NBER WORKING PAPER SERIES

PEER EFFECTS, TEACHER INCENTIVES, 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 anonymous referees and the editor for their suggestions. 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, Teacher Incentives, 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, Revised October 2009
JEL No. I20,O1

ABSTRACT

To the extent that students benefit from high-achieving peers, tracking will help strong students and hurt weak ones. However, all students may benefit if tracking allows teachers to present material at a more appropriate level. Lower-achieving pupils are particularly likely to benefit from tracking if teachers would otherwise have incentives to teach to the top of the distribution. We propose a simple model nesting these effects. We compare 61 Kenyan schools in which students were randomly assigned to a first grade class with 60 in which students were assigned based on initial achievement. In non-tracking schools, students randomly assigned to academically stronger peers scored higher, consistent with a positive direct effect of academically strong peers. However, compared to their counterparts in non-tracking schools, students in tracking schools scored 0.14 standard deviations higher after 18 months, and this effect persisted one year after the program ended. Furthermore, students at all levels of the distribution benefited from tracking. Students near the median of the pre-test distribution benefited similarly whether assigned to the lower or upper section. A natural interpretation is that the direct effect of high-achieving peers is positive, but that tracking benefited lower-achieving pupils indirectly by allowing teachers to teach at a level more appropriate to them.

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

To the extent that students benefit from having higher-achieving peers, tracking students into separate classes by prior achievement could disadvantage low-achieving students while benefiting high-achieving students, thereby exacerbating inequality (Denis Epple, Elizabeth Newton and Richard Romano, 2002). On the other hand, tracking could potentially allow teachers to more closely match instruction to students' needs, benefiting all students. This suggests that the impact of tracking may depend on teachers' incentives. We build a model nesting these effects. In the model, students can potentially generate direct student-to-student spillovers as well as indirectly affect both the overall level of teacher effort and teachers' choice of the level at which to target instruction. Teacher choices depend on the distribution of students' test scores in the class as well as on whether the teacher's reward is a linear, concave, or convex function of test scores. The further away a student's own level is from what the teacher is teaching, the less the student benefits; if this distance is too great, she does not benefit at all.

We derive implications of this model, and test them using experimental data on tracking from Kenya. In 2005, 140 primary schools in western Kenya received funds to hire an extra grade one teacher. Of these schools, 121 had a single first-grade class and split their first-grade class into two sections, with one section taught by the new teacher. In 60 randomly selected schools, students were assigned to sections based on prior achievement. In the remaining 61 schools, students were randomly assigned to one of the two sections.

We find that tracking students by prior achievement raised scores for all students, even those assigned to lower achieving peers. On average, after 18 months, test scores were 0.14 standard deviations higher in tracking schools than in non-tracking schools (0.18 standard deviations higher after controlling for baseline scores and other control variables). After controlling for the baseline scores, students in the top half of the pre-assignment distribution gained 0.19 standard deviations, and those in the bottom half gained 0.16 standard deviations. Students in all quantiles benefited from tracking. Furthermore, tracking had a persistent impact: one year after tracking ended, students in tracking schools scored 0.16 standard deviations higher (0.18 standard deviations higher with control variables). This first set of findings allows us to reject a special case of the model, in which all students benefit from higher-achieving peers but teacher behavior does not respond to class composition.

Our second finding is that students in the middle of the distribution gained as much from tracking as those at the bottom or the top. Furthermore, when we look within tracking schools using a regression discontinuity analysis, we cannot reject the hypothesis that there is no difference in endline achievement between the lowest scoring student assigned to the high-achievement section and the highest scoring student assigned to the low-achievement section, despite the much higher-achieving peers in the upper section.

These results are inconsistent with another special case of the model, in which teachers are equally rewarded for gains at all levels of the distribution, and so would choose to teach to the median of their classes. If this were the case, instruction would be less well-suited to the median student under tracking. Moreover, students just above the median would perform much better under tracking than those just below the median, for while they would be equally far away from the teacher's target teaching level, they would have the advantage of having higher-achieving peers.

In contrast, the results are consistent with the assumption that teachers' rewards are a convex function of test scores. With tracking, this leads teachers assigned to the lower-achievement section to teach closer to the median student's level than those assigned to the upper section, although teacher effort is higher in the upper section. In such a model, the median student may be better off under tracking and may potentially be better off in either the lower-achievement or higher-achievement section.

The assumption that rewards are a convex function of test scores is a good characterization of the education system in Kenya and in many developing countries. The Kenyan system is centralized, with a single national curriculum and national exams. To the extent that civil-service teachers face incentives, those incentives are based on the scores of their students on the national primary school exit exam given at the end of eighth grade. But since many students drop out before then, the teachers have incentives to focus on the students who are likely to take the exam, students at the very top of the first-grade class. Indeed, Glewwe, Kremer, and Moulin (2009) show that textbooks based on the curriculum benefited only the initially higher-achieving students, suggesting that the exams and associated curriculum are not well-suited to the typical student. These features seem common to most educational systems in developing countries.

The model also has implications for the effects of the test score distribution in non-tracking schools. Specifically, it suggests that an upward shift of the distribution of peer achievement will strongly raise test scores for a student with initial achievement at the

top of the distribution, have an ambiguous impact on scores for a student closer to the middle, and raise scores at the bottom. This is so because, while all students benefit from the direct effect of an increase in peer quality, the change in peer composition also generates an upward shift in the teacher's instruction level. The higher instruction level will benefit students at the top; hurt those students in the middle who find themselves further away from the instruction level; and leave the bottom students unaffected, since they are in any case too far from the target instruction level to benefit from instruction. Estimates exploiting the random assignment of students to sections in non-tracking schools are consistent with these implications of the model.

While we do not have direct observation on the instruction level and how it varied across schools and across sections in our experiment, we present some corroborative evidence that teacher behavior was affected by tracking. First, teachers were more likely to be in class and teaching in tracking schools, particularly in the high-achievement sections, a finding consistent with the model's predictions. Second, students in the lower half of the initial distribution gained comparatively more from tracking in the most basic skills, while students in the top half of the initial distribution gained more from tracking in the somewhat more advanced skills. This finding is consistent with the hypothesis that teachers are tailoring instruction to class composition, although this could also be mechanically true in any successful intervention.

Rigorous evidence on the effect of tracking on learning of students at various points of the prior achievement distribution is limited and much of it comes from studies of tracking in the U.S., a context that may have limited applicability for education systems in developing countries. Reviewing the early literature, Betts and Shkolnik (1999) conclude that while there is an emerging consensus that high-achievement students do better in tracking schools than in non-tracking schools and that low-achievement students do worse, the consensus is based largely on invalid comparisons. When they compare similar students in tracking and non-tracking high schools, Betts and Shkolnik (1999) conclude that low-achieving students are neither hurt nor helped by tracking; top students are helped; and there is some evidence that middle-scoring students may be hurt.

Another difficulty is that tracking schools may be different from non-tracking schools. 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. Three recent studies that

tried to address the endogeneity of tracking decisions have found that tracking might be beneficial to students, or at least not detrimental, in the lower-achievement tracks. First, Figlio and Page (2002) compare achievement gains across similar students attending tracking and non-tracking schools in the U.S. This strategy yields estimates that are very different from those obtained by comparing individuals schooled in different tracks. In particular, Figlio and Page (2002) find no evidence that tracking harms lower-achievement students. Second, Zimmer (2003), also using U.S. data, finds quasi-experimental evidence that the positive effects of achievement-specific instruction associated with tracking overcome the negative peer effects for students in lower-achievement tracks. Finally, Lefgren (2004) find that, in Chicago public schools, the difference between the achievement of low and high achieving students is no greater in schools that track that in school that do not.

This paper is also related to a large literature that investigates peer effects in the classroom (e.g., Hoxby, 2000; Zimmerman, 2003; Angrist and Lang, 2004). While this literature has, mainly for data reasons, focused mostly on the direct effect of peers, there are a few exceptions, and these have 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 find that students seem to benefit mainly from having homogeneous peers, which they attribute to indirect effects through teaching practices. 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. Finally, 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, suggesting that the level of teaching may be too high for them.

It is impossible to know if the results of this study will generalize until further studies are conducted in different contexts, but it seems likely that the general principle will hold: it will be difficult to assess the impact of tracking based solely on small random variations in peer composition that are unlikely to generate big changes in teacher behavior. Our model suggests that tracking may be particularly beneficial for low-achieving students when teachers' incentives are to focus on students who are above median achievement levels. Education systems are typically complex, having reward functions for schools and teachers that generate various threshold effects at different test

score levels. But virtually all developing countries teachers have incentives to focus on the strongest students. This suggests that our estimate of large positive impacts of tracking would be particularly likely to generalize to those contexts. This situation also seems to often be the norm in developed countries, with a few exceptions, such the No Child Left Behind program in the U.S.

The remainder of this paper proceeds as follows: Section 2 provides background on the Kenyan education system and presents a model nesting various mechanisms through which tracking could affect learning. Section 3 describes the study design, data, and estimation strategy. Section 4 presents the main results on test scores. Section 5 presents additional evidence on the impact of tracking on teacher behavior. Section 6 concludes and discusses policy implications.

2. Model

We consider a model that nests several different possible channels through which tracking students into two streams (a lower track and an upper track) could affect students' outcomes. In particular, the model allows peers to generate both direct student-to-student spillovers as well as to indirectly affect both the overall level of teacher effort and teachers' choice of the level at which to target instruction.[1] However, the model also allows for either of these channels to be shut off. Within the subset of cases in which the teacher behavior matters, we will consider the case in which teachers' payoffs are convex, linear, or concave in student test scores.

Suppose that educational outcomes for student *i* in class *j*, y_{ij}, are given by:

$$y_{ij} = x_{ij} + f(\overline{x}_{-ij}) + g(e_j)h(x_j^* - x_{ij}) + u_{ij}$$

where x_{ij} is the student's pretest score, $\overline{x}_{-ij}$ is the average scores of other students in the class, e_j is teacher effort, x^* is the target level to which the teacher orients instruction, and u_{ij} represents other i.i.d. stochastic student and class-specific factors that are symmetric and single-peaked. In this equation, $f(\overline{x}_{-ij})$ reflects the direct effect of a student's peers on learning, e.g. through peer-to-peer interactions. For simplicity of exposition, in what follows we remove the class indices.

[1] Epple, Newton and Romano (2002) consider the equilibrium implications of tracking in public school in a model where the indirect effect of peer through teacher effort is shut off, but private school can chose whether or not to track, and students can chose which school to attend.

We will focus on the case when *h* is a decreasing function of the absolute value of the difference between the student's initial score and the target teaching level, and is zero when $x_i - x^* > \theta$, although we also consider the possibility that *h* is a constant, shutting down this part of the model.

The teacher chooses x^* and e^* to maximize a payoff function *P* of the distribution of children's endline achievement minus the cost of effort $c(e)$ where $c(.)$ is a convex function. We assume that the marginal cost to teachers of increasing effort eventually becomes arbitrarily high as teacher effort approaches some level $\bar{e}$. We will also consider the case in which the cost of effort is zero below $\bar{e}$, so teachers always choose effort $\bar{e}$ and this part of the model shuts down. We will consider two kinds of teachers: civil servants, and contract teachers hired to teach the new sections in the ETP program. Contract teachers have higher-powered incentives than civil servants and, as shown in Duflo, Dupas and Kremer (2009) put in considerably more effort. In particular, we will assume that the reward to contract teachers from any increment in test scores equals λ times the reward to civil service teachers from the same increment in test scores, where λ is considerably greater than 1.

The choice of x^* will depend on the distribution of pre-test scores.[2] We assume that within each school the distribution of initial test scores is continuous, quasi-concave, and symmetric around the median. This appears to be consistent with our data (see Figure 1).

With convexity of teachers' payoffs in both student test scores and teacher effort in general, there could be multiple local maxima for teachers' choice of effort and x^*. Nonetheless, it is possible to characterize the solution, at least under certain conditions. Our first proposition states a testable implication of the special case where peers only affect each other directly.

[2] We rule out the possibility that teachers divide their time between teaching different parts of the class. In this case, tracking could reduce the number of levels at which a teacher would need to teach and thus increase the proportion of time students benefited from instruction. If teachers face some fixed cost in changing the focus teaching level, x^*, they will then optimally use some type of Ss process to adjust x^*. In this case, more teachers will change x^* in response to large changes in the composition of student body associated with tracking than in response to small changes associated with random fluctuations in class composition. As discussed below, we think the evidence is consistent with the hypothesis that some teachers change their teaching techniques even in response to random fluctuations in class composition. Fixed costs of changing x^* may not be that great because this change may simply mean proceeding through the same material more slowly or more quickly.

Proposition 1: Consider a special case of the model in which teachers do not respond to class composition because *h( )* is a constant and either *g( )* is a constant or the cost of effort is zero below $\bar{e}$. In that case, tracking will not change average test scores but will reduce test scores for those below the median of the original distribution and increase test scores for those above the median.

Proof: Under tracking, average peer achievement is as high as possible for students above the median and as low a possible for students below the median. ■

Note that this proposition would be true even with a more general equation for test scores that allowed for interactions between students own test scores and those of their peers, as long as students always benefit from higher achieving peers.

Proposition 2: If teacher payoffs, *P*, are convex in post-test scores, in a non-tracked class the target teaching level, x^*, must be above the median of the distribution. If teacher payoffs are linear in post-test scores, then x^* will be equal to the median of the distribution. If teacher payoffs are concave in post-test scores, then x^* will be below the median of the distribution.

Proof: Consider first the convex case. Since the distribution is assumed to be symmetric and quasi-concave, the peak of the distribution must be at the median. To see that x^* must be above the median, suppose that x^* were less than the median. Denote the distance between x^* and the median as D. Now consider an alternative x^*, denoted $x^{*\prime}$, equal to the median plus D. By symmetry of the distribution, the total number of students at any distance from $x^{*\prime}$ equals the total number of students at any distance from x^*. However, the distribution of students within range θ of x'^* first order stochastically dominates the distribution of students within a range θ of x^*. Thus, by convexity of the *P* function the teacher would be better off with the target teaching level $x^{*\prime}$.

To complete the proof for the convex case it is simply necessary to show that the teacher will not choose x^* equal to the median of the distribution. To see this, note that since the distribution is continuous, increasing x^* slightly from the median will lead to a second order reduction in the number of pupils at any distance from the target teaching level but to a first order increase in their test score and thus in the *P* function.

Arguments for the linear and concave case are analogous. ■

Proposition 3:

- If f( ) is increasing in peer test scores, then a uniform increase in peer baseline achievement will raise test scores for any students those with $x > x^*$, and the effect will be the largest for students with $x > x^*$, but $x^* + \theta$; have an ambiguous effects on test scores for students with scores between $x^* - \theta$ and x^*; and increase test scores for students with test scores below $x^* - \theta$, although the increase will be smaller than that for students with test scores greater than x^*.
- If f( ) is a constant, so there is no direct effect of peers, then a uniform increase in peer achievement will cause students with $x > x^*$ to have higher test scores and those with x between $x^* - \theta$ and x^* to have lower scores. There will be no change in scores for those with $x < x^* - \theta$.

Proof: Consider first the case in which f( ) is increasing in peer test scores. A uniform increase in peer baseline achievement will lead to an increase in the focus teaching level. Students with $x > x^*$ and $x < x^* + \theta$ will be closer to the target teaching level. They will thus benefit not only from the direct impact of higher-achieving peers but also from the indirect impact on teachers' choice of target instruction level. Students whose initial test scores were above $x^* + \theta$ are still too far from the target level of instruction, but still benefit from the increase in test scores (note that in the case where the teacher reward is a convex function of student test scores, there may not be any student above $x^* + \theta$, as x^* may have been chosen to be within θ of the top of the distribution).

Students with scores between $x^* - \theta$ and x^* benefit from the higher achievement of their peers and from any increase in teacher effort associated with the higher peer achievement. On the other hand, these students now are further away from the new target teaching level. The overall effect is ambiguous.

Students with scores less than $x^* - \theta$ were not in range of the teacher's instruction prior to the increase in test scores, and are not advantaged or disadvantaged by the change in the target teaching level. However, they benefit from the higher-achievement of their peers. If f( ) is not increasing in test scores (no direct peer effects), the proof follows from the discussion of the indirect effects. ■

Proposition 4: Let x^*_L denote the target teaching level in the lower section in a tracking school and x^*_U denote the target level in the upper section. If payoffs are convex, x^*_L will be within distance θ of x_m, where x_m denotes the 50th percentile of the original distribution. If payoffs are concave, x^*_U will be within distance θ of x_m. If payoffs are linear, both x^*_U and x^*_L will be within distance θ of x_m.

Proof: To see this for the convex case, suppose that $x^*_L < x_m - \theta$. Increasing x^* would both increase the number of students at any distance from x^* and the base score x_i of students at any distance from x^*. Thus it would be preferred. Proofs for the other cases are analogous. ■

Proposition 5: Denote the distance between x_m and the target teaching level in the upper section x^*_U as D_U and denote the corresponding distance between x^*_L and x_m as D_L. If payoffs are convex and the third derivative is non-negative, then $D_U > D_L$, so the median student is closer to the target teaching level in the lower track. If payoffs are linear in student scores then $D_U = D_L$. If teacher payoffs are concave in student test scores and the third derivative is non-positive, then $D_U < D_L$.

Proof: Consider first the case of convex payoffs. Suppose that $D_U = D_L$. In that case, both the teacher teaching the lower track and the teacher teaching the upper track would have the same number of students within any distance, by the symmetry of the original distribution.

The first order necessary condition for an optimum is that increasing x^* marginally reduces the contribution to the *P* function from students to the left of x^* by the same amount it increases the contribution to the *P* function from students to the right of x^*. To see this necessary condition cannot be satisfied simultaneously for both the low achievement class and high achievement class if the target teaching levels in each class are symmetric around the median, note that if x^*_L is within distance θ of x_m and x^*_U is the same distance away from x_m then by quasi-convexity increasing x^*_U will decrease the total number of students at any distance D, whereas marginally increasing x^*_L will increase the total number of students within any distance by the same amount, again by symmetry. Thus increases in x^* will generate relatively more gains for the *P* function to the right of x^* compared to losses on the left in the low achieving class than in the high-achieving class as long as the degree of convexity is non-increasing.

Arguments are analogous for the linear and concave cases. Under linearity, the median student will be equidistant from the target teaching level in the lower and upper section. Under concavity, they will be closer in the top section.■

All of this implies that the median student will be closer to the target teaching level in the bottom of the distribution than in the top of the distribution.

Proposition 6: Teacher effort will be greater in the upper than in the lower section under convexity, equal under linearity, and lesser under concavity. However, for high enough λ, the difference between effort levels of contract teachers assigned to the high- and low-achievement sections will become arbitrarily small.

Proof: Consider the convex case first. Teachers will choose x^* so as to maximize the gain to their *P* function. Take x^*_L as fixed. The teacher of the upper section has the option of choosing a target teaching level x^*_u that is symmetric around x_m. If the teacher does this, then the teacher in the top section will have the exact same number of students at any distance as the teacher in the bottom section. In this case, the marginal gain to the *P* function associated with an increment in their test scores will be greater for the teacher in the upper track by the convexity of the *P* function.

As shown in Proposition 5, under a convex payoff function, the teacher of the top section will in fact choose x^* greater than $x_m + (x_m - x_L)$ and thus will have fewer children within any distance than the teacher of the lower section. However, by revealed preference, the marginal increase in the payoff function from extra effort must be greater than the marginal gain to the *P* function from the impact of extra effort on increasing the scores of students within range of x^*_L. The proofs for the linear and concave cases follow a similar logic.

The second result (that for high enough λ, the difference between effort levels of contract teachers assigned to the high- and low-achievement classes will become arbitrarily small) is due to the assumption that the cost of effort becomes arbitrarily high as a maximum effort level $\bar{e}$ is approached. ■

Proposition 7. Under a linear teacher payoff function, a student initially at the median of the distribution will score higher if assigned to the upper section than the lower section under tracking.

Proof: Under linear teacher payoffs, a student at the median will experience equal teacher effort in the upper and lower sections, and will be equally far from the target teaching level. However, the student will have stronger peers in the top section. ■

Note that under convex teacher payoffs, the student at the median will experience higher teacher effort in the top section (compared to the bottom section) and will have stronger peers but will have teaching which is not as good a match for his or her initial

achievement. The model therefore offers no definitive prediction on whether the median student performs better in the upper or lower track. Similarly, if teacher payoffs, *P*, are concave in student test scores, then the student would have a more appropriate teaching target level but lower teacher effort in the top section.

This model thus nests, as special cases, models with only a direct effect of peers or only an effect going through teacher behavior. It also nests special cases in which teacher payoffs are linear, concave, or convex in students' test scores. Nevertheless, the model make some restrictive assumptions. In particular, teacher effort has the same impact on student test score gains anywhere in the distribution. In a richer model, teacher effort might have a different impact on test scores at different places along the distribution. Student effort might also respond endogenously to teacher effort and the target teaching level. In such a model, ultimate outcomes will be a composite function of teacher effort, teacher focus level, and student effort, which in turn would be a function of teacher effort and teaching level. In this case, we conjecture that the results would go through as long as the curvature assumptions on the payoff function were replaced by curvature assumptions on the resulting composite function for payoffs. Multiplicative separability of e and x* is important to the results, however.

Propositions 1, 2 and 4 provide empirical implications that can be used to test whether the data is consistent with the different special cases.

Below we argue that the data are inconsistent with the special case with no teacher response, the special case with no direct effects of peers, and the special case in which teacher payoffs are linear or concave in students' scores. However, our results are consistent with a model in which both direct and indirect effects operate and teachers' payoffs are convex with student test scores, which is consistent with our description of the education system in Kenya.

Note that this model has no clear prediction for the effect of the variance of initial achievement on test scores in an untracked class or for the interaction between the effect of tracking and the initial variance of the distribution.[3]

[3] To see that changes in the distribution of initial scores that increase variance of these scores could reduce average test scores and the effect of tracking, consider an increase in dispersion so no two students are within distance θ of each other. Then teachers can never teach more than one pupil. Average test scores will be low, and tracking will have no impact on x^* or teacher effort. To see that changes in the distribution

3. The Tracking Experiment: Background, Experimental Design, Data, and Estimation Strategy

3.1. Background: Primary Education in Kenya

Like many other countries, Kenya has a centralized education system with a single national curriculum and national exams. Glewwe, Kremer, and Moulin (2009) show that textbooks based on the curriculum benefited only the initially higher-achieving students, suggesting that the exams and associated curriculum are not well-suited to the typical student.

Most primary-school teachers are hired centrally through the civil service and they face weak incentives. As we show in Section 5, absence rates among civil-service teachers are high. In addition, some teachers are hired on short-term contracts by local school committees, most of whose members are elected by parents. These contract teachers typically have much stronger incentives, partly because they do not have civil-service and union protection but also because a good track record as a contract teacher can help them obtain a civil-service job.

To the extent that schools and teachers face incentives, the incentives are largely based on their students' scores on the primary school exit exam. Many students repeat grades or drop out before they can take the exam, and so the teachers have limited incentives to focus on students who are not likely to ever take the exam. Extrinsic incentives are thus stronger at the top of the distribution than the bottom. For many teachers, the intrinsic rewards of teaching to the top of the class are also likely to be greater than those of teaching to the bottom of the class, as such students are more similar to themselves and teachers are likely to interact more with their families and with the students themselves in the future.

---

that increase variance could increase the impact of tracking, consider moving from a degenerate distribution, with all the mass concentrated at a single point, to a distribution with some dispersion. Tracking will have no effect on test scores with a degenerate distribution, but will increase average scores with tracking. Increases in dispersion could also increase average test scores in the absence of tracking. To see this, suppose teacher payoffs are very convex, so teachers focus on the strongest student in the class. Suppose also that the highest achieving student's initial score exceeds that of the second highest-scoring pupil by more than θ. Consider a move from the initial distribution to a distribution with the same support, but in which some students were pushed to the boundaries of this support. More students will be within range of the teacher and hence teacher effort and average test scores will rise.

Until recently, families had to pay for primary school. Students from the poorest families often had trouble attending school and dropped out early. But recently, Kenya has, like several other countries, abolished school fees. This led to a large enrollment increase and to greater heterogeneity in student preparation. Many of the new students are first generation learners and have not attended preschools (which are neither free nor compulsory). Students thus differ vastly in age, school preparedness, and support at home.

3.2. Experimental Design

This study was conducted within the context of a primary school class-size reduction experiment in Western Province, Kenya. Under the Extra-Teacher Program (ETP), with funding from the World Bank, ICS Africa provided 140 schools with funds to hire an additional first-grade teacher on a contractual basis starting in May 2005, the beginning of the second term of that school year.[4] The program was designed to allow schools to add an additional section in first grade. 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. Duflo, Dupas and Kremer (2009) reports on the effect of the class size reduction and teacher contracts.

We examine the impact of tracking and peer effects using two different versions of the ETP experiment. In 61 schools randomly selected (using a random number generator) from the 121 schools that originally had only one grade 1 section, grade 1 pupils were randomly assigned to one of two sections. We call these schools the "non-tracking schools." In the remaining 60 schools (the "tracking schools"), children were assigned to sections based on scores on exams administered by the school during the first term of the 2005 school year. In the tracking schools, students in the lower half of the distribution of baseline exam scores were assigned to one section and those in the upper half were assigned to another 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.[5] Thus, in the analysis that

[4] 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.

[5] In these schools, the sections that were taught by civil service teachers rather than contract teachers sometimes recombined or exchanged students.

follows, we focus on the 121 schools that initially had a single grade 1 section and exclude 19 schools (10 tracking, 9 non-tracking schools) that initially had two or more.[6]

After students were assigned to sections, the contract teacher and the civil-service teacher were randomly assigned to sections. Parents could request that their children be reassigned, but this only occurred in a handful of cases. The main source of noncompliance with the initial assignment was teacher absenteeism, which sometimes led the two grade 1 sections to be combined. On average across five unannounced school visits to each school, we found the two sections combined 14.4% of the time in non-tracking schools and 9.7% of time in tracking schools (note that the likelihood that sections are combined depends on teacher effort, itself an endogenous outcome, as we show below in Section 5). When sections were not combined, 92% of students in non-tracking schools and 96% of students in tracking schools were found in their assigned section. The analysis below is based on the initial assignment regardless of which section the student eventually joined.

The program lasted for 18 months, which included the last two terms of 2005 and the entire 2006 school year. 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. The fraction of students who repeated grade 1 and thus participated in the program for only the first year was 23% in non-tracking schools and 21% in tracking schools (the p-value of the difference is 0.17).[7]

Table 1 presents summary statistics for the 121 schools in our sample. As would be expected given the random assignment, tracking and non-tracking schools look very similar. Since tests administered within schools prior to the program are not comparable across schools, they are normalized such that the mean score in each school is zero and the standard deviation is one. Figure 2 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 non-normalized peer test scores are not correlated with the student's

[6] 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 sections.

[7] Students enrolled in grade 2 in 2005 and who repeated grade 2 in 2006 were randomly assigned to either the contract teacher or the civil-service teacher in 2006. All the analysis is based on the initial assignment, so they are excluded from the study and excluded from the measures of peer composition at endline. Students who repeated grade 1 in 2006 remain in the data set and are included in the measures of peer composition at endline. 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.

own test score in non-tracking schools but, consistent with the discontinuous assignment at the 50th percentile for most schools, there is sharp discontinuity at the 50th percentile in tracking schools.[8] The baseline exams are a good measure of academic achievement, in that they are strongly predictive of the endline test we administered, with a correlation of 0.47 in the non-tracking schools and 0.49 in tracking schools. In tracking schools, the top section has somewhat more girls and students are 0.4 years older.

3.3 Data

The sample frame consists of approximately 10,000 students enrolled in first grade in March 2005. The key outcome of interest is student academic achievement, as measured by 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 enumerators. 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 generated before enumerators visited the school) after stratifying by their position in the initial distribution. Part of the test was designed by a cognitive psychologist to measure a range of skills students might have mastered at the end of grade 2. Part of the test was written and part was orally administered one-to-one by trained enumerators. 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 those children, however, and the attrition rate on the test was 18 percent. 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). Transfer rates to other schools were similar in

[8] Peer quality is slightly more similar for children below and above 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.

tracking and non-tracking schools. In total, we have endline test score data for 5,796 students.

To measure whether program effects persisted, children sampled for the endline were tested again 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 students had reached grade 3, but repeaters were also tested. The attrition for this longer-term follow-up was 22 percent, only 4 points higher than attrition at the endline test. The proportion of attritors and their characteristics do not differ between the two treatment arms (appendix table 1).

We also collected data on grade progression and dropout rates, and student and teacher absence. Overall, the dropout rate among grade 1 students in our sample was low (below 0.5 percent). Several times during the course of the study, enumerators went to the schools unannounced and checked, upon arrival, whether teachers were present in school and whether they were in class and teaching. On those visits, enumerators also took a roll call of the students.

3.4 Empirical Strategy

a) Measuring the Impact of Tracking

To measure the overall impact of tracking on test scores, we run regressions of the form:

(E1) $$y_{ij} = \alpha T_j + X_{ij}\beta + \epsilon_{ij}$$

where y_{ij} is the endline test score of student i in school j (expressed in standard deviations of the distribution of scores in the non-tracking schools),[9] T_j is a dummy equal to 1 if school j was tracking, and X_{ij} is a vector including a constant and child and school control variables (we estimate 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 a contract or civil-service teacher).

To identify potential differential effects for children assigned to the lower and upper section, we also run:

(E2) $$y_{ij} = \alpha T_j + \gamma T_j * B_{ij} + X_{ij}\beta + \epsilon_{ij}$$

[9] We have also experimented with an alternative specification of the endline 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.

where B_{ij} is a dummy variable that indicates whether the child was in the bottom half of the baseline score distribution in her school (B_{ij} is also included X_{ij}). We also estimate a specification where treatment is interacted 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 endline test scores on baseline test scores in tracking and non-tracking schools, and plot the results.

To understand better how tracking works, we also 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 skills. Finally, we also run regressions of a similar form, using as outcome variable teacher presence in school, whether the teacher is in class teaching, and student presence in school.

b) Non-tracking schools

Since children were randomly assigned to a section in these schools, their peer group is randomly assigned and there is some naturally occurring variation in the composition of the groups.[10] 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 excluding the student him or herself):

(E3) $$y_{ij} = \kappa \overline{x}_{-ij} + X_{ij}\beta + \nu_j + \epsilon_{ij}$$

where $\overline{x}_{-ij}$ is the average peer baseline test score in the section to which a student was assigned.[11] The vector of control variables X_{ij} includes the student's own baseline score x_{ij}. Since students were randomly assigned within schools, our estimate of the coefficient of $\overline{x}_{-ij}$ in a specification including school fixed effects will reflect the causal effect of peers' prior achievement (both direct through peer to peer learning, and indirect through adjustment in teacher behavior to the extent to which teachers change behavior in response to small random variations in class composition). Although our model has no specific prediction on the impact of the variance, we also include the variance of the peers' test scores, as an independent variable in one specification.

[10] On average across schools, the difference in baseline scores between the two sections is 0.17 standard deviation, with a standard deviation of 0.13. The 25th-75th percentiles interval for the difference is [0.7 - 0.24].

[11] 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).

The baseline grades are not comparable across schools (they are the grades assigned by the teachers in each school). However, baseline grades are strongly correlated with endline test scores, which are comparable across schools. Thus, to facilitate comparison with the literature and with the regression discontinuity estimates for the tracking schools, we estimate the impact of average endline peer test scores on a child's test score:

(E4) $$y_{ij} = \kappa \overline{y}_{-ij} + X_{ij}\beta + \nu_j + \epsilon_{ij}$$

This equation is estimated by instrumental variables, using $\overline{x}_{-ij}$ as an instrument for $\overline{y}_{-ij}$.

c) Measuring the Impact of Assignment to Lower or Upper Section

Tracking schools provide a natural setup for a regression discontinuity (RD) design to test whether students at the median are better off being assigned to the top section, as would be true in the special case of the model in which teacher payoffs were linear in test scores.

As shown in Figure 2, students on either side of the median were assigned to classes with very different average prior achievement of their classmates: the lower-scoring member was assigned to the bottom section, and the higher-scoring member 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:

(E5) $$y_{ij} = \delta B_{ij} + \lambda_1 P_{ij} + \lambda_2 P_{ij}^2 + \lambda_3 P_{ij}^3 + X_{ij}\beta + \epsilon_{ij}$$

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

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

(E6) $$y_{ij} = \delta B_{ij} + \lambda_1 P_{ij} + \lambda_2 P_{ij}^2 + \lambda_3 P_{ij}^3 + X_{ij}\beta + \epsilon_{ij} + \nu_j$$

To test the robustness of our estimates to various specifications of the control function, we also run specifications similar to equations (E5) and (E6), estimating the polynomial separately on each side of the discontinuity, and report the difference in test scores across the discontinuity. Finally, we follow Imbens and Lemieux (2007) and use a Fan locally weighted regression of the relationship between endline test scores and baseline percentile on both sides of the discontinuity.

Note that this is an unusually favorable setup for a regression discontinuity design. There are 60 different discontinuities in our data set, rather than just one, as in most regression discontinuity applications, and the number of different discontinuities in principle grows with the number of schools.[12] We can therefore run a specification including only the pair of students straddling the median.

(E7) $$y_{ij} = \delta B_{ij} + X_{ij}\beta + \epsilon_{ij} + \nu_j$$

Since the median will be at different achievement levels in different schools, results will be robust to sharp non-linearities in the function linking pre- and post-test achievement.

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 achievement for the median child in a tracking environment. Specifically, the first stage of this regression is:

$$\overline{y}_{-ij} = \pi B_{ij} + \phi_1 P_{ij} + \phi_2 P_{ij}^2 + \phi_3 P_{ij}^3 + X_{ij}\beta + \epsilon_{ij} + \nu_j$$

where $\overline{y}_{-ij}$ is the average endline test scores of the classmates of student *i* in school *j*. The structural equation:

(E8) $$y_{ij} = \kappa \overline{y}_{-ij} + \lambda_1 P_{ij} + \lambda_2 P_{ij}^2 + \lambda_3 P_{ij}^3 + X_{ij}\beta + \epsilon_{ij} + \nu_j$$

is estimated using B_{ij} (whether a child was assigned to the bottom track) as an instrument for $\overline{y}_{-ij}$.

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 also means that the child is the lowest achieving child of his section (at least at baseline) and having low-achieving peers means that the child is the highest achieving child of his track.

4. Results

In Section 4.1, we present reduced form estimates of the impact of tracking, showing that tracking increased test scores throughout the distribution and thus rejecting the special case of the model in which higher-achieving peers raise test scores directly but there is no indirect effect through changing teacher behavior. In Section 4.2, we use random variation in peer composition in non-tracked schools to assess the implications of

[12] 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.

Proposition 3, and to argue that the data is not consistent with the special case of the model in which there are no direct effects of peers. In Section 4.3, we argue that the data are inconsistent with the special case of the model in which teacher incentives are linear in student test scores, because the median student in tracking schools scores similarly whether assigned to the upper or lower section. We conclude that the data is most consistent with a model in which peer composition affects students both directly and indirectly, through teacher behavior, and in which teachers face convex incentives. In this model, teachers teach to the top of the distribution in the absence of tracking, and teaching can improve learning for all children.

4.1 The Impact of Tracking by Prior Achievement and the Indirect Impact of Peers on Teacher Behavior

A striking result of this experiment is that tracking by initial achievement significantly increased test scores throughout the distribution.

Table 2 presents the main results on the impacts of tracking. At the endline test, after 18 months of treatment, 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 estimated effect is somewhat larger (0.175 standard deviations, with a standard error of 0.077 standard deviations) when controlling for individual-level covariates (column 2). Both sets of students, those assigned to the upper track and those assigned to the lower track, 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). When we look at each quartile of the initial distribution separately, we find positive point estimates for all quartiles (column 4).

Figure 3 provides graphical evidence suggesting that all students benefited from tracking. As in Lee (2008), it plots a student's endline 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. Both in language and math, tracking increases test scores regardless of the child's initial test score in the distribution of test scores.

Overall, the estimated effect of tracking is relatively large. It is similar in magnitude to the effect of being assigned to a contract teacher (shown in Row 6 of Table 4), who, as

we will show in Table 6, exerted much higher levels of effort than civil-service teachers. It is also interesting to contrast the effect of tracking with that of a more commonly proposed reform, class size reduction. In other contexts, studies have found a positive and significant effect of class size reduction on test scores (Angrist and Lavy, 1999; Krueger and Whitmore, 2002). In Duflo, Dupas and Kremer (2009), however, we find that in the same exact context, class size reduction *per se* (without a change in teachers' incentive) generates an increase in test scores of 0.09 standard deviation after 18 months (though insignificant), but the effect completely disappears within one year after the class size reduction stops.

The program effect persisted beyond the duration of the program. When the program ended after 18 months, three quarters of students had then reached grade 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 is slightly larger (0.178 standard deviations) and more significant with control variables (column 2, panel B), and the gains persist both for initially high and low achieving children. A year after the end of the program, the effect for the bottom half is still large (0.135 standard deviations, with a p-value of 0.09), although the effect for students in the bottom quartile is insignificant (Panel B, column 4).

This overall persistence 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 skills in grades 1 and 2 and that this may have helped them learn more later on.[13]

[13] We also find (in results not reported here to save space) that initially low-achieving girls 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 of 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 schools to be 0.16 standard deviations on average. At most, the repetition effect would therefore explain an increase in 0.04*0.16=0.0064 standard deviations in test scores. Furthermore, it is present only for girls, while tracking affects both boys and girls.

Under Proposition 1, this evidence of gains throughout the distribution is inconsistent with the special case of the model in which pupils do not affect each other indirectly through teacher behavior but only directly, with all pupils benefiting from higher scoring classmates.

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. 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.

Panel B present differential effects 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.

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, while contract teachers exert high effort in all situations. This is consistent with the idea that the cost of effort rises very steeply as a certain effort level is approached. Contract teachers are close to this level of effort in any case, and therefore have little scope to increase their effort, while civil service teachers have more such scope.

4.2 Random Variation in Peer Composition and the Direct Effect of Peers

The local random variation in peer quality in non-tracking schools helps us test whether the opposite special case in which peers affect each other only indirectly, through their impact on teacher behavior, but not directly, can also be rejected.

Recall that Proposition 3 implies that the impact of a uniform increase in peer achievement on students at different level of the distribution depends on whether or not

there are direct peer effects. Namely, a uniform increase in peer achievement increases test scores at the top of the distribution in all cases, but effects on students in the middle and at the bottom of the distribution depend on whether there are also direct, positive effects of high achieving peers. In the presence of such effects, the impact on students in the middle of the distribution is ambiguous, while for those at the bottom it is positive, albeit weaker than the effects at the top of the distribution. In the absence of such direct effects, there is a negative impact on students in the middle of the distribution and no impact at the bottom.

The random allocation of students between the two sections in non-tracking schools generated substantial random variation which allows us to test those implications: on average across schools, to assess these implications the difference in baseline scores between the two classes is 0.17 standard deviations, with a standard deviation of 0.14, and the 25^{th}-75^{th} percentiles interval for the difference is [0.7 - 0.24]. [14] 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 4. Similar approaches are proposed by Boozer et al. (2001) in the context of the STAR experiment and Lyle (2007) for West Point Cadets, who are randomly assigned to a group of peers.

On average students benefit from stronger peers: the coefficient on the average baseline test score is 0.35 with a standard error of 0.15 (Table 4 panel A, column 1). 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 endline score (the first stage is presented in panel C). If effects were linear, column 1 would imply that one standard deviation increase in average peer endline test score would increase the test score of a student by 0.52 standard deviations, an effect comparable to that usually found in previous literature, with the exception of Lyle (2007), which finds insignificant peer effects with a similar strategy.[15]

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

[15] Of course, these estimates come from variations in peer test scores that are smaller than one standard deviation and, the extrapolation to one standard deviation may not actually be legitimate: the linear

More interestingly, as shown in columns 6 to 8, the data are consistent with Proposition 3 in the presence of direct peer effects – the estimated effect is 0.9 standard deviations in the top quartile; insignificant and negative in the middle two quartiles, and 0.5 standard deviations in the bottom quartile. The data thus suggest that peers affect each other both directly and indirectly.[16]

4.3 Are Teacher Incentives Linear? The Impact of Assignment to Lower vs. Upper Section: Regression Discontinuity Estimates for Students near the Median

Recall from proposition 7 that under a linear payoff schedule for teachers, the median student will be equidistant from the target teaching level in the upper and lower sections, but will have higher-achieving peers and therefore perform better in the upper section. Under a concave payoff schedule, teacher effort will be greater in the lower section but the median student will be better matched to the target teaching level in the upper section, potentially creating offsetting effects. Finally, if teacher payoffs are convex in student test scores, the median student will be closer to the target teaching level in the lower section but on the other hand will have lower-achieving peers and experience lower teacher effort. These effects go in opposite directions, so that the resulting impact of the section in which the median child is assigned is ambiguous. In this section, we present regression discontinuity estimates of the impact of assignment to the lower or upper section for students near the median in tracking schools. We argue that the test score data are inconsistent with linear payoffs but consistent with the possibility that teachers face a convex payoff function and focus on students at the top of the distribution. (Later, we rule out the concave case.)

The main thrust of the regression discontinuity estimates of peer effects are shown in Figure 3, discussed above. As is apparent from the figure, there is no discontinuity in test scores at the 50th percentile cutoff in the tracking schools, despite the strong discontinuity

---

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. Note that even with this normalization, the results are not quite comparable to those of paper who estimate the effect of a standard deviation in average baseline test scores on endline test scores: those results would be scaled down, relative to the ones we present here, by the size of the relationship between baseline and endline scores.

[16] Controlling for the standard deviation of the test scores (column 2) does not change the estimated effect of the mean, and we cannot reject the hypothesis that the standard deviation of scores itself has no effect, though the standard errors are large.

in peer baseline scores observed in Figure 2 (a difference of 1.6 standard deviations in the baseline scores). The relationship is continuous and smooth throughout the distribution.[17]

A variety of regression specifications show no significant effect of students near the median of the distribution being assigned to the bottom half of the class in tracking schools (Table 5, panel A). Columns 1 and 2 present estimates of equations (E5) and (E6), respectively: the endline 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 (E7), and column 7 adds a school fixed effect. To assess the robustness of these results, columns 3 through 5 specify the control function in the regression discontinuity design estimates in two other ways: column 5 follows Imbens and Lemieux (2007) and shows a Fan locally weighted regression on each side of the discontinuity.[18] The specifications in columns 3 and 4 are similar to equations (E5) and (E6), 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, the marginal students' final test scores do not seem to be significantly affected by assignment to the bottom section.

Panel B shows instrumental variable estimates of the impact of classmates' average test score. We use the average endline score of classmates (because the baseline scores are school specific), and instrument it using the dummy for being in the "bottom half" of the initial distribution. The first stage is shown in panel C, and shows that the average endline test scores of a child's classmates are 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 2, 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 0.002 standard deviations, a point estimate extremely close to zero. The 95 percent confidence interval in this specification is [-0.21; 0.21]. Thus, we are able to reject at 95 percent confidence

[17] This result is robust to a series of specifications. 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 Appendix Figure A1, we reproduce Figure 3 with a quadratic fit for total score in Panel A and also find no discontinuity. We use a Fan locally-weighted regression with a biweight kernel and a bandwidth of 2.0 in Panel B, and we again see no discontinuity at the threshold for being assigned to the bottom track.

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

reasonably modest overall effects of peer average test scores on the median child's test score in a tracking environment.[19]

Overall, these regression discontinuity results allow us to reject the third special case, in which teacher have linear incentives and consequently target the median child in the distribution of the class.

Taken together, the test scores results are consistent with a model in which students influence each other both directly and indirectly through teacher behavior, and teachers face convex payoffs in pupils' test scores, and thus tend to target their teaching to the top of the class. This model can help us interpret our main finding that tracking benefits all students: for higher-achieving students, tracking implies stronger peers and higher teacher effort, while for lower-achieving students, tracking implies a level of instruction that better matches their need. However, we have not yet rejected the possibility that teacher payoffs are concave in student test scores. Recall that under concavity, students in the bottom half of the distribution may gain from greater teacher effort under tracking (proposition 6). The next section examines data on teacher behavior, arguing that it is inconsistent with the hypothesis that teacher payoffs are concave in student test scores, but consistent with the hypothesis that payoffs are convex in student scores..

5. Teacher Response to Tracking

This section reports on tests of implications on the model related to teacher behavior. Subsection 5.1 argues that the evidence on teacher behavior is consistent with the idea that teachers face convex payoffs incentives in pupil test scores and inconsistent with the hypothesis of concavity. Subsection 5.2 presents some evidence that the patterns of changes in test scores are consistent with the hypothesis that teachers change their focus teaching level x^*, in response to tracking.

5.1 Teacher Effort and the Curvature of the Teacher Payoff Function

Estimates of the impact of tracking on teacher's effort are presented in Table 6. Our measure of teacher effort is whether the teacher was present in school during unannounced visits, and whether she was found in class and teaching.

[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 lower section is through classmates' average score.

Recall that the model does not yield a clear prediction for whether tracking should increase or decrease teacher effort overall. However, the model predicts that the effort level might vary across sections (upper or lower) under tracking. Namely, proposition 6 implies that if teacher payoffs are convex in student test scores, then teachers assigned to the top section in tracking schools should exert more effort than those assigned to the bottom section. On the other hand, if payoffs are concave in student test scores, teachers should put in more effort in the lower section than the upper section.

We find that teachers in tracking schools are significantly more likely both to be in school and to be in class teaching than those in non-tracking schools (Table 6, columns 1 and 2).[20] 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. However, the negative coefficient on the interaction term between "tracking" and "bottom half" shows that teacher effort in tracking schools is higher in the upper section than the lower sections, consistent with the hypothesis that teacher payoffs are convex in student test scores.

Recall that the model also suggests that if teachers face strong enough incentives (high enough λ) then the impact of tracking on their effort will be smaller because they have less scope to increase effort. To test this, we explore the impact of tracking on teacher effort separately for civil-service teachers and new contract teachers, who face very different incentives. Contract 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 high job security and promotion depends only weakly on performance. Civil service teachers thus may have more scope to increase effort.

We find that the 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. In contrast, the civil-service teachers are 5.4 percentage points more likely to be in school 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

[20] The specification is similar to equation (E2), though the set of control variables includes teacher age and experience teaching.

a non-tracked group). However, the difference disappears entirely for civil-service teachers assigned to the bottom section: the interaction between tracking and bottom section is minus 7.7 percentage points, and is also significant. The effect is even stronger for finding teachers in their classrooms: 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 on teacher effort also shed light on the differential impact of tracking across students observed in Table 3. Recall that among students who were assigned to civil service teachers, tracking created a larger test score increase in the top section than in the bottom section, but this was not the case for students of contract teachers. What the effort data shows is that, for students of civil service teachers, the tracking effect is larger for the upper stream because they benefit not only from (potentially) more appropriate teaching and better peers, but also from higher effort. For students of contract teachers, the "higher effort" margin is absent.

5.2 Adjustment in the level of teaching and effects on different skills

The model suggests teachers may adjust the level at which they teach in response to changes in class composition. For example, a teacher assigned students with low initial achievement might begin with more basic material and instruct at a slower pace, providing more repetition and reinforcement. 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. Teachers with a heterogeneous class may teach at a relatively high level that is inappropriate for most students, especially those at the bottom.

While we unfortunately do not have direct evidence on the material teachers covered, Table 7 reports specifications similar to equation (E2), but with test scores disaggregated by specific skill for math and language. The differential impact of tracking on strong and weak student's mastery of easy and hard material is consistent with the hypothesis that teachers adjusted their teaching to fit their classroom's composition. 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 skills, and negative for the hardest skills. A chi-square test allows us to reject equality of the coefficients of the interaction in the "easy skills" regression and the "difficult skills" regression at the 5 percent level. Conversely, students assigned to the upper section 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 provides suggestive evidence that tracking allowed teachers the opportunity to focus on the skills that children had not yet mastered, although the estimates are not very precise.[21] An alternative explanation for these results, however, is that weak students stood to gain from any program on the easiest skills (since they had not mastered them yet, and in 18 months they did not have time to master both easy and strong skills), while strong students had already mastered them and would have benefited from any program at the skills they had not already mastered. The ordinal nature of test score data makes regression interaction terms difficult to interpret definitively, which further weakens the evidence.

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

[21] We estimated a version of equation (E2) allowing the effect to vary by quarter of the distribution for each skill, and the patterns are very similar, with progressively weaker students benefiting the most from tracking for the easiest skills, and progressively stronger students benefiting the most for the hardest skills. We also estimated a version of equation (E2) separately by teacher type. We find that the effects observed in Table 7 are much stronger for students assigned to contract teachers than for those assigned to civil-service teachers. This is because lower section students assigned to the civil-service teachers did not benefit from tracking, as seen in Table 3.

to our knowledge this paper provides the first experimental evaluation of the impact of tracking in any context, and the only rigorous evidence in a developing country 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 distribution of student achievement. A regression discontinuity design approach reveals that students who were very close to the 50^{th} percentile of the initial distribution within their school scored similarly on the endline exam whether they were assigned to the top or bottom section. In each case, they did much better than their counterparts in non-tracked schools.

We also find that students in non-tracking schools scored higher if they were randomly assigned to peers with higher initial scores. This effect was very strong for students at the top of the distribution, absent for students in the middle of the distribution and positive but not as strong at the bottom of the distribution. Together, these results suggest that peers affect students both directly and indirectly by influencing teacher behavior, in particular teacher effort and choice of target teaching level. Under the model, the impact of tracking will depend on teachers' incentives, but in a context in which teachers have convex payoffs in student test scores, tracking can lead them to refocus attention closer to the median student.

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. Finally, our results match those of Zimmer (2003), who finds that, in the US, tracking has overall a positive effect on lower-achieving students, for whom the benefit of having more tailored instruction under tracking offsets the reduction in peer quality.

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.

Our results may 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. Furthermore, tracking in public school would affect the equilibrium under these programs. Epple, Newton and Romano (2002) study theoretically how tracking in public schools would affect the decision of private schools to track students, and the welfare of high and low achieving students. They find that, if the only effect of tracking was through the direct effects of the peer group, tracking in public schools would increase enrollment and raise average achievement in public schools, but that high achieving students would benefit at the expense of low achieving students. Our results suggest that, at least in some circumstances, tracking can potentially benefit all students, which would have implication for the school choice equilibrium in contexts with school choices.

Note that since teachers were randomly assigned to each section and class size was also constant, resources were similar for non-tracked classes and the lower and upper-sections under tracking. However, in other contexts, policy makers or school officials could target more resources to either the weaker or stronger students. Piketty (2004) notes that tracking could allow more resources to be devoted to weaker students, promoting catch up of weaker students. Compensatory policies of this type are not unusual in developed countries, but in some developed countries and almost all developing countries, more resources are devoted to stronger students, consistent with the assumption of convex payoffs to test scores in the theoretical framework above. Indeed, even in developed countries, the best teachers are often assigned to the stronger students.

If the best teachers are assigned to the highest achieving students, the initially lower achieving students could be hurt by tracking, so caution is needed in generalizing from these results in which teacher ability was held constant between tracking and non-

tracking schools.[22] Of course tendencies for strong teachers to seek high-achieving, students could perhaps be mitigated if evaluations of a teacher's performance were on a value-added basis, rather than based on endline scores.

It is an open question whether similar results would be obtained in different contexts. The model provides some evidence on features of the context that are likely to affect the impact of tracking: initial heterogeneity, high scope to increase teacher effort (at least through increase presence) and the relative incentives teachers face to teach low- and high-achieving students. For example, in a system where the incentive is to focus on the weakest students, and there is not much scope to adjust teacher effort, tracking could have very strong positive effect on high achievement students, and weak or even negative effect on weak students, who would lose strong peers without the benefit of getting more appropriately focused instruction. Going beyond the model, it seems reasonable to think that the impact of tracking might also depend on availability of extra resources to help teachers deal with different types of students (such as remedial education, teacher aides, lower pupil to teacher ratio, computer-assisted learning, and special education programs).

We believe that tracking might be reasonably likely to have a similar impact in other low income countries in sub-Saharan Africa and South Asia, where the student population is often heterogeneous, and the educational system rewards teachers for progress at the top of the distribution. Our reduced form results may not apply to the US or other developed countries where teachers' incentives may differ. However, we hope that our analysis may still provide useful insights to predict the situations in which tracking may or may not be beneficial in these countries, and on the type of experiments that would shed light on this question.

[22] 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 the civil-service teachers, who also happen to be the worst teachers, 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.

Angrist, Joshua andVictor Lavy (1999). "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." Quarterly Journal of Economics 114, 533-575.

Angrist, Joshua, and Kevin Lang (2004). "Does School Integration Generate Peer Effects? Evidence from Boston's Metco Program," *American Economic Review*, American Economic Association, vol. 94(5), pages 1613-1634

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.

Boozer, Michael, and Stephen Cacciola (2001). "Inside the 'Black Box' of Project Star: Estimation of Peer Effects Using Experimental Data" Yale Economic Growth Center Discussion Paper No. 832.

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

Duflo, Esther, Pascaline Dupas, and Michael Kremer (2009). "Inputs vs. Accountability: Experimental Evidence from Kenyan Primary Schools". Mimeo, MIT.

Epple, Dennis, Elisabeth Newlon and Richard Romano (2002). "Ability tracking, school competition, and the distribution of educational benefits," *Journal of Public Economics* 83:1-48.

Figlio, David and Marianne Page (2002). "School Choice and the Distributional Effects of Ability Tracking: Does Separation Increase Inequality?" *Journal of Urban Economics* 51: 497-514.

Glewwe, Paul W., Kremer, Michael and Moulin, Sylvie. (2009). "Many Children Left Behind? Textbooks and Test Scores in Kenya." *American Economic Journal: Applied Economics*, Vol. 1 (1): pp. 112-35.

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.

Krueger, Alan and Diane Whitmore (2002). "Would Smaller Classes Help Close the Black-White Achievement Gap?" In John E. Chubb and Tom Loveless, eds., *Bridging the Achievement Gap*. Washington: Brookings Institution Press.

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.

Lefgren, Lars (2004). "Educational peer effects and the Chicago public schools," *Journal of Urban Economics* 56: 169-191.

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.

Zimmer, Ron (2003). "A New Twist in the Educational Tracking Debate," *Economics of Education Review* 22: 307-315.

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: Distribution of Initial Test Scores

All schools

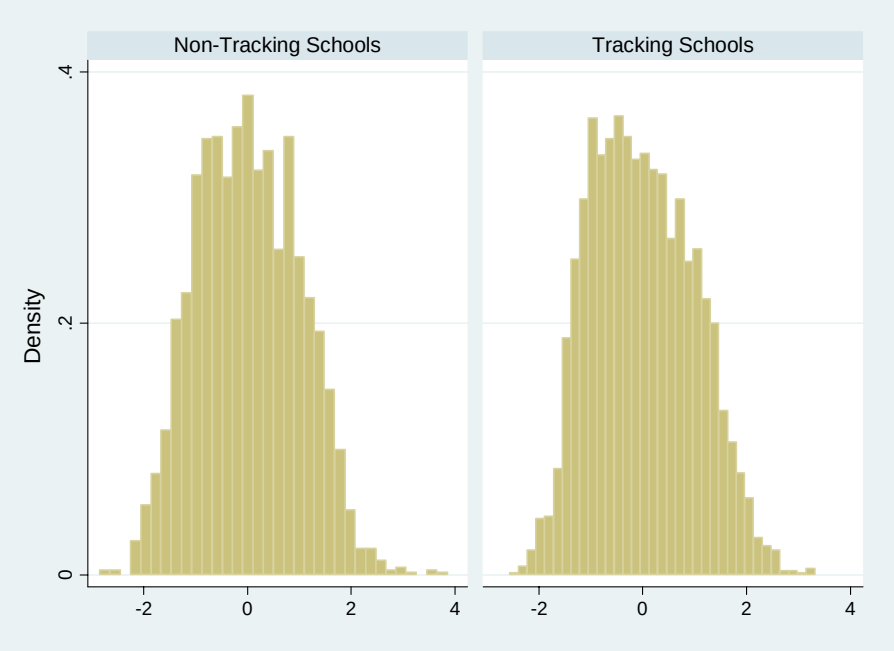
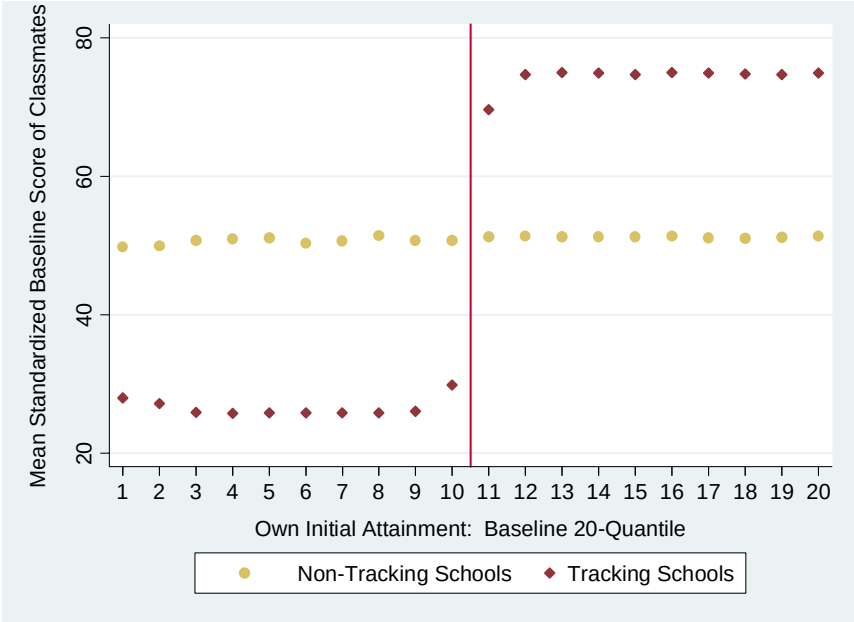


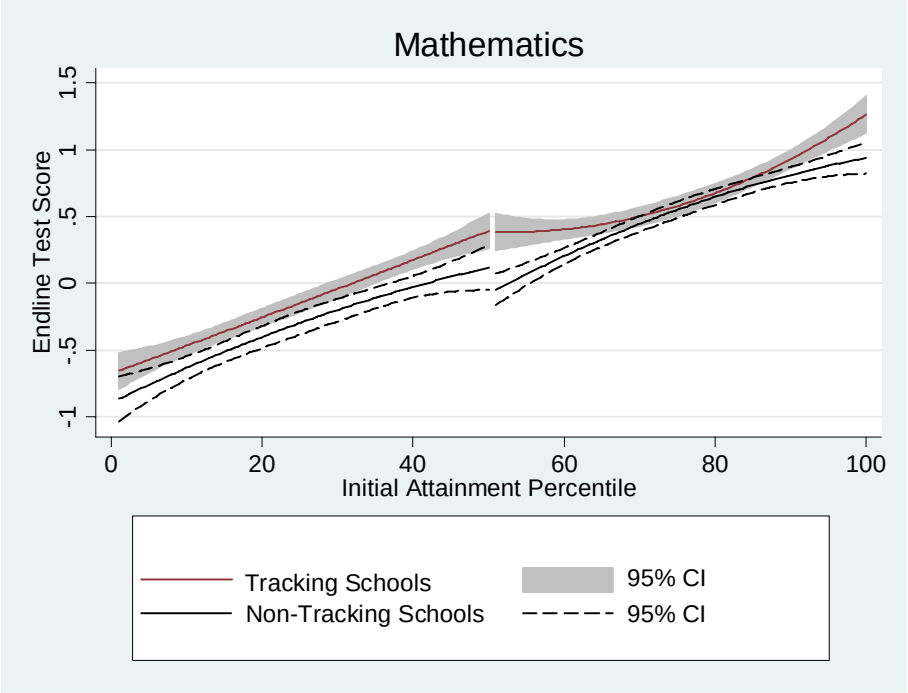
Figure 2: Experimental Variation in Peer Composition

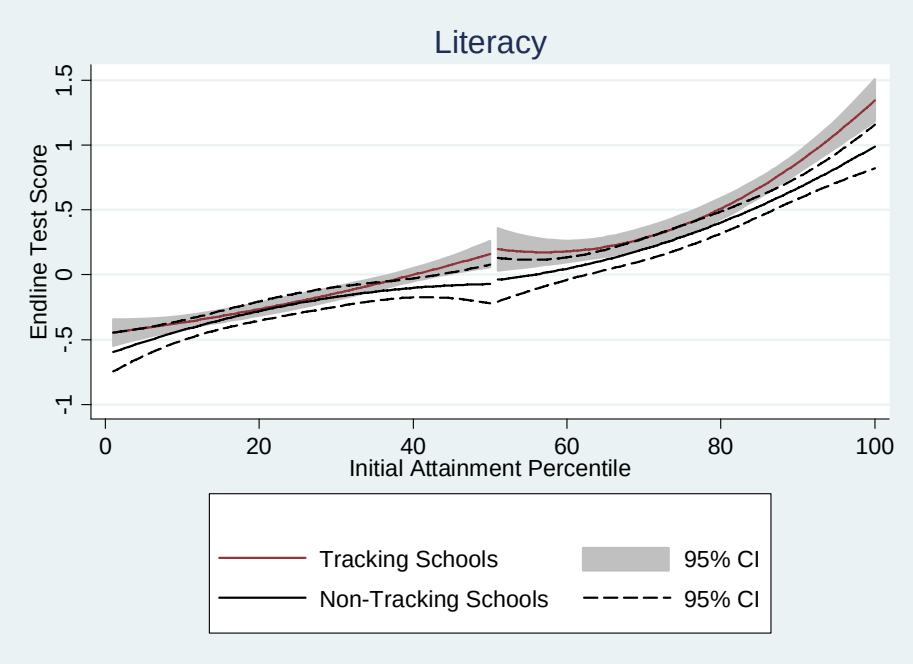
Non-Tracking vs. Tracking Schools



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

Figure 3: Local Polynomial Fits of Endline Score by Initial Attainment

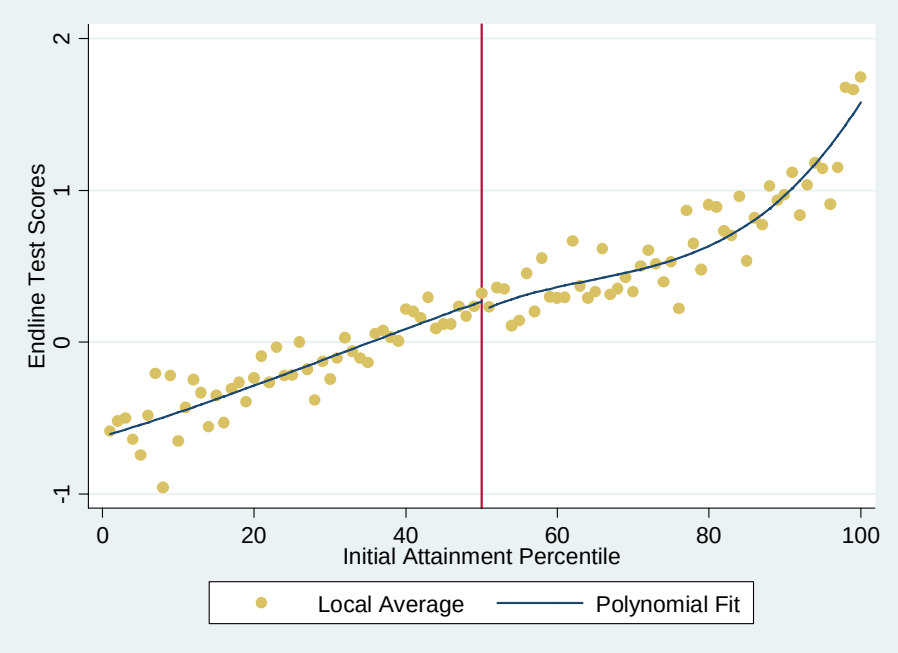




Notes: Fitted values from regressions that include a second order polynomial estimated separately on each side of the percentile=50 threshold.

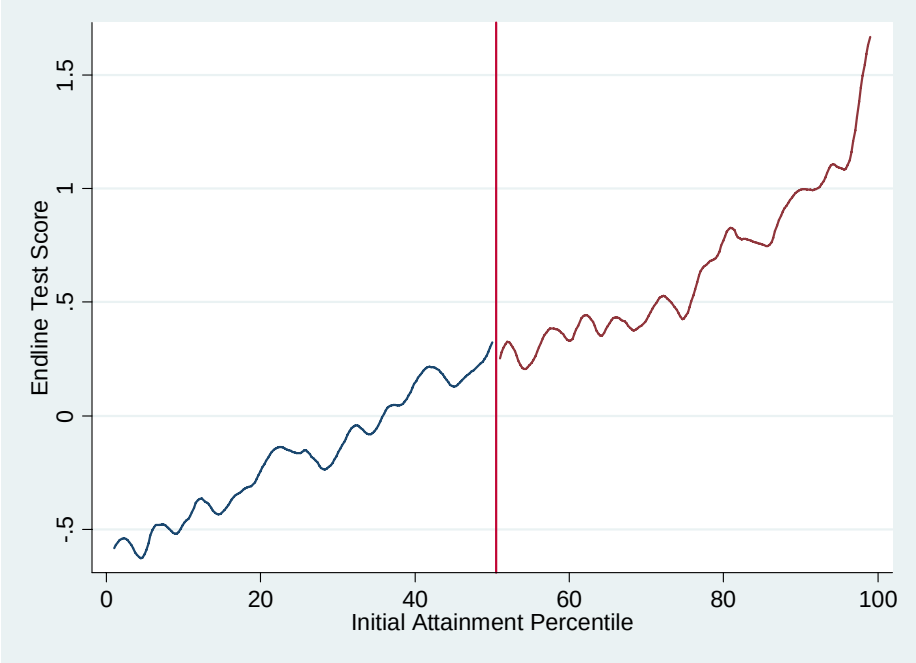
Figure A1: Peer Quality and Endline Scores in Tracking Schools

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.

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

	All ETP Schools				
	Non-Tracking Schools		Tracking Schools		*P-value* Tracking = Non-Tracking
Panel A. Baseline School Characteristics	Mean	SD	Mean	SD	
Total enrollment in 2004	589	*232*	549	*198*	0.316
Number of government teachers in 2004	11.6	*3.3*	11.9	*2.8*	0.622
School pupil/teacher ratio	37.1	*12.2*	35.9	*10.1*	0.557
Performance at national exam in 2004 (out of 400)	255.6	*23.6*	258.1	*23.4*	0.569
Panel B. Class Size Prior to Program Inception (March 2005)					
Average class size in first grade	91	*37*	89	*33*	0.764
Proportion of female first grade students	0.49	*0.06*	0.49	*0.05*	0.539
Average class size in second grade	96	*41*	91	*35*	0.402
Panel C. Class Size 6 Months After Program Inception (October 2005)					
Average class size in first grade	44	*18*	42	*15*	*0.503*
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 second grade	42	*17*	42	*20*	*0.866*
Range of class sizes in sample (second grade)	18-93		21-95		
Number of Schools	61		60		121

	Within Tracking Schools				
	Assigned to Bottom Section		Assigned to Top Section		*P-value* Top = Bottom
Panel E. Comparability of two sections within Tracking Schools	Mean	SD	Mean	SD	
Proportion Female	0.49	*0.09*	0.50	*0.08*	0.38
Average Age at Endline	9.04	*0.59*	9.41	*0.60*	0.00
Average Standardized Baseline Score (Mean 0, SD 1 at school level)	-0.81	*0.04*	0.81	*0.04*	0.00
Average Std. Dev. Within Section in Standardized Baseline Scores	0.49	*0.13*	0.65	*0.13*	0.00
Average Standardized Endline Score (Mean 0, SD 1 in Non-Tracking group)	-0.15	*0.44*	0.69	*0.58*	0.00
Average Std. Dev. Within Section in Standardized Endline Scores	0.77	*0.23*	0.88	*0.20*	0.00
Assigned to Contract teacher	0.53	*0.49*	0.46	*0.47*	0.44
Respected Assignment	0.99	*0.02*	0.99	*0.02*	0.67

	Within Non-Tracking Schools				
	Section A (Assigned to Civil-Service Teacher)		Section B (Assigned to Contract Teacher)		*P-value* A = B
Panel F. Comparability of two sections within Non-Tracking Schools					
Proportion Female	0.49	*0.06*	0.49	*0.06*	0.89
Average Age at Endline	9.07	*0.53*	9.00	*0.45*	0.45
Average Standardized Baseline Score (Mean 0, SD 1 at school level)	0.003	*0.10*	0.002	*0.11*	0.94
Average Std. Dev. Within Section in Standardized Baseline Scores	1.005	*0.08*	0.993	*0.08*	0.43
Average Standardized Endline Score (Mean 0, SD 1 in Non-Tracking group)	0.188	*0.46*	0.047	*0.48*	0.10
Average Std. Dev. Within Section in Standardized Endline Scores	0.937	*0.24*	0.877	*0.24*	0.16

Notes: School averages.

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	0.139	0.176	0.192	0.182	0.125	0.159	0.139	0.156	0.123	0.155	0.198	0.166
	(0.078)*	(0.077)**	(0.093)**	(0.093)*	(0.065)*	(0.064)**	(0.073)*	(0.083)*	(0.08)	(0.083)*	(0.108)*	(0.098)*
(2) In Bottom Half of Initial Distribution			-0.036				0.04				-0.091	
x Tracking School			(0.07)				(0.07)				(0.08)	
(3) In Bottom Quarter				-0.045				0.012				-0.083
x Tracking School				(0.08)				(0.09)				(0.08)
(4) In Second to Bottom Quarter				-0.013				0.026				-0.042
x Tracking School				(0.07)				(0.08)				(0.07)
(5) In Top Quarter				0.027				-0.026				0.065
x Tracking School				(0.08)				(0.07)				(0.08)
(6) Assigned to Contract Teacher		0.181	0.18	0.18		0.162	0.16	0.161		0.16	0.16	0.16
		(0.038)***	(0.038)***	(0.038)***		(0.038)***	(0.038)***	(0.037)***		(0.038)***	(0.038)***	(0.038)***
Individual Controls	no	yes	yes	yes	no	yes	yes	yes	no	yes	yes	yes
Observations	5795	5279	5279	5279	5796	5280	5280	5280	5796	5280	5280	5280
Total effects on bottom half and bottom quarter												
Coeff (Row 1)+Coeff (Row 2)			0.156				0.179				0.107	
Coeff (Row 1)+Coeff (Row 3)				0.137				0.168				0.083
p-value (Total Effect for Bottom = 0)			0.038	0.095			0.016	0.049			0.127	0.237
p-value (Effect for Top quarter = Effect for Bottom Quarter)				0.507				0.701				0.209
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)**	(0.073)**	(0.079)***	(0.088)***	(0.059)**	(0.062)**	(0.064)**	(0.075)**	(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				-0.117				-0.042				-0.152
x Tracking School				(0.09)				(0.10)				(0.085)*
(4) In Second to Bottom Quarter				-0.096				-0.073				-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)
(6) Assigned to Contract Teacher		0.094	0.094	0.094		0.062	0.061	0.061		0.102	0.102	0.103
		(0.032)***	(0.032)***	(0.032)***		(0.031)**	(0.031)**	(0.031)**		(0.031)***	(0.031)***	(0.031)***
Individual Controls	no	yes	yes	yes	no	yes	yes	yes	no	yes	yes	yes
Observations	5490	5001	5001	5001	5490	5001	5001	5001	5496	5007	5007	5007
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)				0.118				0.126				0.089
p-value (Total Effect for Bottom = 0)			0.091	0.229			0.122	0.216			0.117	0.319
p-value (Effect for Top quarter = Effect for Bottom Quarter)				0.365				0.985				0.141

*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	0.130	0.162	0.731	0.084	0.206	0.168
	(0.076)*	(0.100)		(0.083)	(0.084)**	
Girls	0.188	0.222	0.661	0.190	0.227	0.638
	(0.089)**	(0.104)**		(0.098)*	(0.089)**	
Test (Boys = Girls): p-value	0.417	0.470		0.239	0.765	
Panel B: By Teacher Type						
Regular Teacher	0.048	0.225	0.155	0.086	0.198	0.329
	(0.088)	(0.120)*		(0.099)	(0.098)**	
Contract Teacher	0.255	0.164	0.518	0.181	0.246	0.605
	(0.099)**	(0.118)		(0.094)*	(0.103)**	
Test (Regular = Contract): p-value	0.076	0.683		0.395	0.702	

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

	ALL			25th-75th percentiles only	Bottom 25th percentiles	Top 25th percentiles only
	Total Score	Math Score	Lit Score	Total Score	Total Score	Total Score
	(1)	(4)	(5)	(6)	(7)	(8)
Panel A: Reduced Form						
Average Baseline Score of Classmates‡	0.346	0.323	0.293	-0.052	0.505	0.893
	(0.150)**	(0.160)**	(0.131)**	(0.227)	(0.199)**	(0.330)***
Observations	2188	2188	2188	2188	2188	2188
School Fixed Effects	x	x	x	x	x	x
Panel B: IV						
Average Endline Score of Classmates	0.445	0.47	0.423	-0.063	0.855	1.052
(predicted)	(0.117)***	(0.124)***	(0.120)***	(0.306)	(0.278)***	(0.368)***
Observations	2188	2188	2189	1091	524	573
School Fixed Effects	x	x	x	x	x	x
Panel C: First-Stage for IV: Average Endline Score of Classmates						
	Average Total Score	Average Math Score	Average Lit Score	Average Total Score	Average Total Score	Average Total Score
Average (Standardized) Baseline Score	0.768	0.680	0.691	0.795	0.757	0.794
of Classmates‡	(0.033)***	(0.033)***	(0.030)***	(0.056)***	(0.066)***	(0.070)***

*Notes: Sample restricted to the 61 non-tracking 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 school level in parentheses. ***, **, * indicates significance at the 1%, 5% and 10% levels respectively.*

‡ *This variable has a mean of 0.0009 and a standard deviation of 0.1056. 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 5
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	0.010	0.001	-0.045	-0.051	0.088	0.034	0.027
	(0.093)	(0.079)	(0.106)	(0.089)	(0.006)	(0.136)	(0.145)
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	-0.012	-0.002				-0.068	-0.004
	(0.117)	(0.106)				(0.205)	(0.277)
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 2).

*Panel A, columns 1-2 and 6-7: the score was regressed on a dummy "assigned to bottom section" and individual controls (age, gender, dummy for being assigned to contract teacher and, for columns 1 and 2, a polynomial in initial percentile). 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 was 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.*

Regressions in columns 6-7 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 6
Teacher Effort and Student Presence

	All Teachers		Government Teachers		ETP Teachers		Students
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Teacher Found in school on random school day	*Teacher found in class teaching (unconditional on presence)*	*Teacher Found in school on random school day*	*Teacher found in class teaching (unconditional on presence)*	*Teacher Found in school on random school day*	*Teacher found in class teaching (unconditional on presence)*	*Student found in school on random school day*
Tracking School	0.041	0.096	0.054	0.112	-0.009	0.007	-0.015
	(0.021)**	(0.038)**	(0.025)**	(0.044)**	(0.034)	(0.045)	(0.014)
Bottom Half x Tracking School	-0.049	-0.062	-0.073	-0.076	0.036	-0.004	0.003
	(0.029)*	(0.040)	(0.034)**	(0.053)	(0.046)	(0.057)	(0.007)
Years of Experience Teaching	0.000	-0.005	0.002	0.002	-0.002	-0.008	
	(0.001)	(0.001)***	(0.001)*	(0.001)	(0.003)	(0.008)	
Female	-0.023	0.012	-0.004	0.101	-0.034	-0.061	-0.005
	(0.018)	(0.026)	(0.020)	(0.031)***	(0.032)	(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	2098	1633	1633	465	465	44059
Mean in Non-Tracking Schools	0.837	0.510	0.825	0.450	0.888	0.748	0.865
F (test of joint significance)	2.718	9.408	2.079	5.470	2.426	3.674	5.465
p-value	0.011	0.000	0.050	0.000	0.023	0.001	0.000

*Notes: The sample includes 60 tracking and 61 non-tracking schools. 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.*

Table 7
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 tracking and 61 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.*

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 A1
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)	(5)	(6)
	Transferred to other school	If not transferred: missed test	Total Attrition	Total Attrition, Subsample Around Median	Total Attrition	Subsample Around Median
Tracking School	0.003	0.011	0.014	-0.015	0.000	-0.087
	(0.012)	(0.019)	(0.022)	(0.066)	(0.028)	(0.067)
In Bottom Half of Initial Distribution	0.027	-0.013	0.014	-0.006	0.050	0.042
	(0.016)*	(0.020)	(0.025)	(0.075)	(0.028)*	(0.073)
In Bottom Half of Initial Distribution x Tracking School	-0.012	0.02	0.007	0.043	0.006	0.041
	(0.013)	(0.019)	(0.023)	(0.083)	(0.026)	(0.080)
Girl	0.012	0.029	0.041	0.020	0.024	0.055
	(0.009)	(0.014)**	(0.016)**	(0.049)	(0.016)	(0.060)
Girl x Tracking School	0.004	-0.048	-0.044	-0.026	-0.022	0.039
	(0.013)	(0.016)***	(0.019)**	(0.067)	(0.021)	(0.079)
Assigned to Contract Teacher	-0.006	-0.019	-0.025	-0.078	-0.014	-0.086
	(0.010)	(0.011)*	(0.014)*	(0.046)*	(0.015)	(0.055)
Assigned to Contract Teacher x Tracking School	0.018	0.009	0.027	0.056	0.038	0.142
	(0.015)	(0.019)	(0.025)	(0.076)	(0.026)	(0.092)
In Bottom Half of Initial Distribution	0.01	-0.035	-0.025	-0.01	-0.011	0.031
x Assigned to Contract Teacher x Tracking School	(0.019)	(0.030)	(0.036)	(0.093)	(0.037)	(0.104)
Constant	0.034	0.172	0.205	-0.737	0.23	2.11
	(0.018)*	(0.026)***	(0.032)***	(1.664)	(0.034)***	(1.747)
Observations	7345	7345	7345	517	7340	515
Mean	0.057	0.119	0.175	0.175	0.224	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. Columns 4 and 6: restricted to students within 0.1 standard deviation of median at baseline.