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.