STAT 240 Final Project

Rebecca Barter, Andrew Do and Kellie Ottoboni

May 6, 2015

1 Introduction

Over the last two decades developing countries have seen an increase in the number of new primary school entrants, driven in part, by the elimination of school fees. For example, between 1999 and 2004 the number of new entrants to primary school in sub-Saharan Africa increased by more than 30 percent [UNESCO, 2007]. Although this influx of new students is undeniably a positive development, steps need to be taken to ensure that the quality of education is not diminished. For example, by 2005, the average first grade class size in Kenya had swelled to 83, with 28 percent of first grade classes containing more than 100 students [Duflo et al., 2007]. Moreover, many of the new students were significantly less prepared than those in the past.

Unfortunately, little prior work had been undertaken to evaluate the most effective methods of handling such an influx of students. In a randomized experiment, Duflo et al. aim to answer several questions related to a number methods of resource allocation in primary education [Duflo et al., 2011]. In particular, the investigators aim to assess the impact of reduction in pupil-teacher ratios, implementing tracking (separating classes into high and low streams based on prior test scores) and different institutional environments (type of teacher, and whether or not the school undertakes teacher monitoring and education).

In this report, we investigate the data obtained from the study undertaken by Duflo et al, with a focus on evaluating the impact of tracking on student achievement.

2 Tracking

Tracking involves separating pupils by academic ability within schools. In particular, a student is assigned to the high stream if their prior achievement is above the median, and is assigned to the low stream otherwise. There have been multitudes of studies involving the effects tracking in developed countries, with no overall consensus on whether it is beneficial or detrimental to future student achievement. However those who claim that tracking is beneficial to students have several arguments. For example the reduced skill differential has the potential to allow for better lesson execution and time allocation by teachers; they can focus on teaching at a level that will benefit all students in the class, rather than having to cater to a wide range of abilities. Further, it is possible that ensuring that students are placed in classes of the appropriate difficulty levels will reduce behavioral outlashing by students. Finally, proponents argue that the "value-added" is maximized within each group.

In contrast, critics of tracking argue that when students of all levels are integrated, high performing students are given the opportunity to synthesize ideas they've learned by teaching low performers. Further, critics pose the idea that tracking is a self-fulfilling prophecy; students in the low stream will not achieve as highly than they otherwise might simply because they have been placed in a class that implies that they are not as able. This idea is reinforced by the argument that teachers may require less of students in the low stream, and thus that the

education gap will widen between the high and low achieving groups.

Through our analysis of the data provided by Duflo et al., we will provide data-driven evidence for several of these arguments, and present our position on the effects of tracking on disadvantaged schools in Kenya. In particular, we aim to answer whether 1) tracking has a differential effect on different students of different baseline abilities, 2) there is long-term value-added by introducing tracking and 3) whether tracking creates a gap between students of approximately average ability, since these students can be considered to be randomly assigned to the low and high streams.

3 Study Design

The study conducted by Duflo et al. involves data from a randomized experiment spanning 18 months involving the first grade class from 210 primary schools in Western Kenya. These schools have a combined 21,000 students and prior to the experiment each school has only a single first grade class taught by a centrally-hired teacher with civil service protection (hereafter referred to as a "civil service teacher"). The study involves several layers of randomization which are summarized in Figure 1. The "Extra Teacher Program" (ETP) provided funds to 140 schools randomly selected from the pool of 210 schools to hire an extra teacher for first grade classes. These teachers were hired locally, and earned approximately a quarter of the civil service teachers, but had the same academic qualifications. In 70 of these 140 ETP schools, tracking was introduced (these schools are "tracked" schools), whereby the two classes were divided by initial achievement, and the classes were randomly assigned to either a civil service teacher or a contract teacher. In the other half of the ETP schools ("non tracked" schools), students were randomly assigned to either the local contract teacher or the existing civil service teacher. Finally, half of the 70 non-tracked ETP schools and half of the tracked ETP schools were given funds to empower the local school committee to monitor and train teachers (these schools are referred to as the "monitored" schools).

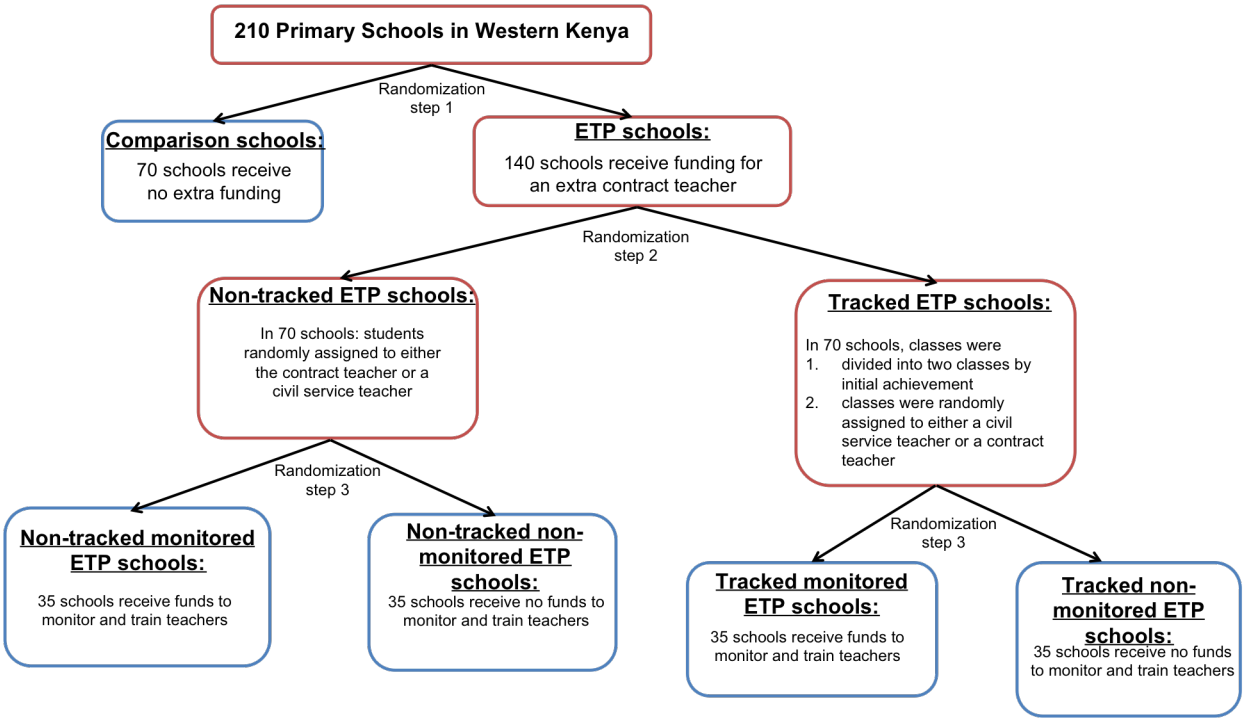

Figure 1: A flowchart describing the randomization steps of the study

Stream assignment in tracked schools was based on initial test scores (baseline test) which were administered locally within schools. Thus these baseline test scores are internally consistent within schools but not comparable across schools. The success of the program was assessed based on scores from a standardized mathematics and language test taken by 60 students from each school (approximately 7000 students all together) after 18 months (the end line test). Another test was also taken by the same students after 24 months (the long-term followup test). We note that the ETP funding ceased after the 18 month endline, so the tracking was no longer in place at the 24 month followup. These tests contained numeracy and literacy questions ranging from counting and identifying letters to subtracting two-digit numbers and writing words.

4 Exploratory Data Analysis

The original dataset contains observations for 7022 students over 100 variables, including individual question scores for each of the endline and follow-up tests, as well as information such as school ID, school district, whether or not the student came from a tracked or non-tracked school, gender, age, teacher-type and stream assignment (for students in the tracked schools). We note that the data contains data for first-grade students whose ages range hugely from 5 to 19, with the median age being 9 (Figure 2).

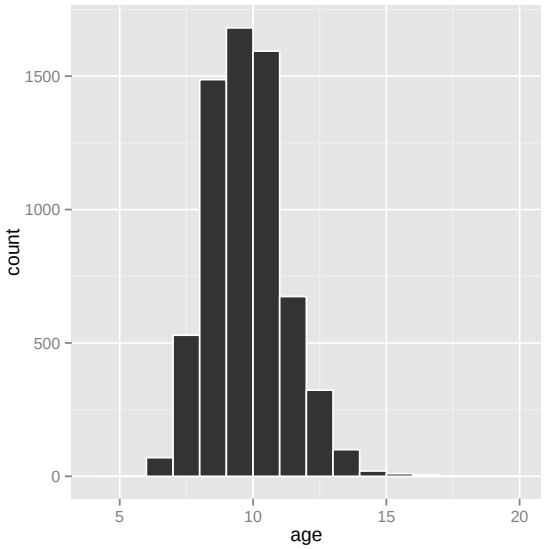

Figure 2: Age ranges of first-grade students

We compared the characteristics of students in tracking and in non-tracking schools (Table 1), and we found that in terms of age, gender and endline attrition (the percentage of students that were not present for the endline test), the students in tracked and non-tracked schools are extremely comparable. We did, however, find that the tracking and non-tracking schools are not distributed equally in terms of location. The 210 schools come from a total of 9 school zones, but we see that for Butere East and Municipality, there are approximately twice as many tracked schools as non-tracked schools, whereas for Khwisero West, there are more than twice as many non-tracked schools as there are tracked schools (Figure 3). Since these regions may have different education standards, we anticipate potential bias by blocking our analyses by school zone.

	Tracked School (n = 3613)	Non-tracked School (n = 3409)
Female (%)	0.49	0.49
Age (mean)	9.36 (1.47)	9.18 (1.46)
Endline attrition (%)	0.17	0.17

Table 1: A comparison of the characteristics of the students in tracked and non-tracked schools. The numbers in parentheses are standard deviations.

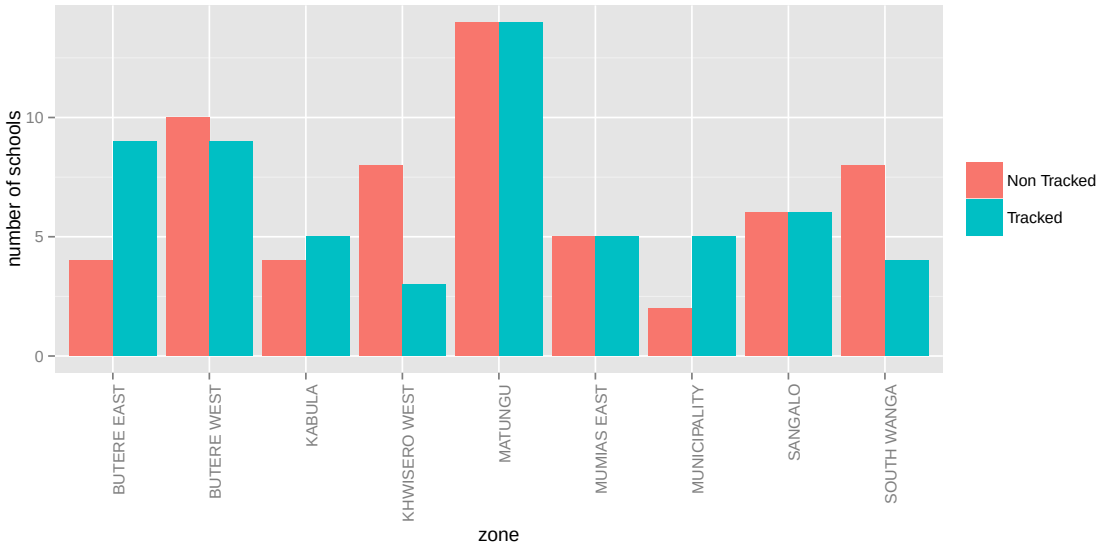

Figure 3: The number of tracked and non-tracked schools in each school zone

Next, for students in tracked schools, we compared the characteristics for students in the high and low stream (Table 2). We found that the students in the high stream were slightly older than those in the low stream and were more likely to be present at the endline test.

	Low stream (n = 1808)	High stream (n = 1805)
Female (%)	0.49	0.50
Age (mean)	9.14 (1.47)	9.58 (1.44)
Endline attrition (%)	0.19	0.16

Table 2: A comparison of the characteristics of the students in high and low streams. The numbers in parentheses are standard deviations.

In addition, we found concerning inconsistencies in the available data. For example the initial grade percentiles provided do not correspond to the percentiles when calculated manually (Figure 4). The codebook that accompanied the data stated that the provided percentiles had been imputed, although what data was used to conduct the imputation and the method of imputation is not described. We thus decided to use our own calculated percentiles for our subsequent analysis.

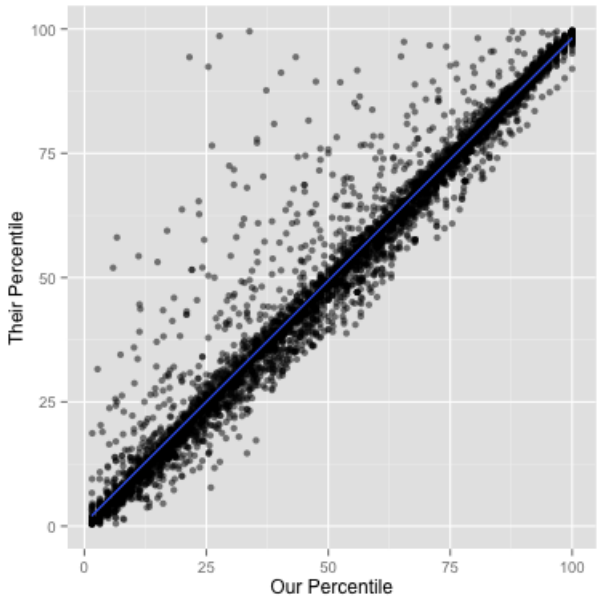

Figure 4: A comparison of the percentiles reported in the data and the percentiles obtained when calculating manually.

Off the 758 students without baseline scores, 756 attended non-tracking schools. To see if the classes at the non-tracking schools mixed their high-performers and low-performers sufficiently, we compared the end-line test percentiles of each student against the average of the classmates. We found that the students' percentiles had no correlation with their peers' average percentiles. Figure 5 shows this result.

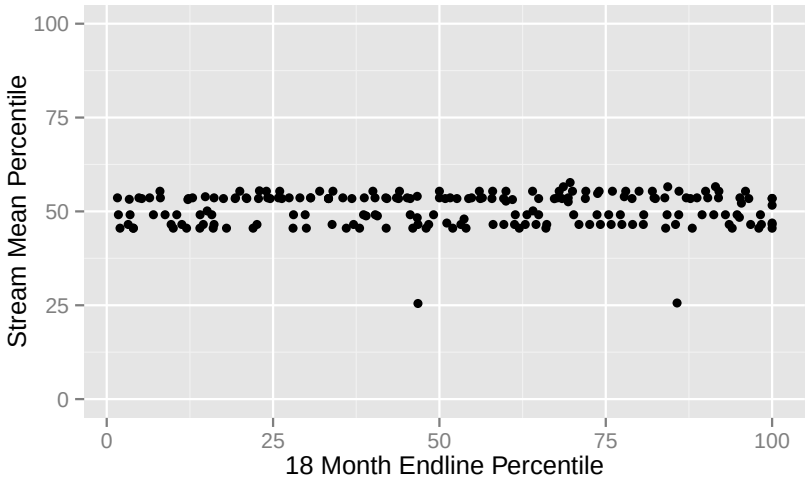

Figure 5: A comparison of endline percentiles and stream mean percentiles for students without baseline scores. Not having a baseline score has no correlation with classroom placement. This turns out to be a nonissue almost all the students with no baseline score attended non-tracking schools. The two deviate points seen in the figure belong to students in tracking schools who were placed in low streams.

Another issue we detected is that within tracking schools, some crossover between low and high streams occurred

(Figure 6), although we note that this crossover is minor (only about 5.3%), it does pose minor issues in our subsequent analyses. Duflo et al. state that the crossover was primarily due to siblings who wanted to remain in the same class, and class assignment was only changed upon a parent's request.

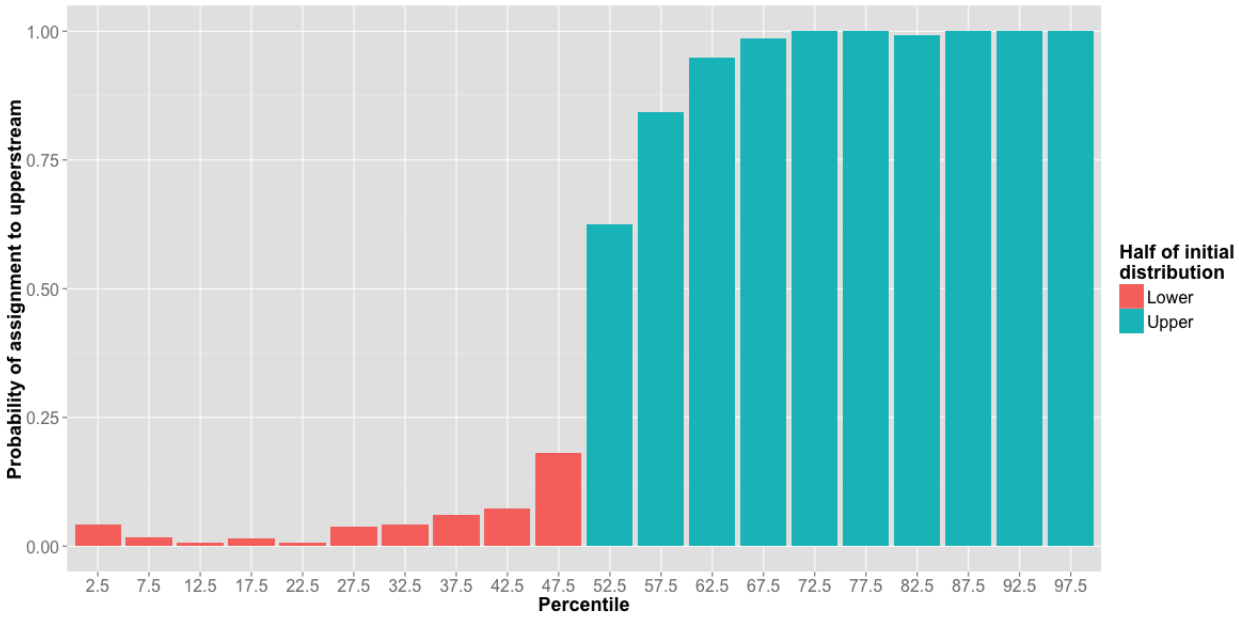

Figure 6: The probability of assignment to high stream versus the baseline score. If no crossover occurred, the probability would be 0 until baseline score reached 0.5, after which the probability would be 1.

5 Analysis

5.1 Value-added over time by tracking

Ideally the data collected would have included standardized baseline scores so we could measure the value-added from tracking by comparing baseline with the 18 month endline scores. Unfortunately, this was not how the study was designed as the baseline test scores are only consistent within schools and not across. As a result, we chose to measure the long-term value-added by comparing the 24 month follow-up scores. We can interpret this comparison as a measure of a lasting effect on students, even after the tracking was no longer in place. However, by using the long-term follow-up data, we are relying purely on randomization to average out baseline abilities. As seen in figure 7, there seems to be a positive effect on the tracked students.

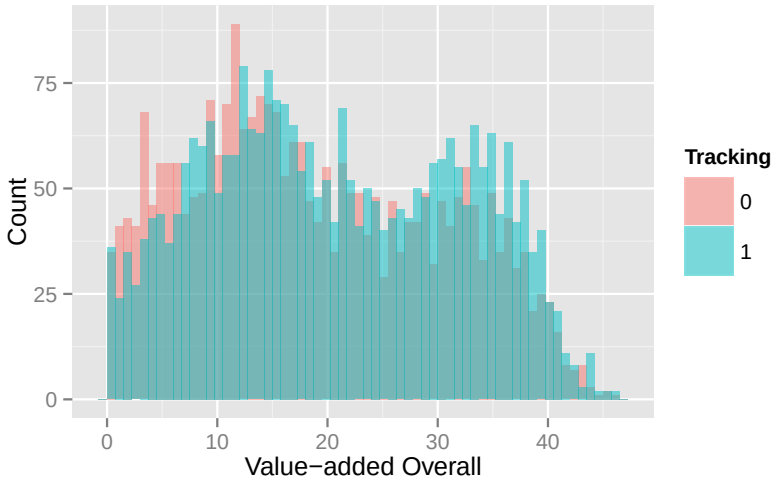

Figure 7: Histogram of value-added at the student level, separated by attendance at tracking and non-tracking schools. There seems to be a slight right-shift in value-added.

We tested the significance of any observed differences in average value-added between tracking and nontracking schools using permutation t-tests. To address the fact that randomization was done at the school-level, we first had to aggregate the student-level data by school, which gave us variance estimates for each of the school-level data points. The t-statistic for the difference in means was calculated using unpooled variance and the null-distribution was approximated using 10000 random shufflings of the data within school zones (to account for the aforementioned imbalances). The observed value-added is significant at $\alpha = 0.1$ level for the overall, word recognition, sentence comprehension, letter recognition, spelling, and literacy scores. Math showed no significant differences, but the design of the endline test may have made it hard to discern any value added as there were only nine items in the arithmetic section compared to the 128 in the literacy sections. The results of the tests can be seen in table 5.1.

	Overall	Word	Sent	Letter	Spell	Literacy	Math
Value Added	1.933	1.566	1.231	4.357	0.509	1.263	0.665
t-statistic	0.130	0.128	0.080	0.141	0.108	0.133	0.095
p-value	0.056	0.069	0.041	0.027	0.061	0.041	0.168

Table 3: Test for differences between 24-month test scores. Shuffling was done within school-zone groups

One could argue that the extra funding for school-based management could be a driving force behind the increase in student performance. However, when we tested the claim, we found no significance in any of the areas. The results can be seen in table 4. In figure 8, we see that when comparing SBM-tracking to SBM-non-tracking schools, we still see a small but visible increase in the tracking schools.

	Overall	Word	Sent	Letter	Spell	Literacy	Math
Value Added	1.022	0.559	0.563	2.709	0.226	0.600	0.421
t-statistic	0.069	0.046	0.037	0.088	0.048	0.063	0.060
p-value	0.160	0.400	0.213	0.124	0.413	0.236	0.106

Table 4: Test for differences 24-month test scores in SBM and non-SBM schools. Shuffling was done within school-zone groups

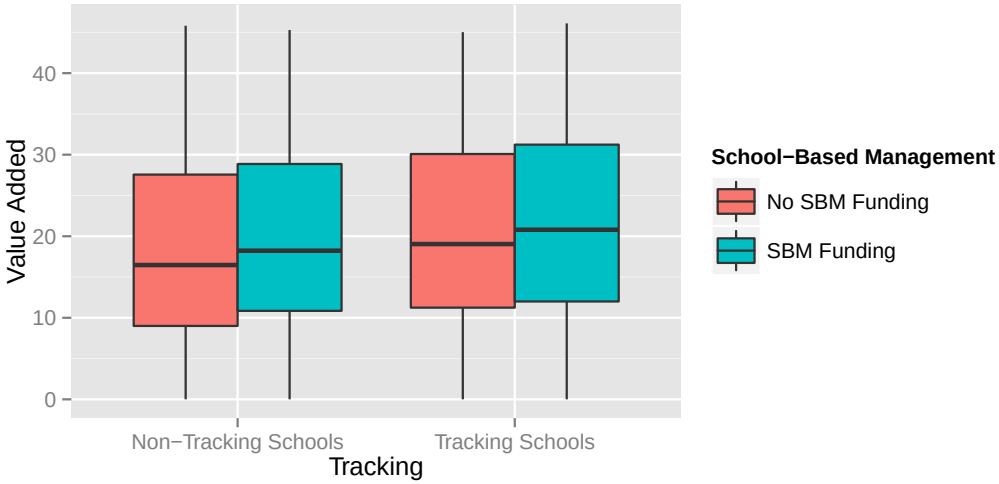

Figure 8: Spread of value added by tracking and participation in SBM. The small difference seen between SBM and non-SBM schools was deemed insignificant by a permutation test. Note that even after accounting for the additional SBM funding, there is still a visible difference between tracking and non-tracking schools.

To assess the practical difference between tracking and non-tracking schools, we focus on the Value Added row of table 5.1. Let us suspend the discussion outside the context of p-values. The difference scores means that on average the students in tracking schools were able to identify 10% more words, understand 5% more sentences, and recognize 10% letters than their non-tracked counterparts. Certainly the effect is not huge, but all the value-added scores are positive, suggesting that tracking, at the very least, does not have a deleterious effect in the classroom on average.

5.2 Stratified permutation tests for comparing tracking with non-tracking schools

We begin with a comparison of scores achieved by students in tracking schools with students in non-tracking schools. To perform this comparison, we tested whether tracking had an effect on the 18 month follow-up scores for students of different baseline abilities using stratified permutation tests with a t-statistic. We split students into four strata based on their percentile on their school's baseline exam score. For this, we used the percentiles that we calculated rather than the ones supplied, due to the inconsistencies we discussed previously. Within each stratum, we calculated the test statistic and permuted treatment assignments. This procedure allowed us to carry out tests first within individual strata, and then for all students by taking a weighted average of the stratum-specific statistics. Figure 9 shows that in each stratum, tracking has a positive effect on follow-up test scores. The effect was significant overall and for each quartile except for the third (Table 5, two-sided alternative).

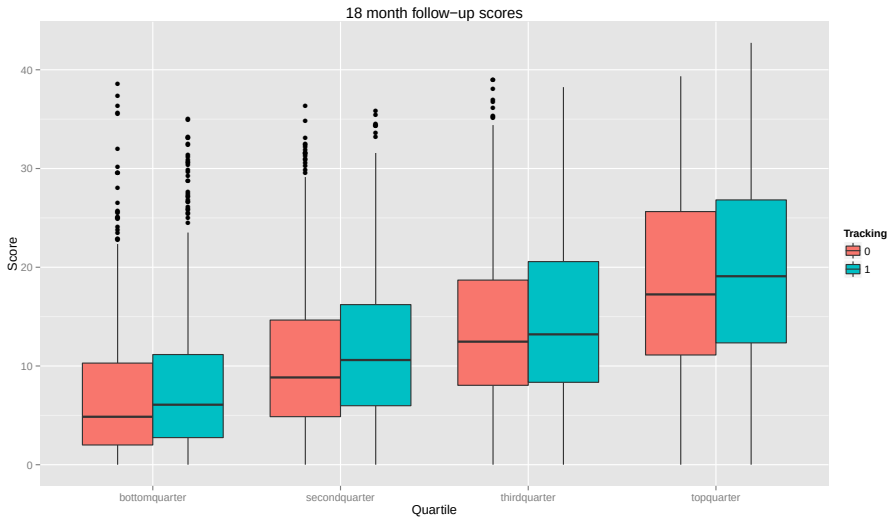

Figure 9: Distribution of follow-up scores in tracking and non-tracking schools, stratified by students' quartile of baseline score

	Bottom quarter	Second quarter	Third quarter	Top quarter	Overall
Difference in means	0.963	1.318	0.798	1.425	1.124
t	2.466	2.998	1.690	2.810	2.493
P-value	0.013	0.003	0.091	0.005	0.000

Table 5: Test for differences in final score, stratified by baseline quartile and overall.

Next, we investigated the effect of tracking on the individual components of the 18 month follow-up score. Table 6 shows the effect of tracking on each topic, by baseline quartile and overall. Tracking increased students' average letter score, spelling score, literacy score, and math score. As with the total score, the positive effects of tracking seemed to be less pronounced in the third quartile of students.

	Bottom quarter	Second quarter	Third quarter	Top quarter	Overall
Word Score	0.048	0.542	1.017	1.084	0.661
	(0.963)	(0.597)	(0.319)	(0.272)	(0.188)
Sentence Score	-0.297	-0.761	0.572	1.173	0.161
	(0.770)	(0.450)	(0.560)	(0.235)	(0.746)
Letter Score	**2.988**	**3.389**	**1.924**	**4.466**	**3.191**
	(0.002)	**(0.001)**	**(0.050)**	**(0.000)**	**(0.000)**
Spelling Score	1.290	1.111	0.740	1.689	**1.210**
	(0.197)	(0.268)	(0.450)	(0.093)	**(0.015)**
Literacy Score	1.504	1.530	1.291	**2.405**	**1.680**
	(0.135)	(0.131)	(0.202)	**(0.015)**	**(0.001)**
Math Score	**2.869**	**3.878**	1.718	**2.508**	**2.750**
	(0.005)	**(0.000)**	(0.086)	**(0.011)**	**(0.000)**

Table 6: t-statistics (p-values) for the test of differences in subject-level final score between tracking and non-tracking schools, stratified by baseline quartile and overall. Significant results are presented in boldface text.

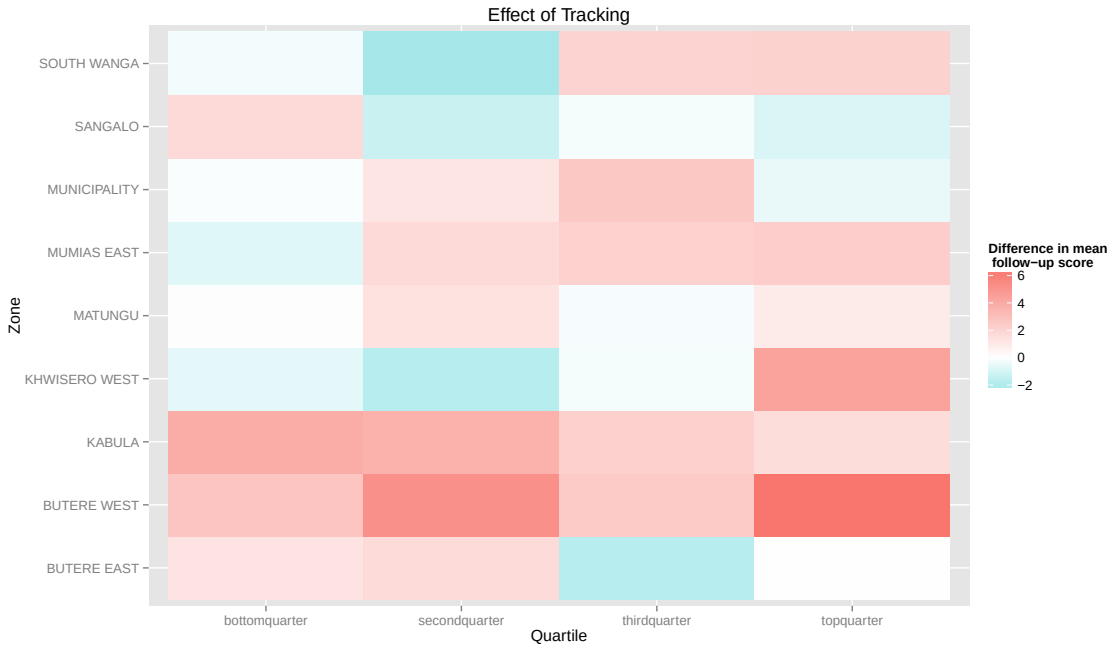

Figure 10: Difference in mean 18 month follow-up scores between students at tracking and non-tracking schools

We next investigated the potential bias in the estimated treatment effect due to differences in school zone. Figure 10 shows the difference in mean 18 month follow-up scores between tracking and non-tracking schools, broken down by school zone. The effect of tracking is heterogeneous across school zones. While most zone-quartile combinations show a positive effect, there are several with a negative effect in the bottom quartile. It appears that two zones in particular, Kabula and Butere West, are driving the positive effect of tracking that we see when all zones are combined.

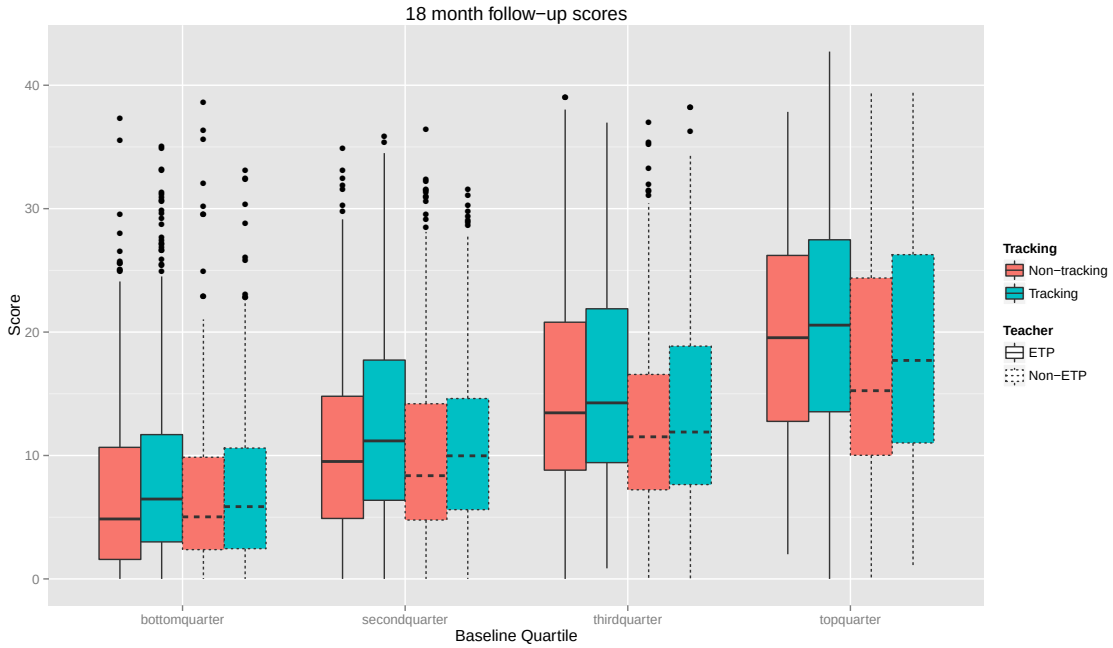

Figure 11: Distribution of follow-up scores in tracking and non-tracking schools, stratified by students' quartile of baseline score and type of teacher

	Bottom quarter	Second quarter	Third quarter	Top quarter	Overall
ETP	**2.506**	**3.695**	0.759	1.918	**2.254**
	(0.011)	**(0.000)**	(0.459)	(0.061)	**(0.000)**
Non-ETP	0.817	0.522	1.906	**2.404**	**1.423**
	(0.421)	(0.611)	(0.061)	**(0.017)**	**(0.004)**

Table 7: t-statistics (p-values) for the test of difference in mean follow-up exam score between students in tracking vs non-tracking schools, stratified by ETP vs civil servant teachers. Significant results are presented in boldface text.

Finally, we considered the possibility that the effect of tracking may be different depending on whether the teacher is a civil servant or an ETP teacher. Duflo et. al. raise the possibility that different types of teachers may teach differently; for example, ETP teachers may have a greater incentive to help the lowest quartile of students in order to secure a teaching position in the future. We did a separate stratified permutation test for each type of teacher to compare the effect of tracking on 18 month follow-up scores, for each baseline quartile and overall. Figure 11 shows that for both types of teacher, tracking increases follow-up exam scores. However, Table 7 presents an interesting trend. Lower quartile students with an ETP teacher in tracking schools had a significantly higher follow-up score than those in non-tracking schools, but the effect was not as strong in the upper half of the distribution. Conversely, students above the median with a civil service teacher had significantly higher follow-up scores in tracking schools compared to non-tracking schools, but the trend did not hold for the lower half of the distribution. In tracking schools, the type of teacher assigned to the high and low stream has an impact on how effective tracking is.

5.3 Value-added over time by tracking

Unfortunately since the baseline test scores were not comparable across different schools, we were unable to measure the value-added from tracking by comparing baseline with the 18 month endline scores in tracking schools. As a result, we thus chose to measure the long-term value-added by comparing 18 month endline scores with the 24 month follow-up scores. This comparison allows us to see if tracking yielded a lasting effect on students, even after the tracking was no longer in place.

5.4 Regression discontinuity for comparing students in the high stream with those in the low stream

We now aim to analyze the effect of stream assignment on students in tracked schools who initially achieved near the cutoff point. This analysis is based on the idea that students who achieved near the cutoff point have approximately equivalent abilities and it was random noise that lead to their assignment into either the high or low stream. For example, if a student slept poorly the night before the baseline test, they may have scored a few points below the cutoff point, whereas they may otherwise have scored a few points above the cutoff point. As a result, we can consider the students whose initial scores fall within a small window of the cutoff point as being randomly assigned into either the high or low stream, and we can thus evaluate the treatment effect of assignment to the high stream (for example) as if it were a random assignment for the students within some cutoff window.
If we plot the baseline score versus endline score aggregated over all students in small initial percentile bins (Figure 12), visually, there does not appear to be a significant jump at the cutoff point (which is what we would expect to see in the presence of a regression discontinuity). In fact, it seems as though the slope becomes less steep, implying that being assigned to the upper stream for students near the cutoff point is actually detrimental to their endline achievements.

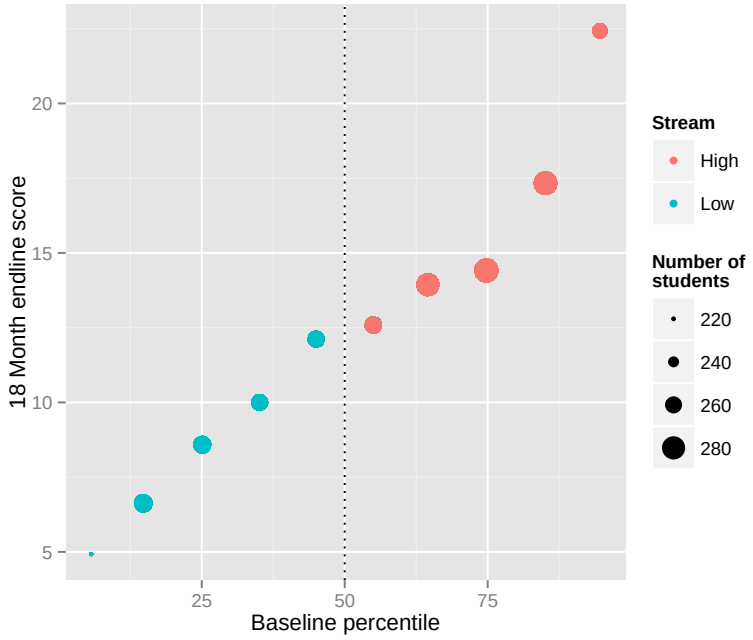

Figure 12: Aggregated scatterplot of baseline score versus 18 month endline score. Each point is the median score of all students who fall within the corresponding score bin, each having a width of 10 percentile points. Each point is colored by the majority stream of the students in each bin.

Moreover, when zooming in on a small window around the cutoff point (Figure 13), we see that there appears to be a discontinuity jump in the negative direction, with those students almost close to the cutoff point who were put into the high stream appearing to perform worse than the students who were put into the low stream. Thus overall, it appears as though the students near the cutoff who are put in the low stream will be better of than those put in the high stream.

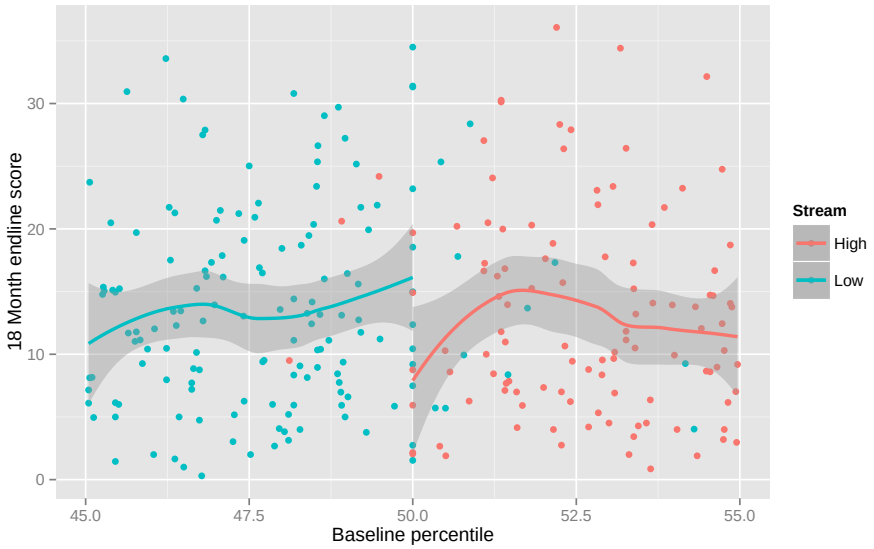

Figure 13: A scatterplot of baseline percentile versus 18-month score within a window of width 10 percentile points (5 percentile points in each direction around the cutoff point) with loess fitted curves for the high and low stream.

Not only do we see this trend in the overall score, but when we restrict to individual subject areas, we see the same trend reflected in Literacy, Math, Spelling and Words. However, in contrast, there is a slight positive jump for Sentences and there appears to be no difference for Letters (Figure 14).

Figure 15 presents the (intention to treat) p-values of a permutation test testing the difference in means for the overall score within each school zone, as well as the combined p-value from the stratified-by-zone permutation test (the rightmost cluster of bars). We see that for most school zones there is no significant difference between the students in the cutoff window who were assigned to the low stream and those who were assigned the upper stream. Further, for most school zones, the difference between high and low streams appears to increase (the p-value decreases) as we widen the window, however, we note that this is the expected result, since as we consider wider windows around the cutoff point, the difference in ability between the high and low stream students being considered increases. Thus it is surprising that for some zones, we see a decrease in difference (an increase in p-value) for example in Kabula and Matungu. The only zones for which we obtain a statistically significant difference is for the South Wanga and Sangalo school zone for the students in the 20-percentile width window about the cutoff point (the window consists of students who scored from the 40th to the 60th percentile in the baseline scores). However, this may simply be a reflection of the window being too large to still consider the students as randomly assigned to the high and low streams, and we may simply be detecting the fact that overall, the students in the high streams perform better than those in the low streams.

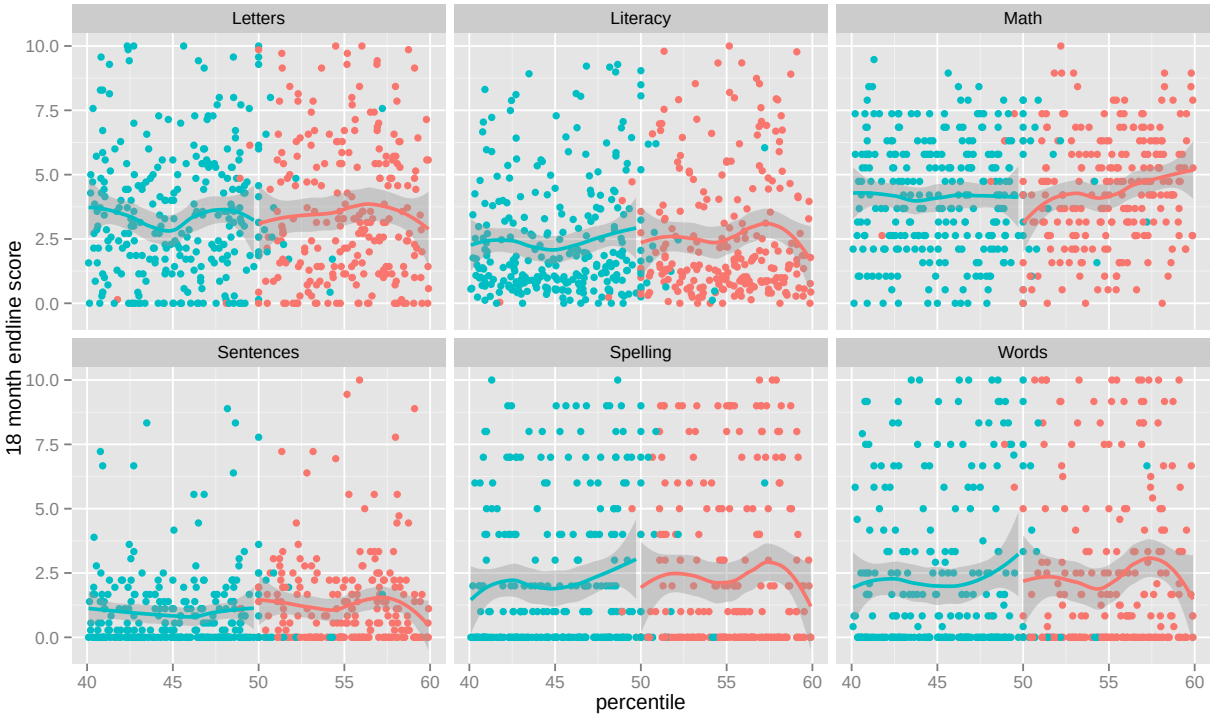

Figure 14: Scatterplot of baseline percentile versus the standardized18-month score for each subject within a window of width 10 percentile points (5 percentile points in each direction around the cutoff point) with loess fitted curves for the high and low stream.

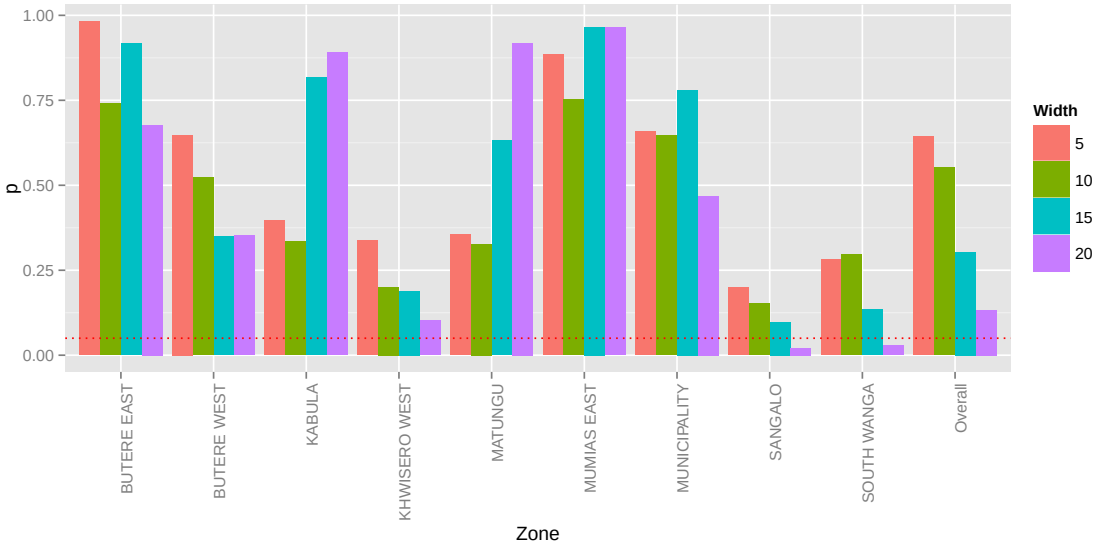

Figure 15: Bar plot of permutation test p-values for differences in means between the low stream and high stream for windows of varying widths (5, 10, 15 and 20 percentile points) about the cutoff point of the 50th percentile. The 5-percentile width window contains a total of 137 students, the 10-percentile width window contains a total of 225 students, the 15-percentile width window contains a total of 405 students and the 20-percentile width window contains a total of 509 students.

6 Conclusion

We conclude that overall tracking, has a small but beneficial effect on students' learning. After a year in a tracking classroom, we see that the students tend to fare better than their non-tracked counterparts in most content-areas regardless of their placement into lower and upper streams. This positive relationship holds true for different types of teacher and for many of the school zones, though the magnitude of the effect varies. However it would seem that the beneficial effects of tracking for math quickly wears off if the program is discontinued as when students were tested again 6 months after the funding ended, only the value-added in literacy was still observable. conclusions about regression discontinuity

Our findings are consistent with [Duflo et al., 2011]. Moreover, we are able to reproduce their results using non-parametric permutation tests, which make no assumptions about the distribution of noise and no assumptions about the functional relationship between tracking and exam scores. In contrast, the p-values that the authors report come from ordinary least squares regression and require normality assumptions that are difficult to check.

References

[Duflo et al., 2007] Duflo, E., Dupas, P., and Kremer, M. (2007). Peer effects, pupil-teacher ratios, and teacher incentives: Evidence from a randomized evaluation in kenya. *Unpublished*.

[Duflo et al., 2011] Duflo, E., Dupas, P., and Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya. *American Economic Review 101*.

[UNESCO, 2007] UNESCO (2007). Strong foundations: Early childhood care and education. *Paris: UNESCO Publishing*.