Economics of Education Review 56 (2017) 40–51

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

Heterogeneous treatment effects in the low track: Revisiting the Kenyan primary school experiment

Joseph R. Cummins

Department of Economics, University of California, Riverside, 900 University Ave., Riverside, CA 92521, United States

ARTICLE INFO

Article history:
Received 25 September 2015
Revised 19 November 2016
Accepted 22 November 2016
Available online 28 November 2016

Keywords:
Ability tracking
Human capital
Economic development

ABSTRACT

I present results from a partial re-analysis of the Kenyan school tracking experiment first described in Duflo, Dupas and Kremer (2011). My results suggest that, in a developing country school system with state-employed teachers, tracking can reduce short-run test scores of initially low-ability students with high learning potential. The highest scoring students subjected only to the tracking intervention scored well below comparable students in untracked classrooms at the end of the intervention. In contrast, students assigned to tracking under the experimental alternative teacher intervention experienced gains from tracking that increased across the outcome distribution. These alternative teachers were drawn from local areas, exhibited significantly higher effort levels and faced different incentives to produce learning. I conclude that although Pareto-improvements in test scores from tracking are possible, they are not guaranteed.

1. Introduction

A recent paper on the effects of school ability tracking by Duflo, Dupas, and Kremer (2011) (henceforth DDK) presents experimental evidence that tracking in Kenyan primary schools improved test scores in both the low-ability and high-ability tracks. DDK conclude that "students at all levels of the initial achievement spectrum benefited from being tracked into classes by initial achievement" (page 1768). The results in DDK constitute the strongest evidence available that tracking improves test scores for children of all ability levels. The results that I present, estimated from the same dataset, constitute the first experimental evidence that tracking in classrooms can lower short-term test scores for some students placed into the low-ability track.

It is no surprise that the DDK analysis, cited over 400 times, has been influential in policy discussion. School ability tracking has long been controversial, usually on grounds related to the distribution of the benefits of tracking. If the strategic distribution of students across classrooms can generate Pareto-improvements in test scores, it would be one of the most cost-effective educational reforms available. However, well-intentioned peer-sorting interventions do not always benefit students ex-post (Carrell, Sacerdote, & West, 2013). Standard economic models of peer effects predict that peer quality affects test scores, and tracking reduces the peer quality of those placed into the low track, thus potentially worsening their learning outcomes (Epple, Newlon, & Romano, 2002). Some non-experimental studies have found evidence that tracking harms low-ability students (Argys, Rees, & Brewer, 1996), although certainly not all studies on the topic (Figlio & Page, 2002). This previous literature relies almost exclusively on evaluating observational studies, and so causal inferences are open to the usual concerns over selection, omitted variables, and measurement (Betts & Shkolnik, 1999).[1] In this environment of uncertainty, DDK's experimental estimates are unusually influential.

DDK interpret the results of the experiment in the framework of an economic model of teacher behavior and child learning. The model incorporates standard mean-peer-quality models where placement in the low track can potentially reduce test scores through decreased quality of peer interactions. However, it also incorporates a decision-making teacher who responds to the ability distribution of their students by adjusting the level of ability to which they target their instruction and the effort they put into teaching. The pattern of heterogeneous treatment effects across pre-intervention test score (pre-score) is then used to infer a set of model parameters consistent with the data. They find relatively large gains for students placed into both the low track and the high track, arguing from this that any negative effect of the decrease in peer quality is offset by behavioral responses on the part of the teacher. They then interpret the null results of a regression discon-

[1] An exception to this is a study of tracking in South African dormitories that finds negative impacts of ability tracking among roommates Garlick (2016), but this is not a classroom intervention.

E-mail address: joseph.cummins@ucr.edu

http://dx.doi.org/10.1016/j.econedurev.2016.11.006
0272-7757/

tinuity across the tracking threshold for pre-score as evidence of teachers targeting their effort towards the top of the within-class ability distribution.

While the economic questions posed by DDK regarding teacher behavioral decisions may be more properly investigated by analyzing heterogeneity in treatment effect across pre-score, there are at least three reasons why the policy question about the value of tracking may be better answered by considering effects across the endline distribution. First, in terms of pure measurement, pre-scores were based on grades from teacher-written tests conducted after 6 months of first grade (Duflo et al., 2011). They are not directly comparable across schools and they likely do not measure a consistent set of skills. In contrast, endline test scores come from standardized tests specifically designed to gauge student learning, were scored by independent graders and are fully comparable across schools. They are much more compelling measures of ability at endline than the pre-scores are measures of ability at baseline. Second, if welfare weights across children are unrelated to pre-intervention ability, then the ex-post distribution of test scores is the relevant measure for policymakers. That is, if policymakers care about the students produced under tracking, as opposed to the students being placed into tracking, then the appropriate counterfactual thought experiment is to compare the distribution of test scores created under tracking to an alternative assignment rule (in this case, random assignment of peers). Third, unlike heterogeneous treatment effects on wages or wealth, which can lead to Pareto improvements in welfare via ex-post targeted transfers, test scores cannot be redistributed across students (Heckman, Smith, & Clements, 1997).

None of these arguments mean that heterogeneity in treatment effects is unimportant or uninteresting. Policy makers have preferences over average scores, but they may also have preferences over tradeoffs between average scores and inequality, or they may put added weight on one of the tails. However, if policymakers do not have preferences over any particular child ex-ante, these tradeoffs relate to comparing the outcome distributions, and not effects across pre-score.

I re-examine the effects of tracking in the Kenyan primary school experiment, but focus on effects across the endline test score distribution. Using quantile treatment effects (QTE) estimators I show that, absent any additional teacher intervention, the highest scoring students placed in the low-ability track scored between 0.35 and 0.45 standard deviations (sd) below the highest-scoring students in the associated comparison group at the e*nd* of the intervention. While there are gains in the middle of the distribution (0.17 sd at the median), point estimates go to 0 around the 80th percentile and are negative and mostly decreasing from the 90th to the 99th percentiles.

I provide some evidence that the difference between the DDK analysis and my own is caused by differential churning of ability ranks induced by tracking. If treatment induces rank change, then the QTE at the 95th percentile does not identify the effect on a person who was in the 95th percentile at the beginning of the program. In the case of test scores in this experiment, the strict rank preservation assumption is not applicable – there are clear changes in test score ranks across rounds. However, since test scores are noisy measures of underlying ability, a more useful thought experiment is to consider rank-similarity. Rank similarity is an assumption about the equal distribution of potential ranks, not realized ranks, across treatment groups (Dong & Shen, 2016).

If test scores are noisy measures of a stable, underlying ability or skill measure, then rank invariance in test scores is likely to be violated, but rank similarity may not be. Empirically, I test whether the distribution of potential ranks for a student with similar pre-scores and observable characteristics is the same in both the treatment and control groups. I provide some evidence that tracking induces differential rank change, rejecting the null hypothesis (at $p < 0.10$) of rank similarity between tracking and control schools on some, but not all, specifications of the test. These tests to tend to reject rank similarity in the middle and upper part of the test score distribution when testing rank similarity among demographic subgroups, in particular those related to student age. I also provide a placebo test (comparing endline and followup scores, when no treatment induced rank change would be possible) and the placebo test fails to reject for any specification.

The main results I focus on (those described above) come from students in classrooms taught by standard Kenyan civil service teachers and are limited to students who were either placed into the low-track or would have been placed into the low track had their school been tracked (they had a pre-score below the in-class median). Researchers and policymakers ought to be especially interested in this group. Low-ability students are usually considered the group in danger of being harmed by tracking, since under the practice they are separated from, and thus cannot learn from, high-ability peers. The focus on students with civil service teachers emphasizes the ceteris paribus effects of instituting tracking as a stand-alone public policy program absent additional alterations to the learning environment.

However, these students comprise only half of the students in the full experiment. Prior to the experiment, all schools had only one classroom. In order to staff the new sections needed to track classrooms, a new "contract teacher" was hired at each school. These contract teachers were recruited and trained separately from the civil service teachers. According to DDK, they exerted much higher levels of effort, had significantly less experience, often came from local areas, and were not employed by (and did not enjoy the employment protections of) the state. In contrast to students of civil service teachers, students of contract teachers who were assigned to the low track experienced gains across the outcome distribution, up to between 0.4 and 0.5 sd for those in the far right tail.

In the absence of this additional intervention, my analysis suggests that tracking in Kenyan primary schools reduced the test scores of a fraction of initially low-ability students with high potential to learn in a mixed-peers environment. The generalizability of this result is unclear and my contribution to the literature is modest. I argue only that the Kenyan experiment does not provide compelling evidence that tracking is likely to generate Pareto improvements in test scores in contexts where teacher effort is low and incentives are misaligned to produce learning for low-ability students, a common but not universal feature of educational systems in developing countries (Chaudhury, Hammer, Kremer, Muralidharan, & Rogers, 2006). Policymakers with competing preferences over the outcome distribution of test scores are thus not freed from considering potential tradeoffs, with increased scores for many students potentially coming at the cost of decreased scores for a few.

2. Background

2.1. Intervention

The school reform program that both DDK and I analyze was designed specifically to test the effectiveness of student ability tracking and was implemented in public schools in Western Kenya. All students from 111 (60 tracking and 51 control)[2] schools were

[2] 10 control group schools are missing pre-score data, and thus cannot be used in my analysis because I cannot assign those students to the proper counterfactual group (I do not know which track they were eligible to be placed in). The regression analysis in DDK similarly drops these schools due to missing pre-scores, but there were in fact 61 control schools for which there are post-intervention grades.

enrolled halfway through first grade. All schools had only one first-grade section prior to the intervention, and this was increased to two sections in both treatment and control schools.[3] Students in control schools were randomly placed into one of these two sections. Students in treatment schools were assigned to either the high or low tracked section based on their pre-score (above/below the school median). Teachers (one regular civil service, one contract) were then randomly assigned to sections.

The distinction between types of teachers is important from both policy and theory perspectives. Civil service teachers face only minimal incentives to produce learning, and DDK argue that those incentives are tied to teaching the most gifted students. They write, "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. Since many students drop out before then, the teachers have incentives to focus on the students at the top of the distribution" (page 1740). Contract teachers, on the other hand, are presumed to exert the relatively high effort witnessed during the trial in hopes of becoming civil-service teachers and/or due to their closer connection to the local community. There is little reason to believe that they would continue to exert such strong effort if hired permanently as civil service teachers.

2.2. Revisiting DDK

Students in ability-tracked classrooms experience a different learning environment than those in non-tracked classrooms. They have a different composition of peers and a teacher optimizing their instructional level under different conditions. DDK provide a simple and elegant economic framework describing how these mechanisms operate. The model begins with a test score production function:

$$Y_{ij} = X_{ij} + f(\bar{X}_{-ij}) + g(e_j)h(X_j^* - X_{ij}) + u_{ij} \quad (1)$$

Student i in classroom j will achieve outcome score Y_{ij} that is a function of their baseline ability (X_{ij}), the mean ability of their peers ($\bar{X}_{-ij}$) from whom they learn directly but in a non-specific manner [$f()$], and their teachers, who choose a target ability level at which to teach (X_j^*) and a level of teaching effort (e_j^*). The closer a student's own ability X_{ij} is to X_j^*, the more the student will learn. Effort then magnifies this effect as a multiplier. Teachers choose X^* and e^* to maximize the function:

$$U(Y_1, Y_2, \ldots, Y_n) - c(e) \quad (2)$$

where $U()$ is a utility function defining preferences over students' outcome test scores. The scores of the students are constrained by the costliness of the effort, $c(e)$, required by the teacher to generate test score improvements.

This model nests the standard linear-in-peer-mean model (setting $f(\bar{X}_{-ij}) = \gamma * \bar{X}_{-ij}$), where mean peer quality affects student performance by changing the quality of interactions between student j and their peers. DDK investigate this directly using random variation in peer quality among control group students. They find that peer mean quality does matter for test scores, and that better peers lead to better grades. Thus, they argue that mean peer quality is part of the test score production function, raising the possibility that tracking can lead to lower test scores for students placed in the low track.

However, the model includes a second set of factors that can counter the effects of peer quality. Teachers can respond to tracking by altering both their target level of instruction and their effort level. They choose their target and effort levels to optimize their utility function over test scores. As stated above, DDK argue that there are theoretical and empirical reasons to believe that teachers set X^* near the top of the within-class ability distribution. Intuitively, they argue that this is in line with the incentives faced by civil service teachers. Empirically, they combine the results on peer mean quality above with the null results obtained in a regression discontinuity. Students on either side of the median pre-score perform equally well, despite the fact that the right-of-median student was assigned to higher-ability peers. Thus, the left-of-median student must have experienced either a teacher teaching more closely to their own ability level, or a teacher exerting significantly more effort.

One further conclusion, inferred from the results on teacher effort presented in DDK Table 6, is relevant to comparing QTEs across sub-groups. Contract teachers were significantly more likely to be in the classroom teaching than civil service teachers (45% of spot checks compared to 75%). Assignment to tracking had no effect on contract teachers. Civil service teachers responded to assignment to the high track with an increased teaching presence, but did not make effort responses to assignment to the low track (relative to their control school counterparts). In the context of this reanalysis, this shows that there is no reason to believe that civil service teachers or contract teachers adjusted their observable effort in response to assignment to the low track compared to similar teachers in untracked classrooms.

2.3. Rank similarity and the test score production function

There are two reasons why the analysis presented in DDK across pre-score, and the one provided here across endline score, may show different patterns of heterogeneity in treatment effects: (1) if rank similarity does not hold, the differences in our analyses may be the results of differences in the kinds of children who learn well in tracked environments and children who learn well with randomly assigned peers; (2) if rank similarity does hold, the differences could simply be the result of compression in test scores near the top of the distribution of students from tracking schools (relative to control schools), maintaining rank order but changing the shape of the distribution. If we can reject rank similarity, it would provide some evidence that the differences in our respective analyses are caused, at least in part, by differential learning patterns for children in tracked and untracked classrooms.

As stated previously, rank preservation is required to interpret the QTEs as indicating that a student at the τth percentile of the control group test score distribution would score δ higher or lower if they were placed into tracking, what is sometimes called the treatment effect on the τth percentile. We cannot know whether Student A at the 97th percentile of the control group distribution would be at the 97th percentile of the treatment group distribution had Student A been assigned a tracked classroom. However, rank change in test scores over time is all but assured in any educational intervention, even in the presence of homogenous learning. Consider the abstracted test score production function:

$$Y_{ij}^T = f^T(X_{ij}; X_{-ij}, e_j^*, X_j^*) + \epsilon_{ij} \quad (3)$$

There are several reasons why rank similarity may be violated under such a general model. First, children with different ability levels may learn differently under tracking, leading to a kind of mechanical violation of rank similarity wherein the differential treatment effects catapult students from one part to another part of the test score distribution. Second, if we consider the vector of X to include not just pre-score but other observable characteristics of children that are related to their learning ability, heterogeneous treatment effects by subgroup could generate rank similarity violations because children of the same pre-score but different observ-

[3] A further experimental treatment arm studied the effects of reduced class size and was reported in Duflo, Dupas, and Kremer (2015).

Table 1
Mean treatment effects by sub-group.

	(1)	(2)	(3)	(4)
	Full	Low	Low-civil	Low-con
	b/se	b/se	b/se	b/se
T	0.17**	0.14*	0.03	0.24**
	(0.08)	(0.08)	(0.09)	(0.10)
Obs	5269	2575	1208	1347
N_schools				
r2	0.24	0.09	0.09	0.10
Comparison	–	1=2	2=3	3=4
Chi2	–	4.79	5.57	5.97
Pval	–	0.09	0.06	0.05

Standard errors clustered at school level. Regressions include linear controls for age and pre-score and a dummy for gender. Column 1 includes a dummy for below median on pre-score.

able characteristics (e.g., gender or age) experience different effects from tracking. Finally, rank similarity could be violated due to unobservable differences in children's propensity to learn in different environments.

While I cannot investigate the last of these possibilities, the first two possibilities are empirically testable by comparing the distributions of observable characteristics across the distribution of endline scores. Intuitively, if $f^T()$ does not depend on treatment assignment T ($f^T() = f()$), the distribution of X across Y should be the same for both treatment and control, with some natural amount of rank-churning in both groups caused by the noisiness of the test score as an ability measure. That is, holding own and peer ability and optimal teacher responses constant, we still expect rank change in test scores (if not unobserved ability, knowledge and skill) due to the random ϵ component. In order to allow for this natural rank-churning caused by ϵ, I focus on testing not for "rank preservation" in test scores, but for "rank similarity". Rank similarity is an defined as the equality of the distribution of potential ranks, not observed ranks. Even in the absence of rank preservation, rank similarity may hold if the changes in rank across rounds are caused only by noise in the measurement of underlying ability.

However, if $f^T()$ varies by treatment status, then rank similarity may be violated. One potential consequence of this violation would be a change in the distribution of pre-scores and covariates (X) across endline scores (Y). People of the same ability and observable characteristics would learn differently in each treatment arm, and thus their scores would be distributed differently within their own treatment group at endline. Using this interpretive framework, I can test whether tracking changed the distribution of pre-score and demographic groups across the endline test score distribution using the test developed in Dong and Shen (2016).

2.4. Data

Data on students' grades, age and gender were collected at baseline, along with identifiers for teacher and school. The randomization produced good balance across treatment groups on student and classroom characteristics (see DDK Table 1). I confirm that means are reasonably well balanced across sub-experiments in Tables A.1 and A.2, though children in tracking schools (with either type of teacher) tend to be slightly older than those in the control schools. To look for differences in the distribution of covariates across initial ability, Fig. A.1 graphs the local mean student age and proportion female across pre-score by treatment group. The randomization procedure produced generally good balance on observables across the pre-score distribution within sub-experiments. Following DDK, I use within-school grade percentile as the measure of pre-score because these scores are the only ones available for all 111 schools.

After 18 months in the program, having remained with the same peers and the same teacher throughout, all children in the sample were given a standardized test comparable across all schools at the end of second-grade (endline). The test contained math and language questions of increasing difficulty levels. The tracking program was then ended, but students were tested once again one year after that (follow-up). Both endline and follow-up tests were graded blindly by research staff and not by the teachers themselves. It is unclear the extent to which students considered the exams to be important to their own interests and the level of effort students put into the exams is similarly unclear. The endline and follow-up scores are standardized against the mean and standard deviation of the control group, as they are in DDK. Treatment effects are thus interpreted as differences in units of control group standard deviations.

3. Empirical methods

In order to compare the analysis I conduct on the effects of tracking across the endline test score distribution with the pre-score based estimates from DDK, I first present a series of regression estimates of mean treatment effects that mimic those in DDK Tables 2 and 3. I depart from DDK by estimating the model separately for various sub-groups instead of via interactions, and by using actual class assignment instead of intention to treat as the definition of treatment group students. I do this for three reasons: first, this ensures the model compares the sub-group of treated students in question (those placed into tracking) with only the relevant control students whose sections were not ability-tracked; second, this more closely mimics the framework for the distributional estimators I employ next, which cannot be estimated from pooled sub-groups; and finally, it allows for more stable bootstrap estimates because there are no remaining clusters with only a single student in any treatment arm. Estimating the model separately for sub-groups makes no substantive impact on the results, and while p-values vary slightly (and criss-cross customary thresholds of statistical significance) across group definitions, there is no qualitative change to the results regardless of treatment group definition.

For the main subgroup of interest, I compare those students in low-track classrooms with civil service teachers to those students with civil service teachers who would have been placed into the low track had their school been randomly assigned tracking. More generally, a sub-group s can be thought of as all students with the same teacher type (or types) and the same ability level (or levels).

3.1. Mean effects

Following DDK I use ordinary least squares, clustering standard errors by school, to estimate:

$$Y_{ij}^s = a + \boldsymbol{X}_{ij}'\beta^s + \delta^s T_j + \epsilon_{ij} \quad (4)$$

where δ^s is the parameter of interest and in OLS regressions is just the coefficient on a treatment dummy. It is super-scripted with an s to indicate that it can be an ability- and teacher-type subgroup specific treatment effect. In an alternative model I allow the treatment effect to vary non-parametrically across pre-score using local linear regressions (as in Kremer, Miguel, & Thornton, 2009). The vector of covariates X includes age, gender and pre-score, but age and gender are excluded for the local linear estimates.

3.2. Quantile treatment effects

My preferred analysis of heterogeneous treatment effects uses unconditional quantile treatment effects. QTEs measure the vertical distance between the inverse CDFs of the treatment and control

group distributions at any given quantile. The method has been used in the past to estimate heterogeneous treatment effects of schooling in the context of evaluating Head Start (Bitler, Hoynes, & Domina, 2014) and highschool quality (Eide & Showalter, 1998), and in contexts where pre-intervention "ranking" data is unavailable or of insufficient quality (Bitler, Gelbach, & Hoynes, 2006). QTE estimates answer the question: how much higher (lower), in units of the outcome variable, is the τ'th percentile of the treatment group distribution than the control group distribution?

Formally, let $F_T^{-1}(\tau)$ represent the inverse CDF of the treatment group distribution, and $F_C^{-1}(\tau)$ the inverse CDF of the control distribution, and define $Y_{T,\tau}$ and $Y_{C,\tau}$ as the value of the outcome at the τ'th quantile of the treatment and control distributions respectively. The QTE estimate for the τ'th quantile is simply the vertical distance between the inverse CDF's: $Y_{T,\tau} - Y_{C,\tau}$. To estimate the QTEs, I use the quantile regression method proposed by Koenker and Bassett Jr (1978) and implemented in *R* via the "quantreg" command. Confidence intervals are generated using the wild gradient bootstrap method proposed by Hagemann (2016), which allows for clustering of errors within schools and for inference across the entire distribution.[4] The regression includes only a constant and an indicator variable for treatment status.

3.3. Rank similarity test

If the differences in the QTE estimates and local linear estimates are sufficiently driven by changes in ability rank generated under tracking, and if ability rank is captured at least partly by pre-score rank, then the distribution of pre-scores across endline score should be different in tracking and control schools. Similarly, if different types of children (based on observables) experience different effects from tracking, then we would expect these observable characteristics to be distributed differently across endline score in tracking and control schools. Either of these possibilities would be inconsistent with a null hypothesis of a similar distribution of potential ranks. It is important to note that, while rejection of the null hypothesis does in fact indicate a violation of rank similarity, failure to reject does not provide credible evidence that rank similarity holds. For example, if there are significant heterogeneities in the effects of tracking related to unobservable aspects of children (e.g., their underlying capacity to learn, personality traits, or learning styles), the test could easily fail to reject rank similarity despite the fact that rank similarity does not hold.

The rank-similarity test I utilize, proposed by Dong and Shen (2016), requires two input decisions from the researcher: a definition of sub-groups to test based on observable characteristics, and a vector of quantiles of the outcome distribution at which to test the distribution of those sub-groups. I use two types of observable characteristics. First, motivated directly by the theoretical model and discussion above, I divide each group up into 5 bins based on pre-score. This allows me to test directly whether pre-scores are distributed similarly across endline score in the two experimental groups. Second, motivated by the fact that different types of children, and different types within any pre-score group, may respond differently to tracking, I include 3 age groups, 2 gender groups and their interactions with pre-score group.. This generates somewhere between 2 and 15 sub-groups, leading to within-group sample sizes of between 40 and 300 per treatment group. I then test the equality of the distribution of these sub-groups across three sets of quantiles and at the mean. I test the middle of the distribution by testing at the 0.2, 0.4, 0.6, and 0.8 quantiles, and I then test the low and high deciles separately (0.1–0.5; 0.5–0.9 in steps of 0.1).

While the test itself remains generally un-tested, in the sense that it has not yet been widely employed in the literature, the experimental design does afford a way in which to test the test itself. I repeat the exercise above, replacing pre-score with endline score and endline score with follow-up score, conceiving of changes in ranks between endline and follow-up as a placebo test. Whatever rank change was induced by tracking is present in the endline score, and there was no difference in the schooling experiences of children between endline and follow-up across treatment groups. Supposing persistence in ability across rounds, we would expect this test to fail to reject if, in fact, the test is appropriate to the data at hand. Though this is an imperfect placebo test, failure to reject rank change between endline and follow-up (coupled with rejection of the comparison from pre-score to endline) would provide some reassurances that at least the test is not simply responding to the general noisiness of the particular measures used in this analysis or likely to always reject in this particular environment.

4. Results

4.1. Replication of mean effects

Regression estimates of mean treatment effects by sub-group are displayed in Table 1. Each column represents a regression on a different set of students: (1) all students; (2) all low-ability students; (3) low-ability students with civil service teachers; and (4) low-ability students with contract teachers. The table is a conceptual replication of Tables 2 and 3 in DDK and qualitatively replicates their findings.

Column 1 shows that, on average, tracking improved test scores by 0.17 sd, a relatively strong effect that is slightly larger than the estimate in DDK (0.139, DDK Table 2 column 1). Column two restricts the sample to initially low-ability students, and their scores on average improved to a similar degree, 0.14 sd (0.156 in DDK Table 2, column 3, sum of first two rows).

However, columns 3 and 4 reveal that the net positive effect for low-ability students is driven exclusively by gains under contract teachers. The point estimate for low-ability students with civil service teachers is only 0.03 sd, and that is not statistically different from 0, while the effects for contract teachers are large and significant at 0.24 sd.[5] These treatment effect estimates are comparable to those in DDK Table 3: 0.048 sd under civil service teachers and 0.255 sd under contract teachers.

DDK further explore heterogeneous effects across ability by interacting pre-score quartile with a treatment dummy and find no negatively affected group (DDK Table 2, panel A column 4). Alternatively, but in the same spirit, I explore the possibility in a continuous pre-score setting that is more closely comparable to the QTE estimates that follow. The results are graphed in Fig. 1 for both of the low-track eligible sub-experiments, with pre-score on the *X*-axis and treatment effect on the *Y*-axis. The treatment effect under civil service teachers hovers close to 0 across the entire pre-score distribution, while the graph for contract teachers shows statistically significant gains at the top and bottom of the pre-score distribution.

[4] I thank Andres Hagemann for providing the *R* code to estimate these confidence intervals. I have also used the cluster robust methods proposed in Parente and Santos Silva (2013) and implement in the Stata package qreg2, (Machado, Parente, & Silva, 2014). The point estimates and *p*-values are similar. Point estimates for the QTEs are also easily confirmed by manually computing the individual percentiles using Stata's "summarize, detail" command and subtracting the treatment and control group estimates at each quantile. Alternative bootstrap and randomization test approaches produce similarly sized *p*-values and confidence intervals.

[5] A seemingly unrelated regressions test rejects that low-ability students with regular teachers did as well as those with contract teachers under tracking, shown in the lower rows of Table 1.

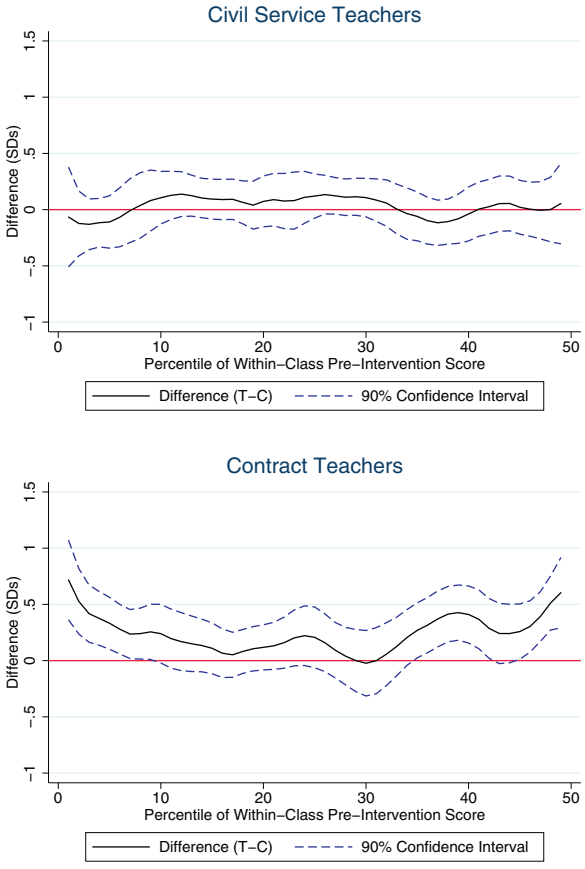

Fig. 1. Mean effects across pre-score.

4.2. Distributional effects

The unconditional QTE estimates (*Y*-axis) at each outcome percentile (*X*-axis) are graphed in Fig. 2. Most percentiles of the treatment group distribution are at higher levels than the corresponding percentiles of the control group distribution, up through approximately the 80th percentile. The QTE at the 50th percentile (a median regression estimate) is 0.17 sd and significantly different from 0. The QTE estimates for the higher outcome percentiles, though, decline to 0 around the 80th percentile and decrease rapidly after the 90th percentile. The QTE estimates for the 97th and 98th percentiles are −0.42 and −0.36 sd, respectively, with the latter being statistically significant at the 90% level. This can be directly interpreted as saying that children at the 98th percentile of the tracking group test score distribution scored more than 0.3 sd lower than children at the 98th percentile of the control group distribution. The conclusion that tracking was not Pareto improving on test scores follows directly from this result. Whoever the highest scoring children in the control group are, they would have scored lower had they been placed into tracking (since no one in the tracking group scored as high as they did).

The bottom panel of Fig. 2 shows the QTE estimates for low track eligible children assigned to contract teachers. In contrast to the civil service teacher arm, the QTE estimates here show that initially low-ability students test scores increased across the outcome distribution, up to around 0.5 sd near the top. This result is fully consistent with the possibility that tracking generated Pareto improvements in test scores for children assigned to contract teachers. Furthermore, the improvements in test scores under contract teachers strongly support the underlying argument in DDK that peer mean effect models without teacher behavioral adjustments are insufficiently nuanced to predict the effects of ability tracking.

4.2.1. Effects by subject and difficulty level

A section of DDK not yet discussed examines the effects of the intervention on the specific test questions that children were more or less likely to answer correctly.[6] I perform a related exercise. Fig. 3 plots the subject-specific difference in scores between treatment groups, separately by teacher type, at each percentile of the total endline score distribution. A clear pattern emerges: the dynamics seen in the QTEs are the result of changes in language scores. Where the QTE estimates diverge from each other near the top of the unconditional outcome score distribution, the language scores for students at those percentiles diverge from the respective control group scores.

4.2.2. Follow-up test scores

The initial tracking intervention lasted 18 months. One year after the end of the program and their return to un-tracked classrooms, students were re-examined to test the persistence of the effects of the intervention. I test for the persistence of the effect in two ways. First, the top panel of Fig. A.2 in the Appendix shows QTE estimates using follow-up scores. The apparent negative effects of tracking on the distribution of outcome scores seen in the endline QTEs does not persist after the program ceased.

Second, I test directly whether the effects found at endline persist on those particular students who were at the top of their respective endline distributions. To do this, I rank students by their endline score percentile and estimate local effects across follow-up score in the same manner I estimated the effects at endline across pre-score. The results are displayed in the bottom panel of Fig. A.2. The graph for civil service teachers looks very similar to the QTE graph. Treatment effect estimates are increasing up to the middle of the distribution, then decrease and become negative for the students with the highest endline scores. This implies that the specific children at the top of the control group endline distribution continued to do better than those at the top of the treatment group endline distribution, even 12 months after the program ceased.

4.3. Rank preservation

I present the results from several specifications of the rank similarity test described above in the top half of Fig. 4. The *X*-axis shows the set of sub-groups whose distribution across *Y* I am testing, with the ability-related measures to the left of the vertical center line and the demographic-only measures to the right. The test based purely on pre-score does not reject the null hypothesis of rank similarity under any choice of test quantiles. However, when groups are defined to include either age or gender, some tests do reject the null hypothesis. The interaction of pre-score and age leads to rejection at the mean and when considering the entire width of the high-density part of the distribution (the 20–80th percentiles). The interaction of pre-score and gender leads to rejection at the high end of the distribution (50–90th percentiles). The purely demographic variables also tend to reject the null. Equality at the mean is rejected or borderline for all three demographic subgroup definitions, while testing the central parts and high end of the distribution leads to rejection for both age alone and the age-gender interaction. There is no clear evidence of violations of

[6] The test covers 7 difficulty levels: three in math (addition, subtraction, multiplication) and four in language (letter recognition, word recognition, spelling, sentence comprehension). At each percentile of the total outcome score (maintaining the *x*-axis from Fig. 2), I estimate the average subject-x-difficulty score separately for treatment and control groups using local linear regressions and graph the difference.

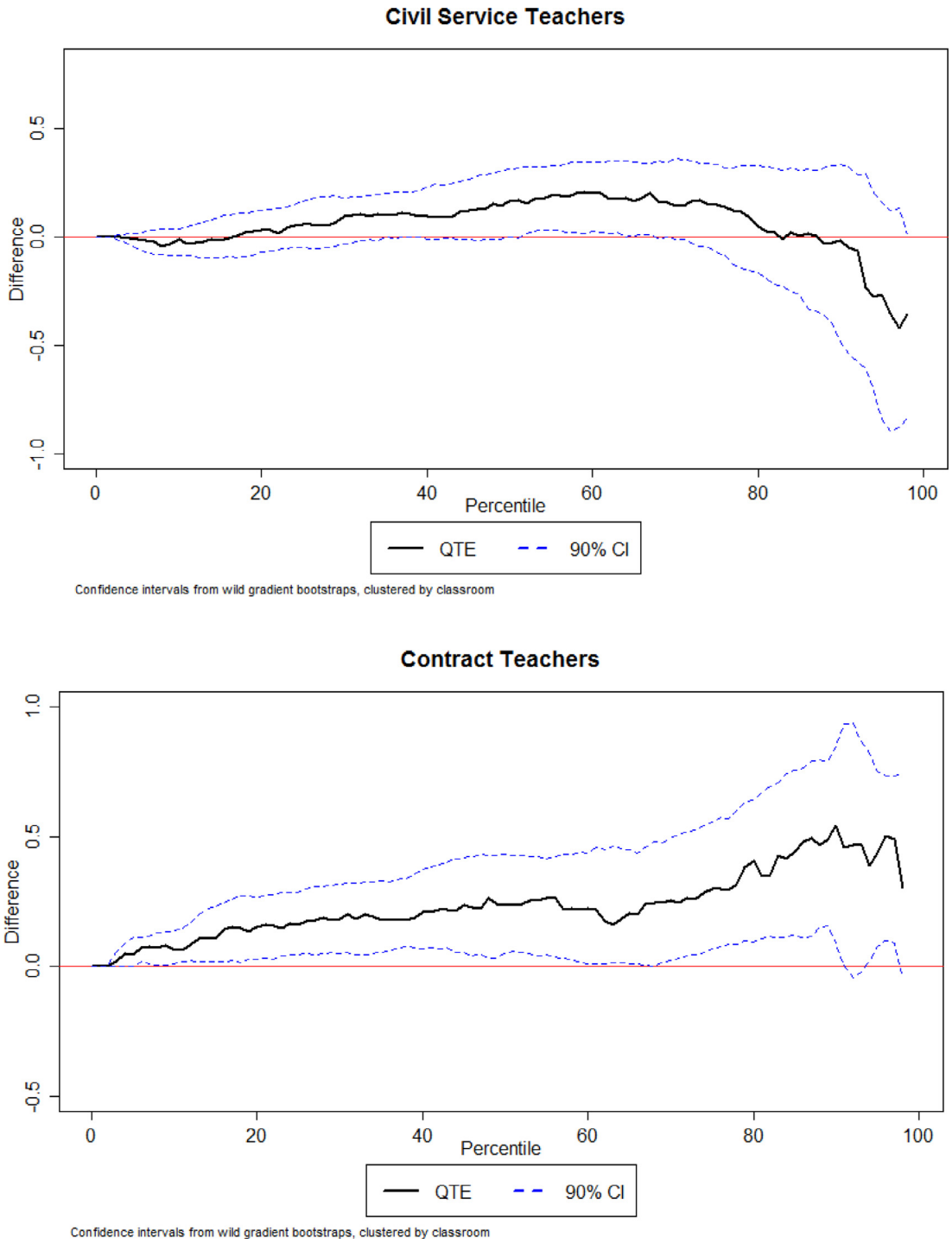

Fig. 2. Quantile treatment effects.

rank similarity at the low end of the distribution for any sub-group definitions.

It is reasonable to interpret these results with caution. First, the test rejects the null hypothesis only for certain combinations of variables at certain percentiles. Second, the test is new and we lack institutional knowledge in the field regarding its properties and the appropriate choices of test inputs. However, the thrust of the results seem to indicate that some differential rank-churning was induced by treatment, particularly in the middle and at the top of the outcome distribution. This differential rank churning seems to be less a function of initial ability itself than of observable characteristics of the students. That is, younger/older students, and male and female students, seem to have different response to being placed in the low track, a differential effect that is more powerful (statistically) than any rank change induced by differences in pre-score that are unconditional on demographic characteristics.

Several possible mechanisms for such a rank-similarity violation have been proposed in the broader education literature. Tracking could change students' perceptions of their abilities relative to their peers, leading them to re-adjust their efforts or alter their learning strategies (Bandura, 1993). Analogously, parental inputs may change in response to assignment to the high or low track

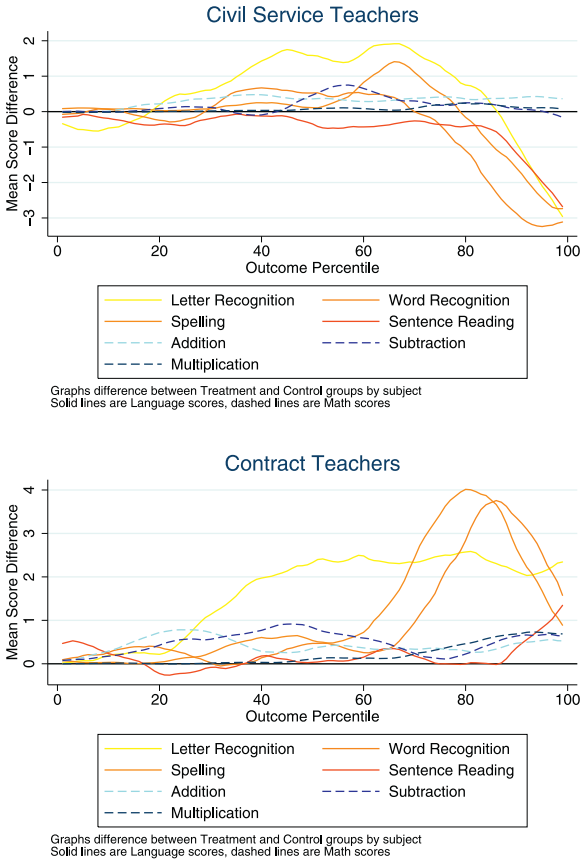

Fig. 3. Mean differences in subject scores across outcome distribution. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article).

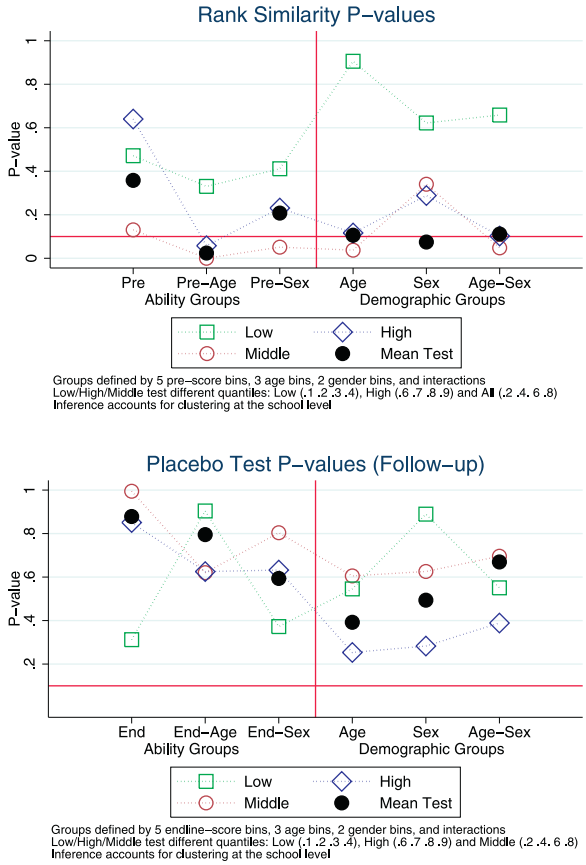

Fig. 4. *P*-values for rank similarity test. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article).

(Pop-Eleches & Urquiola, 2013). Children from different social or racial groups can also exhibit different learning trajectories and respond differently to the same environments; for instance, in the United States, black and white children display fundamentally different learning trajectories even controlling for initial ability (Fryer & Levitt, 2006), an effect attributed to both student and teacher behaviors (Ferguson, 2003). And of course, the nature of the test and the aspects of ability it measures can muddy the distinction between testing whether children changed ability ranks due to tracking, or just learned different things from the two different practices (Lord, 1952). While I cannot identify the mechanisms at play from among the candidate explanations, and though I cannot fully rule-out rank similarity, the evidence does raise the possibility that tracking may benefit different kinds of students than those who benefit from randomly-assigned peer classrooms.

In order to provide evidence that the test itself is working as it is supposed to, I repeat the exercise above, but replacing pre-score with endline score and endline score with follow-up score, conceiving of changes in ranks between endline and follow-up as a placebo test. Whatever rank change was induced by tracking is present in the endline score, and there was no difference in the schooling experiences of children between endline and follow-up across treatment groups. Thus, we would expect this test to fail to reject if, in fact, the test is working properly.

The bottom panel of Fig. 4 shows the results from the placebo rank test. With ability defined post-experiment, the results to the left of the vertical red center line have the interpretation of a (reasonably clean) placebo test. Those results based only on demographic characteristics, which may carry forward information on rank-churning induced by tracking since the group definitions have not changed, have a less clean interpretation, but I include them for completeness. Regardless, all tests in the bottom panel fail to reject the null hypothesis of rank similarity. The ones which come closest to rejecting are the tests based on unchanging demographic subgroups, and are likely carrying residual information from any rank change induced by tracking. The placebo tests based on comparing endline scores across follow-up all fail to reject, and only one test generates a *p*-value less than 0.4.

5. Limitations and robustness

I address three potentially important limitations to the analysis presented here: (1) the lack of a pre-specified analysis plan and the problem of ex-post sub-group analysis; (2) concerns regarding the control group; and (3) imprecision in the estimates of the QTEs near the top of the distribution.

First, this analysis was performed ex-post, without an explicit pre-analysis plan, and thus these findings are open to concerns of phishing, p-hacking and multiple comparisons. Phishing and p-hacking concerns can be alleviated to some extent by the public dissemination of all analysis code on the Open Science Framework (OSF).[7] I provide all replication code, written so as to be altered to check robustness in various ways. Furthermore, the processed data provided by DDK was not altered in any way and only minimal decisions about assignment rules were made, severely reducing "research degrees of freedom." Concerns about multiple comparisons are certainly valid, but are also assuaged to some degree by the nature of the comparisons undertaken; each of these sub-group comparisons are implicitly or explicitly made in DDK, and

[7] https://osf.io/4xncv/.

the experimental design is such that the disaggregations between high/low ability and civil service/contract teachers are natural to the data and the experimental manipulation.

Second, there are concerns about the pre-scores of students in the control group around the median. The regression discontinuity graph provided in the Corrigendum to DDK and reproduced using my sample definitions in Fig. A.3, shows that control group students just to the right of the median did worse than the students just to the left. This apparent discontinuity at the 50th pre-score percentile for the control group with civil service teachers is a failed placebo test since students on either side of the cutoff were assigned to the same classrooms, teachers and peers. In Fig. A.4, I present several robustness tests to test the sensitivity of my results to the control group students near the cutoff. Dropping the observations near the cutoff or swapping control group students between the 45th and 49th pre-score percentile of the control group with students from the 51st to 55th percentiles does not qualitatively affect the results. Restricting the sample to schools where full pre-scores (and not just percentiles) are available, or using an intentent-to-treat definition of treatment status (based only on pre-score and teacher type), also does not qualitatively change results.

Finally, reasonable readers may be concerned that the results for the QTEs near the top of the distribution are only "marginally significant", and certainly there are group assignments and sample selection choices that can make the effect "not significant" (see Appendix). These concerns are reasonable, but I argue they are misplaced. First, the group definitions and statistical inference methods presented here are quite conservative, and I encourage skeptical readers to evaluate the stability of the effect for themselves using the replication files provided on the OSF. Second, the argument provided in this paper that tracking alone did not generate Parteo improvements in test scores for low track students does not rest on the rejection of some particular right-tail percentile being statistically significantly below zero. The argument relies instead on the fact that the right-tail of the treatment group distribution, as a whole, is systematically shorter than the right-tail of the control group distribution. While I have not found a suitable method to directly test the equality of the tails, the pattern of the QTEs shows clearly that, in the sample from which DDK conclude universally positive effects from tracking, the highest scoring children in the low track under-performed relative to the highest scoring children with randomly assigned peers. I leave judgements regarding external validity to the reader.

6. Conclusion

My analysis is intended to highlight an effect of the Kenyan school tracking experiment that was overlooked in the original analysis presented by DDK. Both analyses conclude that tracking improved average endline scores among low track eligible students with contract teachers, but had no measurable effect on mean scores of students assigned to civil service teachers. However, my results suggest that the highest-scoring students assigned to civil service teachers in the control schools would have performed less well at the end of the intervention had they been placed into ability-tracked classrooms.

DDK provide convincing evidence that tracking need not be harmful to students placed in the low track. However, researchers and policy makers should not conclude from the Kenyan tracking experiment that the mechanisms through which tracking affects learning are always likely induce Pareto improvements in test scores.

Acknowledgments

I would like to thank Marianne Bitler, Scott Carrell, Esther Duflo, Pascaline Dupas, Andreas Hagemann, Hilary Hoynes, Michael Kremer, Mindy Marks, Doug Miller, Marianne Page, Shu Shen, Stephen Vosti, Brandon Wales and two anonymous referees for helpful comments and discussions. All mistakes, conceptual shortcomings, and methodological imperfections are my own.

Appendix A

This Appendix includes results from balance tests, follow-up test scores, and robustness checks.

A.1. Balance

Table A1
Balance – civil service teachers.

	Control	Tracking	Difference	*P*-value
Prescore	26.20	27.06	−0.86	0.15
	(0.50)	(0.33)	(0.60)	
Age	8.85	9.24	−0.39	0.00
	(0.09)	(0.10)	(0.13)	
Female	0.46	0.51	−0.04	0.14
	(0.02)	(0.02)	(0.03)	
SBM	0.52	0.62	−0.10	0.40
	(0.07)	(0.09)	(0.12)	
N	529	692	1221	

Standard errors in parentheses, clustered by school.

Table A2
Balance – contract teachers.

	Control	Tracking	(1) vs. (2)	*P*-value
Prescore	26.96	26.33	0.62	0.32
	(0.51)	(0.36)	(0.62)	
Age	8.91	9.16	−0.25	0.05
	(0.08)	(0.09)	(0.12)	
Female	0.45	0.49	−0.03	0.31
	(0.02)	(0.02)	(0.03)	
SBM	0.48	0.44	0.05	0.70
	(0.07)	(0.09)	(0.12)	
N	581	780	1361	

Standard errors in parentheses, clustered by school.

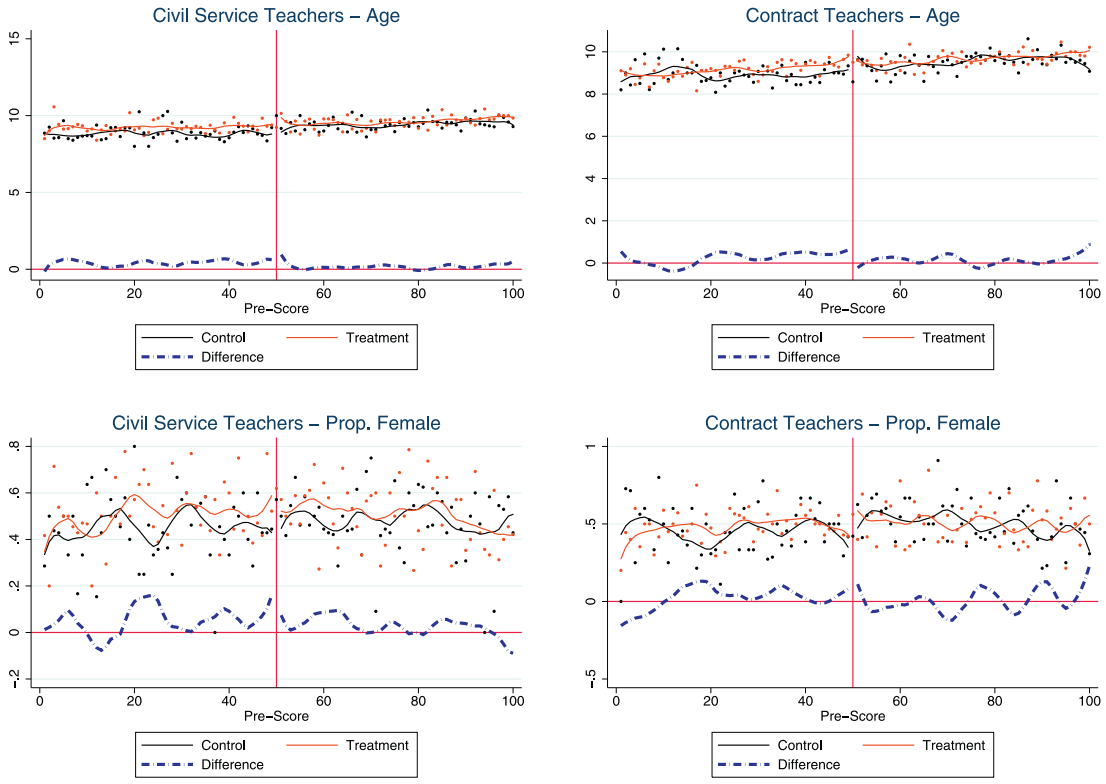

Fig. A1. Balance of covariates across pre-score.

A.2. Follow-up results

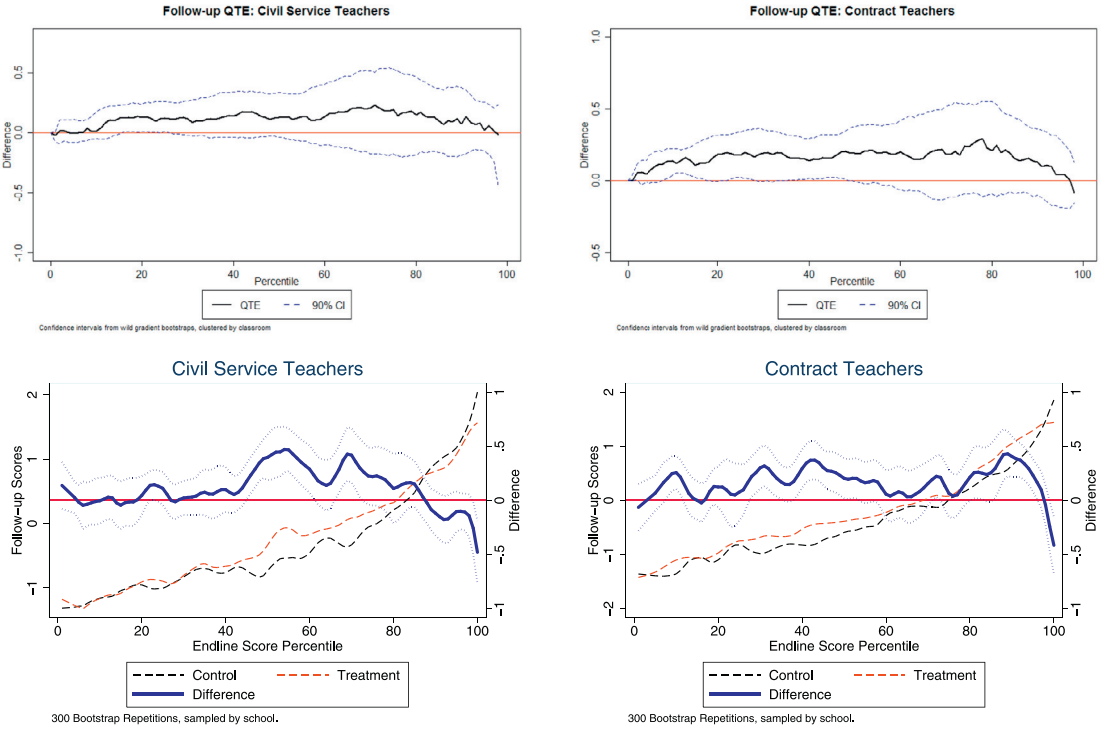

Fig. A2. Follow up scores (1 year later).

A.3. Robustness checks

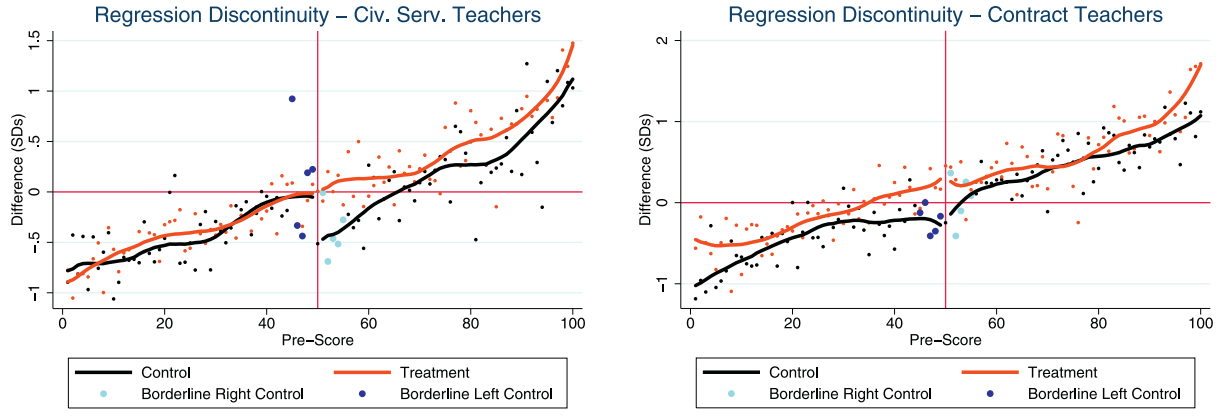

Fig. A3. Regression discontinuity.

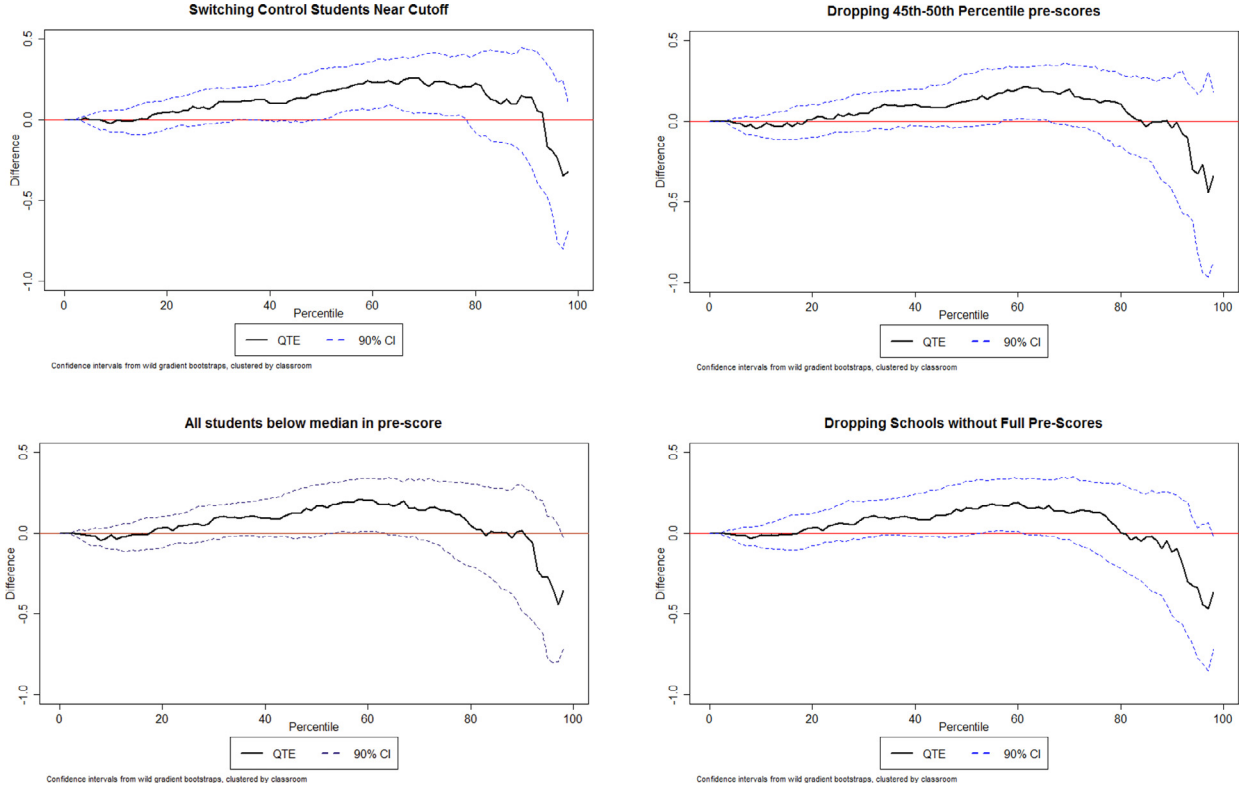

Fig. A4. Alternative sample/assignment QTEs.

References

Argys, L. M., Rees, D. I., & Brewer, D. J. (1996). Detracking America's schools: Equity at zero cost? *Journal of Policy Analysis and Management, 15*(4), 623–645. doi:10.1002/(SICI)1520-6688(199623)15:4⟨623::AID-PAM7⟩3.0.CO;2-J.

Bandura, A. (1993). Perceived self-efficacy in cognitive development and functioning. *Educational Psychologist, 28*(2), 117–148.

Betts, J. R., & Shkolnik, J. L. (1999). The effects of ability grouping on student achievement and resource allocation in secondary schools. *Economics of Education Review, 19*(1), 1–15.

Bitler, M. P., Gelbach, J. B., & Hoynes, H. W. (2006). What mean impacts miss: Distributional effects of welfare reform experiments. *American Economic Review, 96*(4), 988–1012.

Bitler, M. P., Hoynes, H. W., & Domina, T. (2014). Experimental evidence on distributional effects of head start. *Technical Report*. National Bureau of Economic Research.

Carrell, S. E., Sacerdote, B. I., & West, J. E. (2013). From natural variation to optimal policy? The importance of endogenous peer group formation. *Econometrica, 81*(3), 855–882.

Chaudhury, N., Hammer, J., Kremer, M., Muralidharan, K., & Rogers, F. H. (2006). Missing in action: Teacher and health worker absence in developing countries. *The Journal of Economic Perspectives, 20*(1), 91–116.

Dong Y. Shen S.,Testing for rank invariance or similarity in program evaluation: The effect of training on earnings revisited 2016.

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American Economic Review, 101*(5), 1739–1774.

Duflo, E., Dupas, P., & Kremer, M. (2015). School governance, teacher incentives, and pupil–teacher ratios: Experimental evidence from Kenyan primary schools. *Journal of Public Economics, 123*, 92–110.

Eide, E., & Showalter, M. H. (1998). The effect of school quality on student performance: A quantile regression approach. *Economics Letters, 58*(3), 345–350.

Epple, D., Newlon, E., & Romano, R. (2002). Ability tracking, school competition, and the distribution of educational benefits. *Journal of Public Economics, 83*(1), 1–48.

Ferguson, R. F. (2003). Teachers' perceptions and expectations and the black–white test score gap. *Urban Education, 38*(4), 460–507.

Figlio, D. N., & Page, M. E. (2002). School choice and the distributional effects of ability tracking: Does separation increase inequality? *Journal of Urban Economics, 51*(3), 497–514.

Fryer, R. G., & Levitt, S. D. (2006). The black–white test score gap through third grade. *American Law and Economics Review, 8*(2), 249–281.

Garlick, R. (2016). Academic peer effects with different group assignment policies: Residential tracking versus random assignment. *Economic Research Initiatives at Duke (ERID) Working Paper, (220)*.

Hagemann, A. (2016). Cluster-robust bootstrap inference in quantile regression models. *Journal of the American Statistical Association*, 1–30 just-accepted.

Heckman, J. J., Smith, J., & Clements, N. (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies, 64*(4), 487–535.

Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society, 1*, 33–50.

Kremer, M., Miguel, E., & Thornton, R. (2009). Incentives to learn. *The Review of Economics and Statistics, 91*(3), 437–456.

Lord, F. M. (1952). *A theory of test score (psychometric monograph no. 7)* (p. 35). Iowa City, IA: Psychometric Society.

Machado, J. A. F., Parente, P. M., & Silva, J. S. (2014). Qreg2: Stata module to perform quantile regression with robust and clustered standard errors. *Statistical Software Components*.

Parente, P. M., & Santos Silva, J. (2013). Quantile regression with clustered data. *Journal of Econometric Methods*.

Pop-Eleches, C., & Urquiola, M. (2013). Going to a better school: Effects and behavioral responses. *The American Economic Review, 103*(4), 1289–1324.