefer teaching these students or because they have incentives based on performance of students who stay in school and complete the primary-school leaving exam (see Glewwe, Kremer and Moulin, 2007), then weak students will receive a lower share of teacher effort if they are in a class with strong students than if they are in a class with other weak students Alternatively, if teachers adapt pedagogy to the median student, then students will get more attention if they are in the middle of the distribution of their class. Moreover, e_{ij} is likely to also depend on the heterogeneity of achievement levels of students in the classroom. Teachers with very heterogeneous classes, may, for example, divide the students into groups, and divide attention among the different groups (because teaching only material appropriate to the median students would not help the strong and weak students). When heterogeneity is sufficiently reduced, the teacher may be able to separate the class into fewer groups, and therefore devote more time to material that is relevant to each student. This suggests that a reduction in heterogeneity (measured for example by variance in test scores) which is sufficiently large and sufficiently salient to change teacher time allocation would help all students.

To sum up, suppose that teacher's attention received by student i is given by:

$$e_{ij} = g(\overline{x_j}, var(x_j), r_{ij}, \epsilon_{ij})$$

where r_{ij} is the rank of student i in the class and ϵ_{ij} are unmodeled factors (incentives, pupil teacher ratio, teacher motivation) which affect the teacher's overall effort and relationship with student i.

In this framework, to understand the impact of changing peers on educational outcomes, it is important to look at the total derivative of *f* with respect to average quality of peers and their heterogeneity. In particular, the framework suggests that small and large changes in peer quality or in the variance of peers may have very different impacts:

if teacher behavior changes only when there is a large enough or salient enough change in peer quality or homogeneity, small changes in the composition of the peer group will not affect e_{ij}, and will help us identify the partial derivative of f with respect to $\overline{x_j}$. However, going from no tracking to tracking may affect $\overline{x_j}$, $var(x_j)$ and r_{ij} sufficiently to change e_{ij}, and the effect of $\overline{x_j}$ and $var(x_j)$ identified by a tracking experiment could therefore potentially be quite different from that identified from small changes.

In particular, small random decreases in $\overline{x_j}$ due to random peer assignment may decrease test scores, but large reductions in $\overline{x_j}$ accompanied by large reductions in $var(x_j)$, such as those induced by tracking for students initially in the lower half of the achievement distribution may increase test scores (both by reducing heterogeneity and by leading teachers to teach material better suited toward these students). Thus, tracking could benefit both stronger and weaker students if the positive impact of homogeneity dominates.

The model suggests that tracking could be beneficial to students (both strong and weak), but it has a number of additional predictions which we also explore below: (1) the impact of a small change in peer composition, which identifies the direct effect of the peer group, may be very different from the impact of a large and salient change in peer composition (even keeping the variance in the peer group constant); (2) teacher effort may be different in tracking and non tracking environment, and this could partly mediate the impact of tracking; (3) if the gains from tracking are due to a change in the content of what is being taught in class, this may imply that initially high and low achieving students in tracking schools may make progress relative to their counterparts in non-tracking schools on different types of skills.

2.3.2 Measuring the Impact of Tracking

Given our experimental setup, measuring the overall impact of tracking on test scores is straight-forward. We run regressions of the form:

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

where y_{2ij} is the end line test score of student i in school j (expressed in standard deviations of the distribution of scores in the non-tracking schools),[12] 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 that controls for 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 civil-service 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 that indicates whether the child was in the bottom half the baseline score distribution in her school (B_{ij} is also included in the vector X_{ij}). We also estimate a specification where we interact the treatment with the initial quartile of the child in the baseline distribution. 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 end line 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

[12] We have also experimented with an alternative specification of the end line test score for math, which uses item response theory to give different weights to questions of different levels of difficulty (the format of the language score was not appropriate for this exercise). The results were extremely similar (results available from the authors) so we focus on the standardized test scores in this version.

of a similar form, using as outcome variable teacher presence in school, whether the teacher is in class teaching, and student presence in school.

2.3.3 Peer effects

This setup provides two separate opportunities to identify peer effects in the classroom:

a) Regression Discontinuity Design

Tracking schools provide a natural setup for a regression discontinuity (RD) design estimate of the impact of average peer scores. As shown in figure 1, the two students close to the median were assigned to classes where the average prior achievement 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 students, the median student was randomly assigned to one of the sections).

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:

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

To test the robustness of our estimates to various specifications of the control function, we also run similar specifications as equations (3) and (4), allowing the polynomial to be estimated separately for each side of the discontinuity, and we report the difference in test scores on both side of the discontinuity. Finally, we also follow

Imbens and Lemieux (2007) and use a Fan locally weighted regression of the relationship between end line test scores and baseline percentile on both sides of the discontinuity.[13]

Note that this is an unusually favorable setup for an RD design. There are 60 discontinuities in the data set, 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).[14] We can thus do what the RD framework suggests should be done asymptotically (but cannot happen in practice in finite data sets), 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).

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

These reduced form results are of independent interest, and 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 median 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 end line test scores of the classmates of student i in school j. The structural equation:

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

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 child in a tracking environment, where having high achieving peers on average

[13] The results turned out to be insensitive to the bandwidth, so we did not implement the cross-validation proposed by Imbens and Lemieux (2007). We report the result with a bandwidth of 2.

[14] Black, Galdo and Smith (2007) also exploit a series of sharp discontinuities in their estimation of a re-employment program across various sites in Kentucky.

also means that the child is the lowest achieving child of his track (at least at baseline) and having low-achieving peers means that the child is the highest achieving child of his track. It is useful to go back to the conceptual framework to help us think through the interpretation of the findings. The rank of students in the class may affect how much teacher attention they get. The strongest student in a section may receive much more teacher attention than the weakest student.. Thus, even if the direct effect of peers' average effort on student achievement is positive, the median children assigned to the low class may perform better, if the rank effect dominates. Another characteristic of this setup is that the variation in peers' scores around the threshold is very large (1.6 standard deviation for the baseline score, and 0.73 standard deviations for the end line test score), and explicitly acknowledged by schools and teachers: thus, the estimate captures the impact of the teacher response to the change in peer composition. Thus, if the teachers teach less advanced material in the bottom track, and if the median students benefits more from being taught less advanced material than more advanced material, any direct effect of peers may be more than compensated by the change in the material taught by the teacher.

b) Random Assignment to Peers

To estimate the impact of smaller variations in peer achievement in a non-tracking environment, we 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.

In the sample of non-tracking schools, we start by estimating the effect of a student's peer average baseline test scores by OLS (this is the average of the section minus the student's own score):

(7) $$y_{2ij} = \kappa A(y_{1ij}) + X_{ij}\beta + \varepsilon_{ij,} + \upsilon_j + \varepsilon_{ij},$$

where $A(y_{1ij})$ is the average baseline test score of the peers in the section a student was assigned.[15] The vector of control variables X_{ij} includes a measure of the student's own baseline score. Since students were randomly assigned within schools, our estimate of the coefficient of $A(y_{1ij})$ in a specification including school fixed effects will reflect the causal effect of peers' prior achievement.

To capture heterogeneity, we also run regressions with the variance of the peers' test scores, noted $V(y_{2ij,})$, as the independent variable (as well as a specification including both):

(8) $$y_{2ij} = \lambda V(y_{1ij}) + X_{ij}\beta + \varepsilon_{ij,} + \upsilon_j + \varepsilon_{ij},$$

The baseline grades are not comparable across schools (they are the grades assigned by the teachers in each school). Therefore, the magnitude of the results in equation (7) is a bit difficult to interpret. However, baseline grades are strongly correlated with end line test scores, which are comparable across schools. Thus, to provide results that can be compared with the literature and with the regression discontinuity result we present for the tracking schools, we estimate the impact of average end line peer test scores on a child's test score:

(9) $$y_{2ij} = \kappa A(y_{2ij}) + \lambda_1 P_{ij} + X_{ij}\beta + \varepsilon_{ij,}$$

This equation is estimated by instrumental variables, using $A(y_{1ij})$ as an instrument for $A(y_{2ij})$.[16]

[15] There were very few re-assignments, but we always focus on the initial random assignment: that is, we consider the test scores of the other students *initially assigned* to the class to which a student was *initially assigned* (regardless of whether they eventually attended that class).

[16] This is similar to equation (6), except that there is no polynomial in test scores.

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 at end line (panel A shows effects after 18 months of continuous tracking) as well as in the long run (panel B shows effects a year after the end of the program). We find that tracking by initial achievement significantly increased test scores. At the end line test, students in tracking schools scored 0.138 standard deviations (with a standard error of 0.078 standard deviations) more than students in non-tracking schools overall (table 2, column 1, panel A). The effect becomes somewhat larger (0.175 standard deviations, with a standard error of 0.077 standard deviations) when controlling for individual-level covariates (column 2).

The program effect persists beyond the duration of the program. When the program ended after 18 months, most students had then reached class 3, and in all schools except five, there was only one class for grade 3. The remaining students had repeated and were in grade 2 where, once again, most schools had only one section (since after the end of the program they did not have funds for additional teachers). Thus, after the program ended, students in our sample were not tracked any more (and they were in larger classes than both tracked and non-tracked students had experienced in grade 1 and 2). Yet, one year later, test scores of students in tracking schools were still 0.163 standard deviations greater (with a standard error of 0.069 standard deviations) than those of students in non-tracking schools overall (table 2, column 1, panel B). The effect are slightly larger (0.178 standard deviation) and more significant with control variables (column 2, panel B). Thus, this program has very persistent long term effects. This is striking, since in many evaluations, the test score effects of even successful interventions tend to fade over time (e.g., Banerjee, et al. (2007), Andrabi, et al. (2008)). This indicates that tracking may

have helped students master core competencies in grades 1 and 2 and that this may have helped them learn more later on.[17]

Table 2 also allows us to see whether the effect of tracking by initial achievement differs between initially high-scoring students (who are grouped with other strong students in tracking schools) and initially low-scoring students (who are grouped with other low-scoring students). We find that both sets of students benefited from tracking (in row 2, column 3, panel A, the interaction between being in the bottom half and in a tracking school cannot be distinguished from zero, and the total effect for the bottom half is 0.155 standard deviations, with a p value of 0.04). A year later, the effect for the bottom half is still large (0.135, with a p value of 0.09). When we look at each quartile of the initial distribution separately, we find positive point estimates for all quartiles, although the effect for students in the bottom quartile is insignificant a year after the end of the program (column 4).

There is no significant difference in the impact of the program between math and language scores, though the effects are more precisely estimated for math than for language (columns 5 through 8 and columns 9 through 12).

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 end line test score as a function of the baseline test score using a second-order polynomial estimated separately on either side of the cutoff in both the tracking and non-tracking schools. This figure shows that, both in language and in math, tracking seems to

[17] We also find (in results not reported here to save space) that initially low achieving students in tracking school are 4 percentage points less likely to repeat grade 1. Since the program continued in grade 2, students who repeated lost the advantage of being in a small class, and being more likely to be taught by a contract teacher. Part of the effect of tracking after the end of grade 1 may be due to this. In the companion paper, we estimate the effect of the class size reduction program in non tracking school to be 0.16 standard deviations. At the most, the repetition effect would therefore explain an increase in 0.04*0.16=0.0064 standard deviation in test scores. Furthermore, it is present only for girls, while the tracking effect is there both for boys and girls.

increase test scores 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 25^{th} or 75^{th} percentile, who became the median students in the tracking schools.

Table 3 tests for heterogeneity in the effect of tracking. We present the estimated effect of tracking separately for boys and girls in panel A, and for students taught by civil-service teachers and contract teachers in panel B. This distinction is important, since the impact of tracking could be affected by teacher response, and contract and civil-service teachers have different experience and incentives. The two treatments were stratified. Finally we look at the interaction between tracking and class size in panel C. In each case, we separately present the effect for students that started in the bottom half of the distribution and those that started in the top half.

Although the coefficients are not significantly different from each other, point estimates suggest that the effects are larger for girls in math (panel A). For both boys and girls, initially weaker students benefit as much as initially stronger students.

While tracking increases test scores for students at all levels of the pre-test distribution assigned to be taught by contract teachers (indeed, initially low-scoring students assigned to a contract teachers benefited even more from tracking than initially high-scoring students), initially low-scoring students did not benefit from tracking if assigned to a civil-service teacher. In contrast, tracking substantially increased scores for initially high-scoring students assigned to a civil-service teacher. Below, we will present evidence that this may be because tracking led civil-service teachers to increase effort when they were assigned to the high-scoring students, but not when assigned to the low-scoring students.

Panel C shows the interaction between tracking and class size. Since class sizes are very large in Kenya, it could be concluded that the impact of tracking is large in this setting just because tracking is effective when class sizes are large. However, the impact of tracking seems to slightly decline with class size. While the range of class sizes in the sample does not allow us to conclude that tracking would be as effective with the very small class sizes observed in rich countries, there is thus no prima facie evidence that the effect is only driven by very large classes.

These estimates suggest that a large decrease in class heterogeneity has a positive impact on student's learning, even for initially low achieving students for whom it is coupled with a large decline in peer quality. The next section focuses on the median students, to separately identify the impact of a large change in peer quality.

3.2 The Effects of Large Change in Peer Quality: Regression Discontinuity Estimates

The main thrust of the regression discontinuity estimates of peer effects are shown in figure 3. In panel A, following Lee (2008), we regress test scores on a second-order local polynomial in initial percentile separately for students below the 50^{th} percentile cutoff and students above the 50^{th} 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 assignment to the bottom section in tracking schools (being at the 50^{th} percentile score). As is apparent from the figure, there is no discontinuity in test scores at the 50^{th} percentile cutoff in the tracking schools, 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 distribution. When we use a linear fit, rather than a polynomial, we again do not see an effect of the section 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 section.

We examine this result in a regression framework in table 4. In panel A, we estimate the effect of being assigned to the bottom section for the median student in tracking schools. Columns 1 and 2 present estimates of equations (3) and (4), respectively: the end line test score is regressed on a cubic of original percentile of a child in the distribution in his school, and a dummy for whether he is in the bottom half of the class. Column 6 presents estimates of equation (5), and column 7 adds a school fixed effect. Neither of these specifications shows any significant effect of being in the bottom half of the class in tracking schools. To assess the robustness of these results, columns 3 through 5 experiment with two other ways to specify the control function in the regression discontinuity design estimates: column 5 follows Imbens and Lemieux (2007) and estimate a Fan locally weighted regression on each side of the discontinuity. [18] The specifications in columns 3 and 4 are similar to equations (3) and (4), but the cubic is replaced by a quadratic allowed to be different on both sides of the discontinuity. The results confirm what the graphs show: despite the big gap in average peer achievement (1.6 standard deviations of the baseline school grade), the marginal students' final test scores do not seem to be significantly affected by assignment to the bottom section.

Panel B shows the instrumental variable estimate of the impact of classmates' average test score on a child's test score. We use the average end line 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

[18] Since the result is completely insensitive to the choice of bandwidth, we do not implement the cross-validation strategy they recommend.

shown in panel C, and shows that the average of the end line test scores of a child's classmates is about 0.76 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. For example the specification in column 3, which has school fixed effects and uses all the data, suggests that an increase in one standard deviation in the classmates' average test score *reduces* a child's test score by 0.007 standard deviations, a point estimate extremely close to zero. The 95 percent confidence interval in this specification is [-0.19; 0.19]. Thus, we are able to reject at 95 percent confidence reasonably modest direct effects of the average of one's peers on a student's performance for the median child in a tracking environment.[19]

In table 5 we combine regression discontinuity estimates in tracking and non-tracking schools to estimate impact of tracking for a median student assigned to either the top or bottom section. Specifically, we estimate the specifications reported in table 4 for the non-tracking schools, and we then report the predicted difference in test scores in tracking versus non-tracking schools for a student at the 50^{th} percentile student from the right, and from the left (in column 2 the tracking school equation and the non-tracking school equations are estimated jointly). Consistently across specifications, we find that students near the median of the pre-test distribution have higher test scores at the end line in tracking schools than non-tracking schools, and that the effect of tracking is very similar regardless of the section to which the child was assigned.

These results are quite striking. They imply that being the best student 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.

[19] With the caveat, mentioned above, that there may be a direct impact of one's rank in the class, which would violate the identification assumption that the only channel of impact of being assigned to the bottom track is through classmates' average score.

This rejects not only a model of peer effects that is purely "linear in means," but also a broader class of models in which peer effects are monotonic in peer achievement. Beyond this, the findings also seems difficult to reconcile with the idea that the reason heterogeneity in prior achievement reduces average test scores is that the teacher is teaching to the median student in the class and learning drops off the farther students are from this achievement level. Under this model, the student originally at the 50^{th} percentile would do better without tracking. It suggests that students benefit from homogeneity because the teacher can more narrowly tailor teaching to the class (called the "boutique" or "focus" models of teaching by Hoxby and Weingarth, 2006), and does not need to spend time addressing the needs of students at very varied levels. We will provide some additional evidence of what could be happening in the classroom in Section 4. To shed more light on whether small variations in the peer group may have a different impact than large and salient changes (as in the tracking case), we now turn to estimating the effect of peer on test scores using the small variations generated by the random assignment of peers in the non-tracking schools.

3.3. Peer Effects: Random Variation in Peer Composition

The regression discontinuity design approach we have discussed generates large differences in average initial peer achievement, to which teachers are potentially able to react by adjusting teaching methods. As our conceptual framework suggests, teachers are less likely to adjust teaching methods in response to smaller variation in students' achievement, such as that stemming from year-to-year variation in the characteristics of the student population in a school. This may explain the difference between our results and those suggesting that students benefit from academically stronger peers (e.g., Hoxby 2000).

To investigate this further, we exploit another source of variation in peer group average achievement generated by the design of our study, which produces local random variation in peer quality. In non-tracking schools, students were randomly assigned to sections and very few were reassigned later.[20] 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 class composition similar to those introduced by Hoxby (2000), with the difference that we use actual random variation in peer group composition, but have lower sample size. The results are presented in table 6.

The reduced form results (panel A, column 1) show that students benefit from stronger peers: the coefficient on the average base line test score is 0.42 with a standard deviation of 0.14. This coefficient is not comparable with other estimates in the literature since we are using the school grade sheets, which are not comparable across schools, and so we are standardizing the baseline scores in each school. Thus, in panel B, we use the average baseline scores of peers to instrument for their average end line score (the first stage is presented in panel C). Column 1 suggests that a one standard deviation increase in average peer end line test score would increase the test score of a student by 0.53 standard deviations, an effect comparable to that found in previous literature, and if anything on the higher side.[21]

Taken together with the results on the impact of tracking, these results suggest that children impact their peers through two very distinct channels: directly, and indirectly, through their impact on teacher behavior. The direct impact of having strong peer is

[20] We used only the initial assignment (which was random) in all specifications, not the section the student eventually attended.

[21] Of course, these estimates come from variations in peer test scores that are smaller than one standard deviation and, as our results warn us, the interpolation to one standard deviation may not actually be legitimate: the linear approximation is valid only locally. However, presenting the results in term of the impact of a one standard deviation change in peers' test scores allows us to compare our results to that of the literature, which also uses local variation in average test scores, and generally expresses the results in terms of the impact of one standard deviation increase in average test scores.

positive. But the indirect impact of having peers of one's own level is sufficiently strong to compensate it. Interestingly, combining the direct effect of peers with the fact that the median children did not suffer from being affected to the bottom track suggest that either median children benefit more from being taught easier material, or that the best students in a class get more teacher attention, and benefit from it, than the worst students.

Taken together, these results suggest that peer effect estimates arising from "local" variations in the composition of the student population, while interesting in their own right, are not a particularly helpful guide to what would happen in a tracking environment because they do not capture endogenous teacher responses to the changes in class composition generated by tracking.

Columns 2 and 3 show no effect of the variance of peer test scores on a student's own scores, but the standard errors are large: it is therefore difficult to conclude whether small changes in class heterogeneity are too small to motivate a teacher's response (and there is no direct impact of peer heterogeneity on results), or whether there is an effect of even small changes in peer variance on test scores, but our sample size does not allow us to detect it.

4. Why Did Tracking Work? Exploring the Channels

The basic results suggest that tracking increased test scores for all students, regardless of the group to which they were assigned, and regardless of their place in the initial achievement distribution. The fact that students near the median of the pre-test distribution were not hurt by being assigned to the bottom section, contrasted with the positive peer effects from being randomly assigned somewhat higher-scoring classmates in the non-tracking schools, suggests that students may have benefited from more focused teaching and perhaps more teacher effort. In this section, we explore two additional

pieces of evidence that shed some light on this question. First, we look at teacher presence and effort (do they spend more time in class and teaching?). Second, we disaggregate the test score 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 includes teacher age and experience teaching). We present the results separately for civil-service teachers and new contract teachers because they face very different incentives. The new teachers are on short-term (one year) contracts, and have incentives to work hard to increase their chances both of having their short-term contracts renewed, and of eventually being hired as civil-service teachers. In contrast, the civil service teachers have weak extrinsic incentives, and may be more sensitive to intrinsic motivation.

Teachers in tracking schools are significantly more likely to both be in school and be in class teaching than teachers in non-tracking schools (table 7, columns 1 to 3). Overall, teachers in tracking schools are 9.6 percentage points (19 percent) more likely to be found in school and teaching during a random spot check than their counterparts in non-tracking schools.

There are, however, large differences across teachers. The new contract teachers attend more than the civil-service teachers, are more likely to be found in class and teaching (74 percent versus 45 percent for the civil-service teacher), and their absence rate is unaffected by tracking. The civil-service 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 points, and is also significant. Conditional on being in school, civil-service 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 civil-service teachers are 11 percentage points more likely to be in class and teaching when they are assigned to the top section in tracking schools than when they are assigned to non-tracking schools. This represents a 25 percent increase in teaching time. When civil-service teachers are assigned to the bottom section, they are about as likely to be teaching as their counterparts in non-tracking schools. Students' attendance is not affected by tracking or by the section they were assigned to (column 10).

These results suggest that teachers may be more motivated to teach a group of students with high initial scores than a group with low initial scores or a heterogeneous group. Recall from table 3 that students assigned to the top section with a civil-service teacher benefited more from tracking than those assigned to the bottom section with a civil-service teacher. Increased teacher effort may help explain why tracking raised test scores for high-scoring students assigned to civil-service teachers.

4.2 Focused Teaching

Another 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 more homogeneous group of students might allow teachers to adjust the material covered and the pace of instruction to

students' needs. For example, a teacher might begin with more basic material and instruct at a slower pace, providing more repetition and reinforcement when students are initially less prepared. 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. With a heterogeneous group, they may be compelled to cover both simple and advanced material, spending less time on each, which would hurt all students.

A way to capture this is to see whether children at different levels of the distribution gained from tracking differentially at different competencies. Table 8 reports specifications similar to equation (2), but 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 the estimates for math 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 times bottom half is positive for the easiest competencies, and negative for the hardest competencies. A chi-square test allows us to reject equality of the coefficients of the interaction in the "easy competencies" regression and the "difficult competencies" regression at the 5 percent 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.[22] An alternative explanation for these results, however, is that weak students stood to gain from any program on the easiest competencies (since they had not mastered them yet, and in 18 months they did not have time to master both easy and strong competencies), while strong students had already mastered them and would have benefited from any program at the competencies they did not master. To check that this is not the reason for the program's success, we present in appendix table 2 a "placebo" test, where we run the same specifications using the assignment to a contract teacher (which was also random) as the treatment variable. Being assigned to a contract teacher, rather than a regular teacher, also raises test scores: students who were assigned to the new contract teachers scored much higher at end line than those assigned to the regular teachers. However, there is no evidence that being assigned to a contract teacher disproportionately raised scores for initially highly ranked students on the most difficult competencies, or for initially low ranked students on the easiest competencies. This suggests that the effects found in table 8 are not mechanical, and that tracking raised scores by allowing teachers to focus instruction on tasks in which students needed help.

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

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

to our knowledge this paper provides the first experimental evaluation of the impact of tracking in any context.

After 18 months, the point estimates suggest that the average score of a student in a tracking school is 0.14 standard deviations higher than that of a student in a non-tracking school. These effects are persistent. One year after the program ended, students in tracking schools performed 0.16 standard deviations higher than those in non-tracking schools.

Moreover, tracking raised scores for students throughout the initial rank distribution. A regression discontinuity design approach reveals that students who were very close to the 50^{th} percentile of the initial rank distribution within their school scored similarly at the end line whether they were assigned to the top or bottom section. In each case, they did much better than their counterparts in non-tracked schools. In contrast, students in non-tracking schools scored higher if they were randomly assigned to peers with somewhat higher initial scores. This suggests that the impact of small variation in peer quality in a non-tracking environment is not a good indication of the potential effect of tracking, because it does not take into account endogenous teacher response to a more homogeneous peer group.

Greater homogeneity is presumably beneficial because it allows teachers to adapt material better to the students in their class. Consistent with this, we find that students assigned to the bottom section in the tracking school gained most in the easiest competencies, and least in the hardest competencies. The effort of civil-service teachers also increased in tracking schools when the teacher was assigned to the top section, suggesting that the typical civil-service teacher is more motivated when teaching students with greater initial achievement. 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 they are at least as motivated when teaching a homogeneous group, even if it is weaker on average.

These conclusions echo those reached by Borman and Hewes (2002), who find positive short- and long-term impacts of "Success for All." One of the components of this program, first piloted in the United States by elementary schools in Baltimore, Maryland, is to regroup students across grades for reading lessons targeted to specific performance levels for a few hours a day. Likewise, Banerjee, et al. (2007), who study a remedial education and computer-assisted learning programs in India, found that both programs were very effective, mainly because they allowed students to learn at their own levels of achievement. In the Kenyan context, Glewwe and Kremer (2007) find that provision of government textbooks raised test scores only for those students with high initial test scores, but not for those with more typical scores. They interpret these results as consistent with the idea that instruction is not oriented toward the typical student in many cases. 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 pertinent for them could potentially have large positive effects with little or no additional resource cost.

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

Our results may also have implications for debates over school choice and voucher systems. A central criticism of such programs is that they may wind up hurting some students if they lead to increased sorting of students by initial academic achievement and if all students benefit from having peers with higher initial achievement. If, on the other hand, tracking is beneficial, this might be less of a concern.

Two limitations regarding external validity are worth highlighting. First, in this program, teachers were randomly assigned to sections. If the best teachers are assigned to the highest achieving students, the initially lower achieving students could be hurt.[23] Explicit and implicit incentives for teachers in Kenya and many other countries are based on end line performance of the group of students, rather than value added. If evaluations of a teacher's performance were on a value-added basis, teachers might be happier to work with initially lower-achieving students.

Second, it is an open question whether similar results would be obtained in different contexts. In developed countries for example, much smaller class sizes may allow more tailored instruction even without tracking, and extra resources (such as remedial education, computer-assisted learning, and special education programs) may already provide tools to help teachers deal with different types of students. Tracking may not be as beneficial in this type of environment. Our results are more likely to be directly applicable to settings where classes are large, student population is very heterogeneous, and there are little additional resources available to teachers, which are conditions found in many developing countries where primary enrollment has rapidly expanded in the last decade.

[23] Note, however, that in our setting it seems likely that if choice had been allowed, the more powerful teachers would have been assigned to the stronger group, and since the more powerful teachers are the civil-service teachers, who are also the worst teachers in this settings, this would have benefited the weak students.

References

Andrabi, Tahir, Jishnu Das, Asim Khwaja, and Tristan Zajonc (2008). Do Value-Added Estimates Add Value ? Accounting for Learning Dynamics. Mimeo, Harvard University.

Black, Dan A., Galdo, Jose and Smith, Jeffrey A. (2007) "Evaluating the Worker Profiling and Reemployment Services System Using a Regression Discontinuity Approach." *American Economic Review*, May (*Papers and Proceedings)*, 97(2), pp. 104-107.

Banerjee, Abhijit, Cole, Shawn, Duflo, Esther and Linden, Leigh.(2007) "Remedying Education: Evidence from Two Randomized Experiments in India." *Quarterly Journal of Economics*, August, 122(3), pp. 1235-1264.

Borman, Geoffrey D. and Hew, Gina M. (2002) "The Long-Term Effects and Cost-Effectiveness of Success for All." *Educational Evaluation and Policy Analysis*, Winter, 24(4), pp. 243-266.

Betts, Julian R. and Shkolnik, Jamie L. (1999) **"**Key Difficulties in Identifying the Effects of Ability Grouping on Student Achievement." *Economics of Education Review,* February, 19(1), pp. 21-26.

Clark, Damon. (2007) "Selective Schools and Academic Achievement." Institute for the Study of Labor (IZA) Working Paper No. 3182, November.

Ding, Weili and Lehrer, Steven F. (2007) "Do Peers Affect Student Achievement in China's Secondary School?" *The Review of Economics and Statistics*, February, 89(2), pp. 300-312.

Glewwe, Paul W., Kremer, Michael and Moulin, Sylvie. (2007). "Many Children Left Behind? Textbooks and Test Scores in Kenya." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 13300.

Hoxby, Caroline. (2000) "Peer Effects in the Classroom: Learning from Gender and Race Variation." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 7867.

Hoxby, Caroline and Weingarth, Gretchen. (2006) "Taking Race Out of the Equation: School Reassignment and the Structure of Peer Effects." Unpublished manuscript, Harvard University.

Imbens, Guido and Lemieux, Thomas. (2007). "Regression Discontinuity Designs: A Guide to Practice." National Bureau of Economic Research (Cambridge, MA) Working Paper No. 13039.

Lavy, Victor, Daniel Paserman and Analia Schlosser (2008) "Inside the Black Box of Ability Peer Effect: Evidence from Variation of Low Achiever in the Classroom" NBER working paper No 14415

Lee, David S. (2008). "Randomized experiments from non-random selection in U.S. House elections". *Journal of Econometrics*, 142(2), pp. 675-697.

Lyle, David S. (2007). "Estimating and Interpreting Peer and Role Model Effects from Randomly Assigned Social Groups at West Point." *Review of Economics and Statistics*, May, 89(2), pp. 289-299.

Manning, Allen and Pischke, Jörn-Steffen. (2006). "Comprehensive Versus Selective Schooling in England & Wales: What Do We Know?" Centre for the Economics of Education (LSE) Working Paper No. CEEDP006.

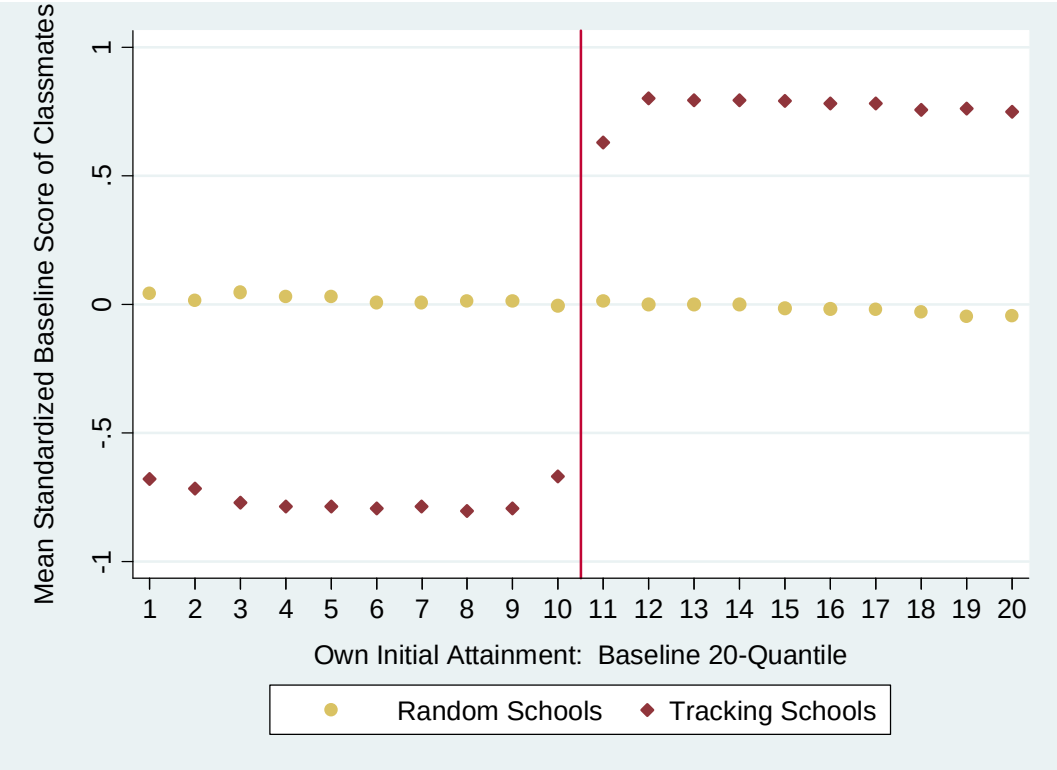
Piketty, Thomas. (2004) "L'Impact de la taille des classes et de la ségrégation sociale sur la réussite scolaire dans les écoles françaises : une estimation à partir du panel primaire 1997. " Unpublished manuscript, PSE, France.

UNESCO. (2006). United Nations Education, Scientific and Cultural Organization. *Fact Book on Education for All*. Nairobi: UNESCO Publishing, 2006.

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

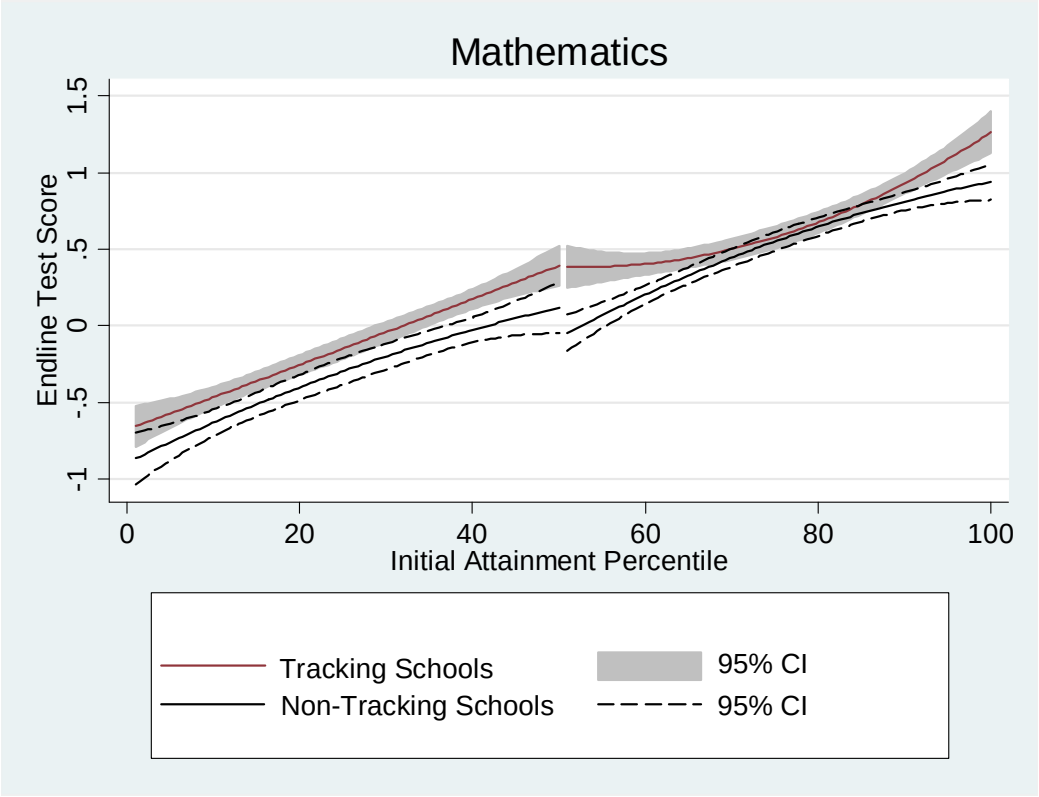
Zimmerman, David J. (2003). "Peer Effects in Academic Outcomes: Evidence from a Natural Experiment." *The Review of Economics and Statistics*, November, 85(1), pp. 9-23.

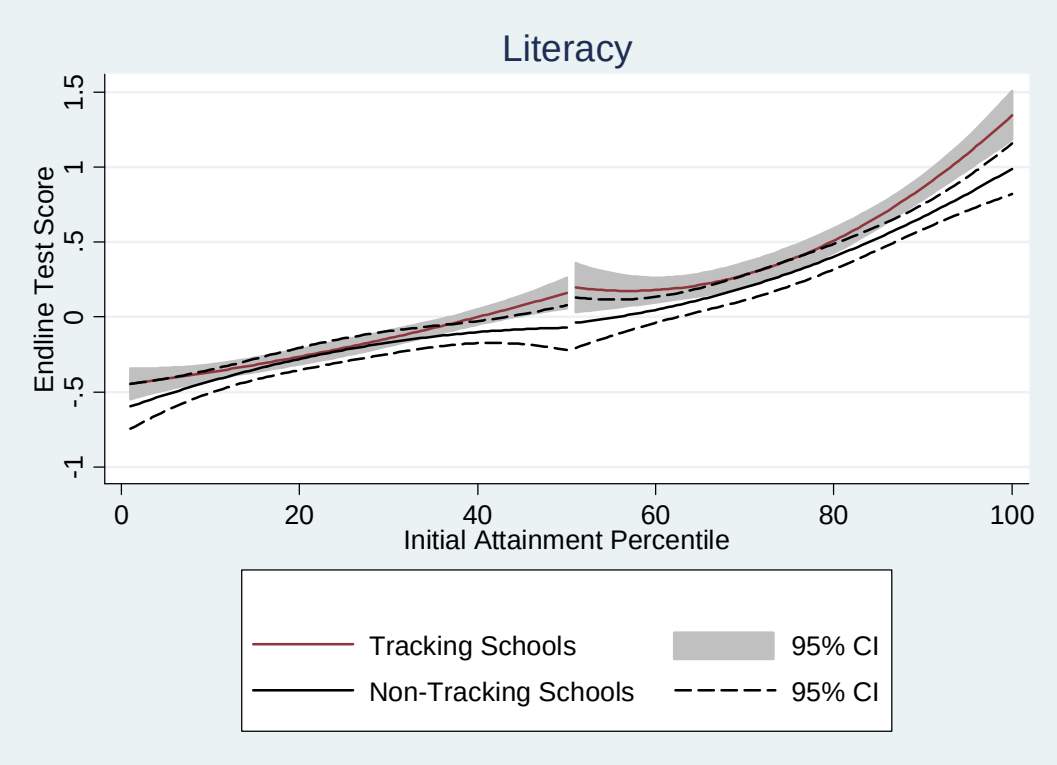
Figure 1: Experimental Variation in Peer Composition
Tracking vs. Non-Tracking Schools



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 End Line Score by Initial Attainment

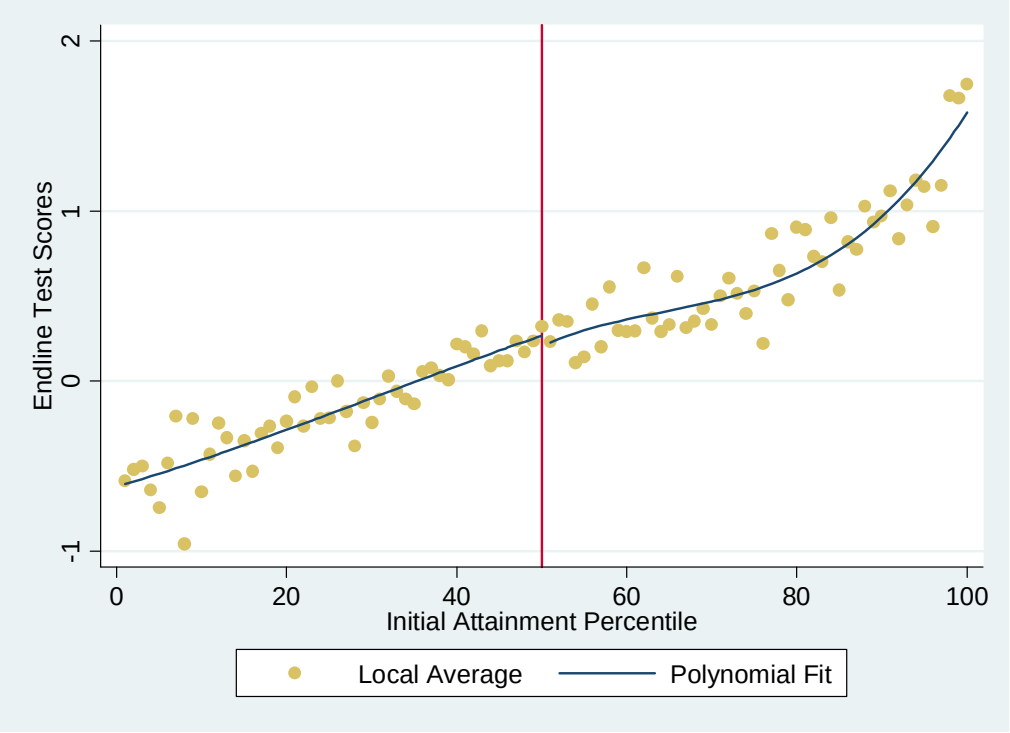




Notes: Fitted values from regressions that include a second order polynomial estimated separately on each side of the percentile=50 threshold.

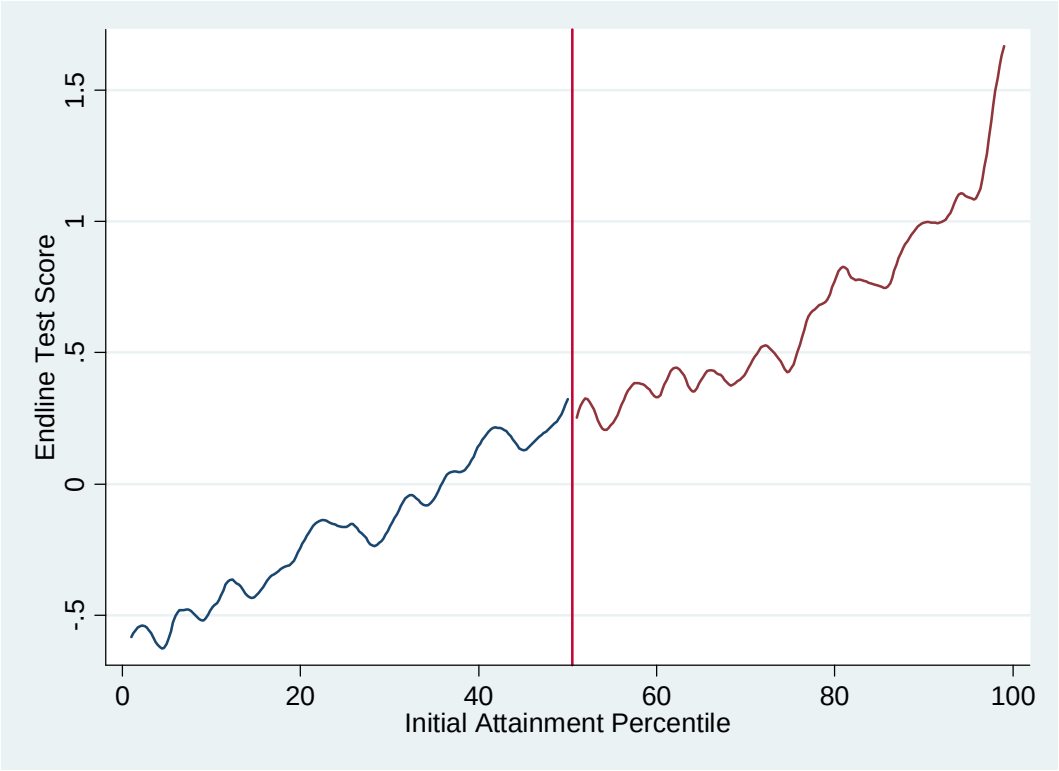
Figure 3: Peer Quality and End Line Scores in Tracking Schools

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.

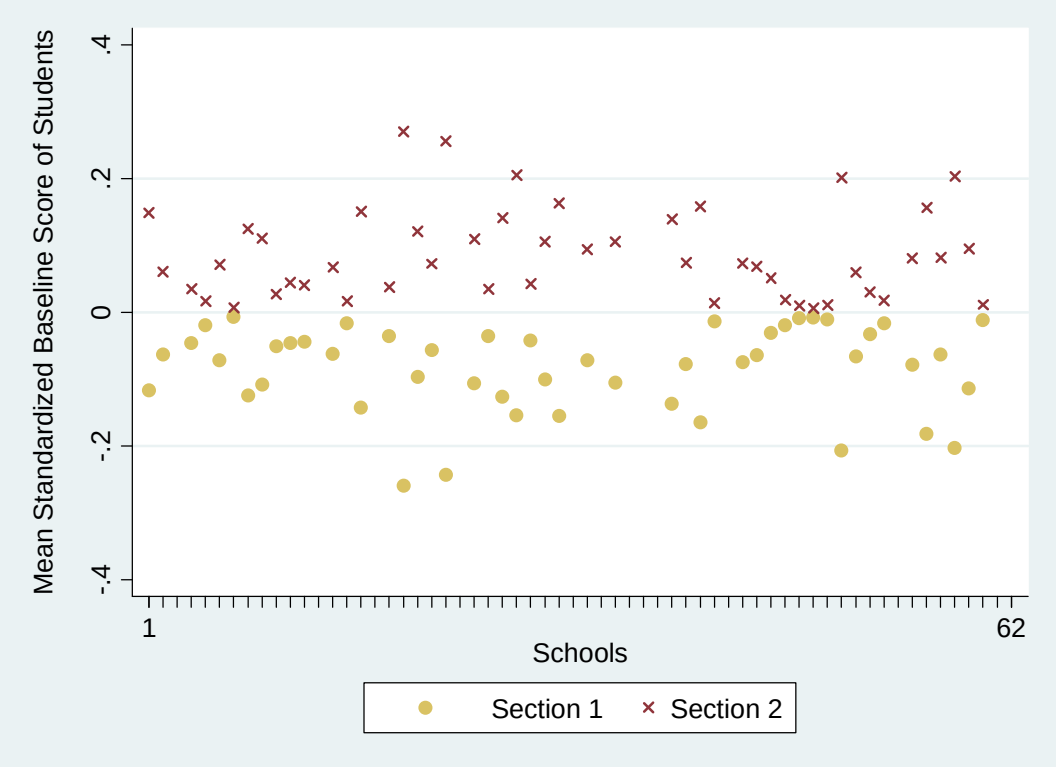
Panel B. Fan Locally-Weighted regression



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
(Non-Tracking Schools)



Note: Schools are ordered alphabetically along 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			
	Non-Tracking Schools		Tracking Schools		Bottom Section		Top Section	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Panel A. Baseline School Characteristics								
Total enrollment in 2004	589	*232*	549	*198*				
Number of government teachers in 2004	11.6	*3.3*	11.9	*2.8*				
School pupil/teacher ratio	37.1	*12.2*	35.9	*10.1*				
Performance at national exam in 2004 (out of 400)	256	*24*	258	*23*				
Panel B. Class Size Prior to Program Inception (March 2005)								
Average class size in first grade	91	*37*	89	*33*				
Proportion of female first grade students	0.49	*0.06*	0.49	*0.05*				
Average class size in second grade	96	*41*	91	*35*				
Panel C. Class Size 6 Months After Program Inception (October 2005)								
Average class size in first grade	47	*17*	45	*15*				
Range of class sizes in sample (first grade)	19-98		20-97					
Panel D. Class Size in Year 2 of Program (March 2006)								
Average class size in first grade	78	*32*	75	*28*				
Average class size in second grade	46	*15*	45	*15*				
Range of class sizes in sample (second grade)	18-93		21-95					
Panel E. Within Tracking Schools: Students Characteristics by Tracking Status								
Proportion Female					0.48	*0.50*	0.49	*0.50*
Average Age at Endline					8.92	*1.46*	9.39	*1.46*
Was in preschool in 2004					0.07	*0.25*	0.04	*0.20*
Was in grade 1 in 2004					0.01	*0.08*	0.01	*0.12*
Average Standardized Baseline Score (Mean 0, Std. Dev. 1 at school level)					-0.80	*0.51*	0.80	*0.66*
Average Standardized Endline Score (Mean 0, Std. Dev. 1 at school level)					-0.44	*0.83*	0.42	*0.95*
Number of Schools	**61**		**60**		**60**			

Table 2: Overall Effect of Tracking

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Total Score				Mathematics Score				Literacy Score			
Panel A: Short-Run Effects (after 18 months in program)												
(1) Tracking School	0.138	0.175	0.191	0.18	0.124	0.158	0.138	0.155	0.122	0.154	0.197	0.165
	(0.078)*	(0.077)**	(0.093)**	(0.092)*	(0.065)*	(0.064)**	(0.073)*	(0.082)*	(0.08)	(0.083)*	(0.107)*	(0.098)*
(2) In Bottom Half of Initial Distribution			-0.036				0.039				-0.09	
x Tracking School			(0.07)				(0.07)				(0.08)	
(3) In Bottom Quarter				-0.044				0.011				-0.081
x Tracking School				(0.08)				(0.09)				(0.08)
(4) In Second to Bottom Quarter				-0.014				0.025				-0.042
x Tracking School				(0.07)				(0.08)				(0.07)
(5) In Top Quarter				0.028				-0.025				0.066
x Tracking School				(0.08)				(0.07)				(0.08)
Initial attainment percentile		0.018	0.02	0.022		0.017	0.019	0.021		0.014	0.017	0.019
		(0.001)***	(0.001)***	(0.002)***		(0.001)***	(0.001)***	(0.002)***		(0.001)***	(0.001)***	(0.002)***
Individual Controls	no	yes	yes	yes	no	yes	yes	yes	no	yes	yes	yes
Observations	5796	5282	5282	5282	5797	5283	5283	5283	5797	5283	5283	5283
Total effects on bottom half and bottom quarter												
Coeff (Row 1)+Coeff (Row 2)			0.155				0.177				0.107	
Coeff (Row 1)+Coeff (Row 3)				0.136				0.166				0.084
F Test: Total Effect = 0			4.39	2.864			5.93	3.925			2.37	1.455
p-value			0.04	0.093			0.02	0.05			0.13	0.23
Panel B: Longer-Run Effects (a year after program ended)												
(1) Tracking School	0.163	0.178	0.216	0.235	0.128	0.131	0.143	0.168	0.16	0.18	0.231	0.241
	(0.069)**	(0.072)**	(0.079)***	(0.088)***	(0.059)**	(0.062)**	(0.064)**	(0.074)**	(0.075)**	(0.078)**	(0.089)**	(0.096)**
(2) In Bottom Half of Initial Distribution			-0.081				-0.027				-0.106	
x Tracking School			(0.06)				(0.06)				(0.06)	
(3) In Bottom Quarter				-0.116				-0.042				-0.151
x Tracking School				(0.09)				(0.10)				(0.085)*
(4) In Second to Bottom Quarter				-0.096				-0.074				-0.091
x Tracking School				(0.07)				(0.07)				(0.07)
(5) In Top Quarter				-0.028				-0.04				-0.011
x Tracking School				(0.07)				(0.06)				(0.08)
Initial attainment percentile		0.017	0.019	0.021		0.016	0.018	0.02		0.015	0.017	0.018
		(0.001)***	(0.001)***	(0.002)***		(0.001)***	(0.001)***	(0.002)***		(0.001)***	(0.001)***	(0.002)***
Individual Controls	no	yes	yes	yes	no	yes	yes	yes	no	yes	yes	yes
Observations	5490	5003	5003	5003	5490	5003	5003	5003	5496	5009	5009	5009
Total effects on bottom half and bottom quarter												
Coeff (Row 1)+Coeff (Row 2)			0.135				0.116				0.125	
Coeff (Row 1)+Coeff (Row 3)				0.119				0.126				0.09
F Test: Total Effect = 0			2.92	1.48			2.43	1.56			2.50	1.03
p-value			0.09	0.23			0.12	0.22			0.12	0.31

*Notes: The sample includes 60 tracking and 61 non-tracking schools. The dependent variables are normalized test scores, with mean 0 and standard deviation 1 in the non-tracking schools. Robust standard errors clustered at the school level are presented in parentheses. ***, **, * indicates significance at the 1%, 5% and 10% levels respectively. Individual controls included: age, gender, being assigned to the contract teacher, dummies for initial half/quarter, and initial attainment percentile. We lose observations when adding individual controls because information on the initial attainment could not be collected in some of the non-tracking schools.*

Table 3
Testing for Heterogeneity in Effect of Tracking on Total Score

	Short-Run: After 18 months in program			Longer-Run: a year after program ended		
	Effect of Tracking on Total Score for		*Test (Top = Bottom)*	*Effect of Tracking on Total Score for*		*Test (Top = Bottom)*
	Bottom Half	Top Half	p-value	Bottom Half	Top Half	p-value
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: By Gender						
Boys	0.128	0.162	0.724	0.084	0.206	0.168
	(0.075)*	(0.100)		(0.083)	(0.084)**	
Girls	0.187	0.221	0.666	0.190	0.227	0.641
	(0.089)**	(0.104)**		(0.098)*	(0.089)**	
Test (Boys = Girls): p-value	0.410	0.475		0.238	0.767	
Panel B: By Teacher Type						
Regular Teacher	0.049	0.225	0.157	0.090	0.199	0.345
	(0.087)	(0.120)*		(0.099)	(0.098)**	
Contract Teacher	0.252	0.163	0.521	0.177	0.245	0.588
	(0.099)**	(0.118)		(0.094)*	(0.103)**	
Test (Regular = Contract): p-value	0.079	0.679		0.437	0.713	
Panel C: Interaction between Tracking and Class Size						
Coefficient on interaction	-0.005	-0.008	0.370	-0.006	-0.009	0.331
Tracking x Class size	(0.004)	(0.005)*		(0.004)	(0.005)*	

*Notes: The sample includes 60 tracking and 61 non-tracking schools. The dependent variables are normalized test scores, with mean 0 and standard deviation 1 in the non-tracking schools. Robust standard errors clustered at the school level are presented in parentheses. ***, **, * indicates significance at the 1%, 5% and 10% levels respectively. Individual controls included: age, gender, being assigned to the contract teacher, dummies for initial half, and initial attainment percentile.*

Table 4
Peer Quality: Regression Discontinuity Approach (Tracking Schools Only)

	Total Score						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Specification 1: With third order polynomial in baseline attainment		Specification 2: With second order polynomial in baseline attainment estimated separately on either side		Specification 3: With local linear regressions (Fan)	Specification 4: Pair around the median	
Panel A: Reduced Form							
Estimated Effect of Bottom Section at 50th percentile	0.009	0.001	-0.045	-0.051	0.089	-0.057	-0.043
	(0.093)	(0.079)	(0.106)	(0.089)	(0.329)	(0.141)	(0.157)
Observations (Students)	2959	2959	2959	2959	2959	149	149
School Fixed Effects	no	yes	no	yes	no	no	yes
Panel B: IV							
Mean Total score of Peers	-0.013	-0.001				-0.073	-0.005
	(0.127)	(0.102)				(0.221)	(0.236)
Observations (Students)	2959	2959				149	149
School Fixed Effects	no	yes				no	yes
Panel C: First Stage for IV	*Dep. Var: Average Total Score of Peers*					*Dep. Var: Average Total Score of Peers*	
In Bottom Half of Initial Distribution	-0.731	-0.743				-0.612	-0.607
	(0.047)***	(0.021)***				(0.090)***	(0.058)***
Observations (Students)	2959	2959				149	149
R-squared	0.42	0.78				0.25	0.57
School Fixed Effects	no	yes				no	yes

*Notes: Sample restricted to the 60 tracking schools (where students were tracked into two sections by initial attainmnent). Students in the bottom half of the initial distribution were assigned to the "bottom section" where the average peer quality was much lower than in the top section (see Figure 1). Panel A, columns 1-2 and 6-7: the score was regressed on a dummy "assigned to bottom section". We present the estimated coefficient of the dummy "assigned to bottom section". Standard errors clustered at school level. ***, **, * indicates significance at the 1%, 5% and 10% levels respectively.*
*Panel A, columns 3-5: The estimated effect of being assigned to the bottom section is the difference between the estimates of the expectation function estimated separately on either side of the 50th percentile. In columns 3-4, the score was regressed on a second order polynomial in initial percentile fully interacted with a dummy for "bottom section". In column 5, the score estimated through local linear regression (bandwidth = 2). Bootstrapped standard errors clustered at the school level. ***, **, * indicates significance at the 1%, 5% and 10% levels respectively.*
All regressions also include individual controls (age, gender, and dummy for being assigned to the contract teacher).

Regressions in columns 7-8 include 1 pair of students per school: The top student in the bottom section and the bottom student in the top section. 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 5

Interactions between Tracking and Peer Quality: Regression Discontinuity Approach

All ETP schools (Tracking and Non-Tracking)

	Total Score		
	(1)	(2)	(3)
	Specification 1: With third order polynomial in baseline attainment	Specification 2: With second order polynomial in baseline attainment estimated separately on either side	Pair around 50th percentile
Estimated Effect of Tracking at	0.156	0.497	0.392
50th percentile, from the left	(0.035)***	(0.290) *	(0.147)***
Estimated Effect of Tracking at	0.192	0.596	0.392
50th percentile, from the right	(0.034)***	(0.291) **	(0.139)***
Observations	5282	5282	306

Notes: Sample includes 61 non-tracking schools and 60 tracking Schools. 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).

Column 1: Estimated effect of "Tracking" at the 50th percentile computed as follows: The score was regressed on a third order polynomial in initial percentile, and a dummy for "bottom section" fully interacted with a dummy for "tracking". The estimated effect "from the left" is the difference between the estimates of the expectation function (Tracking vs. Non-Tracking schools) on the left side of the 50th percentile, i.e. for the median student assigned to the top stream. The estimated effect "from the right" is the difference between the estimates of the expectation function (Tracking vs. Non-Tracking schools) on the right side of the 50th percentile, i.e. for the median student assigned to the bottom stream. Bootstrapped standard errors clustered at the school level.

Column 2: The score was regressed on a second order polynomial in initial percentile fully interacted with a dummy for "bottom section", as well as a individual controls (age, gender, and dummy for being assigned to the contract teacher), separately for Tracking and Non-Tracking schools. The estimated effect "from the left" is the difference between the estimates of the expectation function (Tracking vs. Non-Tracking schools) on the left side of the 50th percentile, i.e. for the median student assigned to the top stream. The estimated effect "from the right" is the difference between the estimates of the expectation function (Tracking vs. Non-Tracking schools) on the right side of the 50th percentile, i.e. for the median student assigned to the bottom stream. Bootstrapped standard errors clustered at the school level

*Column 3: sample restricted to 1 pair of students per school: the two students on either side of the 50th percentile. The number of observation is above 242 because of ties. The score was regressed on a dummy for "tracking" and the interaction "tracking x bottom half of initial distribution", as well as individual controls (age, gender, a dummy for being assigned to the contract teacher, and a dummy for the bottom half of the initial distribution). Standard errors are clustered at the school level. ***, **, * indicates significance at the 1%, 5% and 10% levels respectively.*

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

	Total Score			Math Score	Lit Score
	(1)	(2)	(3)	(4)	(5)
Panel A: Reduced Form					
Average Baseline Score of Classmates‡	0.41		0.41	0.38	0.36
	(0.14)***		(0.14)***	(0.16)**	(0.12)***
Std. Dev. in Baseline Score of Classmates		0.05	0.02	-0.23	0.24
		(0.17)	(0.17)	(0.18)	(0.16)
Observations	2187	2187	2187	2187	2188
School Fixed Effects	x	x	x	x	x
Panel B: IV					
Average Endline Score of Classmates (predicted)	0.534		0.533	0.525	0.541
	(0.183)***		(0.186)***	(0.223)**	(0.181)***
Std. Dev. in Baseline Score of Classmates (predicted)		0.14	0.07		
		(0.52)	(0.52)		
Observations	2187	2187	2187	2187	2187
School Fixed Effects	x	x	x	x	x
	Dep. Var. Average Endline Score of	*Dep. Var.* Std. Dev. In Endline Score of		*Dep. Var.* Average Endline Score	*Dep. Var.* Average Endline Score
Panel C: First-Stage for IV	Classmates	Classmates		of Classmates	of Classmates
Average (Standardized) Baseline Score of Classmates‡	0.771			0.702	0.674
	(0.032)***			(0.033)***	(0.030)***
Std. Dev. in Baseline Score of Classmates		0.336			
		(0.025)***			

*Notes: Sample restricted to schools where students were randomly assigned to a section. Individual controls included but not shown: gender, age, being assigned to the contract teacher, and own baseline score. Robust standard errors clustered at the section level in parentheses. ***, **, * indicates significance at the 1%, 5% and 10% levels respectively.*

‡ *This variable has a mean of 0.000 and a standard deviation of 0.107. We define classmates as follows: two students in the same section are classmates; two students in the same grade but different sections are not classmates.*

Table 7
Teachers and Students Presence

	All Teachers			Government Teachers			ETP Teachers			Students
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Teacher Found in school on random school day	*If in school, found in class teaching*	*Teacher found in class teaching (unconditional on presence)*	*Teacher Found in school on random school day*	*If in school, found in class teaching*	*Teacher found in class teaching (unconditional on presence)*	*Teacher Found in school on random school day*	*If in school, found in class teaching*	*Teacher found in class teaching (unconditional on presence)*	*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	-0.015
	(0.021)**	(0.039)**	(0.038)**	(0.025)**	(0.047)**	(0.044)**	(0.034)	(0.036)	(0.045)	(0.014)
Bottom Half x Tracking School	-0.049	-0.034	-0.062	-0.073	-0.036	-0.076	0.036	-0.034	-0.004	0.003
	(0.029)*	(0.042)	(0.040)	(0.034)**	(0.059)	(0.053)	(0.046)	(0.050)	(0.057)	(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	-0.005
	(0.018)	(0.027)	(0.026)	(0.020)	(0.035)***	(0.031)***	(0.032)	(0.037)	(0.043)	(0.004)
Assigned to Contract Teacher										0.011
										(0.005)**
Assigned to Contract Teacher x Tracking School										0.004
										(0.008)
Observations	2098	1782	2098	1633	1367	1633	465	415	465	44059
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
F (test of joint significance)	2.718	7.693	9.408	2.079	4.414	5.470	2.426	2.570	3.674	5.465
p-value	0.011	0.000	0.000	0.050	0.000	0.000	0.023	0.016	0.001	0.000

*Notes: Linear probability model regressions. Multiple observations per teacher and per student. Standard errors clustered at school level. ***, **, * indicates significance at the 1%, 5% and 10% levels respectively. Region and date of test dummies were included in all regressions but are not shown. In colums 10 and 11, a dummy for "Bottom Half" is included in the regression but 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	-0.49		-3.86	-4.05	-4.15	-1.15
	(0.09)***	(0.08)***	(0.05)***		(0.33)***	(0.42)***	(0.40)***	(0.21)***
(2) Tracking School	0.15	0.16	0.21	X^2 = 0.66	1.63	1.00	1.08	0.38
	(0.10)	(0.12)	(0.10)**	p-value = 0.417	(0.65)**	(0.78)	(0.75)	(0.34)
(3) In Bottom Half of Initial Distribution	0.18	0.08	-0.10	X^2 = 3.97	-0.42	-0.61	-0.39	-0.44
x Tracking School	(0.14)	(0.12)	(0.08)	p-value = 0.046	(0.46)	(0.61)	(0.56)	(0.30)
Constant	4.93	1.82	0.57		11.64	10.06	10.12	3.94
	(0.23)***	(0.22)***	(0.16)***		(1.00)***	(1.20)***	(1.12)***	(0.56)***
Observations	5284	5284	5284		5283	5279	5284	5284
Maxiumum possible score	6	6	6		24	24	24	24
Mean in Non-Tracking Schools	4.16	1.61	0.67		6.99	5.52	5.00	2.53
Std Dev in Non-Tracking Schools	2.02	1.62	0.94		6.56	7.61	7.30	3.94
Total effect of tracking on bottom half:								
Coeff (Row 2)+Coeff (Row 3)	0.33	0.24	0.11	X^2 = 2.34 p-value = 0.126	1.21	0.39	0.69	-0.06
F Test: Coeff (Row 2)+Coeff (Row 3) = 0	3.63	6.39	4.42		4.74	0.70	1.82	0.09
p-value	0.06	0.01	0.04		0.03	0.40	0.18	0.76

*Notes: The sample includes 60 Tracked Schools and 62 Untracked Schools. Robust standard errors clustered at the school level are presented in parentheses. ***, **, * indicates significance at the 1%, 5% and 10% levels respectively.*

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?

	At Endline Test (after 18 months in program)			At Long-Run Follow-up Test (a year after program ended)
	(1)	(2)	(3)	(4)
	Transferred to other school	If not transferred: missed test	Total Attrition	Total Attrition
Tracking School	0.000	0.019	0.019	0.002
	(0.012)	(0.016)	(0.020)	(0.026)
In Bottom Half of Initial Distribution	0.027	-0.013	0.014	0.05
	(0.016)*	(0.020)	(0.025)	(0.028)*
In Bottom Half of Initial Distribution x Tracking School	-0.008	0.003	-0.005	0
	(0.013)	(0.012)	(0.017)	(0.021)
Girl	0.012	0.029	0.041	0.024
	(0.009)	(0.014)**	(0.016)**	(0.016)
Girl x Tracking School	0.004	-0.048	-0.044	-0.022
	(0.013)	(0.016)***	(0.020)**	(0.021)
Assigned to Contract Teacher	-0.006	-0.019	-0.025	-0.014
	(0.010)	(0.011)*	(0.014)*	(0.015)
Assigned to Contract Teacher x Tracking School	0.023	-0.008	0.014	0.032
	(0.013)*	(0.014)	(0.018)	(0.020)
Constant	0.034	0.171	0.205	0.23
	(0.018)*	(0.026)***	(0.032)***	(0.035)***
Observations	7345	7345	7345	7340
Mean	0.057	0.119	0.175	0.224

Notes: OLS Regressions; standard errors clustered at school level. Additional controls not shown: a third degree polynomial in the student's percentile in the initial attainment distribution.

Appendix Table 2
Falsification Test for Table 8: Effect of *Contract Teachers* 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	-1.25	-1.05	-0.46		-4.01	-4.05	-4.14	-1.40
	(0.12)***	(0.09)***	(0.06)***		(0.51)***	(0.54)***	(0.54)***	(0.25)***
(2) Assigned to Contract Teacher	0.26	0.37	0.21	X^2 =0.32	0.72	1.43	1.36	0.45
	(0.08)***	(0.10)***	(0.09)**	(Prob > X^2) = 0.572	(0.54)	(0.65)**	(0.63)**	(0.28)
(3) In Bottom Half of Initial Distribution	-0.16	-0.23	-0.16	X^2 =0.00	-0.16	-0.74	-0.50	-0.01
x Contract Teacher	(0.18)	(0.16)	(0.12)	(Prob > X^2) = 0.965	(0.88)	(0.89)	(0.94)	(0.38)
Constant	4.82	1.69	0.55		11.82	9.82	9.88	3.93
	(0.23)***	(0.22)***	(0.16)***		(1.07)***	(1.23)***	(1.14)***	(0.58)***
Observations	5284	5284	5284		5283	5279	5284	5284
Maxiumum possible score	6	6	6		24	24	24	24
Mean (regular teacher)	4.19	1.57	0.67		7.36	5.43	4.92	2.40
Std Dev (regular teacher)	1.98	1.57	0.93		6.69	7.55	7.20	3.79
Total effect of tracking on bottom half:								
Coeff (Row 2)+Coeff (Row 3)	0.10	0.14	0.05	X^2 = 0.21 (Prob > X^2) = 0.65	0.56	0.69	0.86	0.44
F Test: Coeff (Row 2)+Coeff (Row 3) = 0	0.48	2.38	0.96		1.47	3.21	4.10	5.94
Prob > F	0.49	0.13	0.33		0.23	0.08	0.05	0.02

*Notes: The sample includes 60 Tracked Schools and 62 Untracked Schools. Robust standard errors clustered at the school level are presented in parentheses. ***, **, * indicates significance at the 1%, 5% and 10% levels respectively.*

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