**Pulling Out the Gifted can Benefit All: Theory and Empirical Evidence from a Randomized Experiment in the U.S.**[1]

Bradley Carl, Daniel Marlin, Cemile Sahin, Jennifer Vadas[2]

March 2019

**Abstract**

This paper presents a theoretical model of ability grouping and its empirical estimation using data from randomized control trial we conducted in a sample of U.S. middle schools. Starting with the premise that learning is a concave function of instructional level and that previous achievement and instructional levels are complements, we show that dividing students based on their ability raises not only total learning but also learning in both groups. Furthermore, we show that while lower ability students in the lower ability group and the higher ability students in the higher ability group gain from such split, students close to the cut-off ability level in both groups may lose. We empirically test these predictions using the randomized control trial we designed to evaluate the effectiveness of a gifted program in schools with disadvantaged populations. We show that in one semester, average growth in treated schools is more than double that in control schools. As predicted, this is due to large gains experienced by students at the top and bottom of the ability distribution. However, also consistent with our predictions, we find that students close to the cutoff ability level do not gain. Using the guidance of the theoretical model was key to unearth these heterogeneous treatment impacts.



[1] The funding for this project came from the U.S. Department of Education, Office of Elementary and Secondary Education, Jacob J. Javits Gifted and Talented Students Education Program. The implementation of the program and the data for this project was the product of a team effort at UW-Madison and one of our large urban partner districts. We are particularly indebted to the people implementing the program for always making considerable efforts to keep the rigor of the experimental design, which as we all know can be very challenging.

[2] Bradley Carl, University of Wisconsin-Madison WCER, bradley.carl@wisc.edu; Daniel Marlin, University of Wisconsin-Madison WCER, daniel.marlin@wisc.edu; Cemile Sahin, cemilexsahin@gmail.com; Jennifer Vadas University of Wisconsin-Madison WCER, jennifer.vadas@wisc.edu.

## I. Introduction

The reading scores of U.S. 15-year-olds in the 2015 PISA assessment show that students from disadvantaged economic backgrounds are almost five times less likely to excel in school and about three times more likely to fall behind[3]. The mismatch between student ability levels and the classroom instruction level could contribute to the problem. A way to reduce potential mismatches is to group students based on ability and adjust levels of instruction for each group accordingly. By lining up students' ability with instruction, ability grouping can help struggling students catch up and advanced learners stay engaged.

Nonetheless, some argue that ability grouping is detrimental to equity as students from lower socio-economic backgrounds will be sentenced to low ability groups and might never be able to catch up (See Oakes (2005) for a criticism of student grouping). The peer effects literature supports these claims by arguing that students are better off if their peers have higher average ability levels. Such argument implies that grouping students based on ability could benefit high ability students assigned to groups with high average ability and hurt lower ability students assigned to groups with low average ability. In other words, ability grouping is a zero-sum game where the higher ability students gain at the expense of the lower ability ones. The underlying assumption of this argument is that peer effects are monotonically increasing: the higher the ability of peers, the higher the academic performance regardless of distance between own ability and average ability. (See, Hoxby (2000), Lefgren (2004), Carrell et al. (2008) and Imberman et al. (2012).)

However, the empirical evidence does not support the assumption that peer effects are monotonically increasing. Carrell et al. (2012) finds negative and significant treatment effects on low ability students grouped with high ability students in a randomized experiment. Lavy et al. (2012a) finds that low ability students impose a negative externality on all students and finds no evidence of significant positive effects of average peer quality. Moreover, both Gibbons and Telhaj (2008) and Burke and Sass (2013), the former for British middle schoolers and the latter for students in Florida public schools, find that higher achieving students gain from an increase in average peer ability while low ability students are hurt by it. More directly in line with the ability grouping argument, Hoxby and Weingarth (2006) reports that students of all abilities in North Carolina benefit from being together with other similar students.

While these studies cast a doubt on the validity of positive and linear peer effects in developed countries, evidence from developing countries documents that there is little learning when there is a mismatch between the instruction level and the student ability level. For instance, Glewwe et al. (2009) argues that the distribution of free textbooks to Kenyan school children

[3] The 2015 PISA assessment reports 17.9 percent of advantaged students and 3.7 percent of disadvantaged students scoring at the top. In the low scorer category, the percentages are 9 and 28.9 respectively.

improved educational outcomes only for higher ability students. They argue that the gap between the lexical range of the books and the reading levels of lower achievers was too great to benefit them. Likewise, Pritchett and Beatty (2015) present simulation results and show that over-ambitious curricula relative to the level of student ability harms the prospects of human capital accumulation in developing countries.

Additionally, other studies in developing countries provide evidence of the effectiveness of "teaching at the right level." In a controlled experiment, Duflo et al. (2011) finds that students of all ability levels in tracked classrooms improve their academic achievement as compared to students in the controlled group with mixed ability classrooms. Muralidharan et al. (2016) documents the effectiveness of personalized computer-assisted learning to improve the achievement of students of all ability levels, but especially of the students in the bottom third of the ability distribution.

In order to make sense of these seemingly conflicting results, we provide a theoretical framework of ability grouping. We assume that individual learning increases with instruction level until a certain point beyond which learning decreases. We also assume that ability level and instruction level are complements. Thus, the optimal level of instruction, which maximizes learning, increases with the student's ability. Deviations from the optimal level, in either direction, will reduce learning. Based on these assumptions, we show that dividing students based on their ability raises not only total learning but also learning in both groups. Furthermore, we show that lower ability students in the lower ability group and higher ability students in the higher ability group gain from this split. Nevertheless, we also show that students close to the cut-off ability level in both groups may lose.

We test the model using a randomized control trial we conducted in a partner school district with large proportions of minority and economically disadvantaged students. The primary goal of the trial is to evaluate the effects of a high-quality gifted programming in writing and reading on students' academic performance. We selected sixteen middle schools and paired them based on students' observable demographic characteristics and achievement. In each pair, we randomly assigned one school to receive a challenging writing curriculum for its gifted sixth graders. Hence, treated schools implemented a two-groups-two-curricula setting during language arts while there was no such ability grouping in the control schools. We measure learning by taking the difference between the pre-test and post-test scores in a standardized language arts test for all students.

Our data source enables us to test our model because the educational system allows teachers to tailor instruction to their classroom make-up, and school effectiveness measures incentivize school staff to maximize growth in learning.

Our empirical results validate the predictions of the theoretical model. Average test score growth is more than double in treated schools than in control schools and the difference is statistically significant. Furthermore, both groups gain from the separation. As anticipated by our model, the most and the least advanced learners have greater gains in the treated schools than in the control schools. We believe this happens because departure of the more advanced students creates an opportunity for teachers to teach at a level that is more appropriate for less advanced students. However, also consistent with our predictions, we find that students close to the cutoff ability level, which are students in the upper end of the ability distribution within the less advanced learner group and at the lower end of the ability distribution within the more advanced learner group, do not gain.[4]

The organization of the paper is as follows. Sections 2 exposes the theoretical model, and section 3 explains the empirical strategy we use to estimate the size of the effects. Section 4 then describes the gifted program and the experiment designed for its evaluation. Section 5 presents the empirical results and section 6 contains concluding remarks.

## II. Theoretical Model

Few would argue against the idea that students gain the most when instructional levels are commensurate with their current achievement levels. Students with elementary reading skills will be ill-served if they are assigned to read Shakespeare, and those who are learning addition will gain little if they are exposed to integration. Conversely, students who excelled in their Language Arts classes can study literary texts, and those proficient in algebra can progress to calculus. Our theoretical model is based on this basic premise.

Let $a_i$ denote student i's current achievement level, which can be measured by current test score, S denote school instructional level, and $g(a_i, S)$ denote the achievement gains function[5].

The achievement gains function, $g(a, S)$, obeys certain rules in compliance with our basic premise. We assume that a) starting from no schooling, achievement gains increase from an increase in instructional level, b) the achievement level and the instruction level reinforce each other in achievement gains, and c) students' achievement gains first increase with the instruction level and then decreases. Going back to the earlier argument, students who just learned how to read will learn more if they are assigned chapter books; assigning readings from a literary journal will not help them as much.

Mathematically, we can express these three assumptions as follows:

***Assumption 1a:*** At S = 0, $\frac{\partial g(a,S)}{\partial S} > 0$.

***Assumption 1b:*** Student's current achievement level and the instruction level are complements in students' achievement gains: $\frac{\partial^2 g(a,S)}{\partial S \partial a} > 0$.

[4] The theoretical model predicted a loss for students around the cutoff point but the empirical evidence suggests a zero effect. This could be due to a smaller class size effect as well.

[5] Achievement gains are a function of other variables, such as student individual and family characteristics, previous achievement gains, and other school inputs. We suppress all these arguments to keep our notation simple.

***Assumption 1c:*** The achievement gains function is concave in S, $\frac{\partial^2 g(a,S)}{\partial S^2} < 0$. Thus, $\frac{\partial g(a_i,S)}{\partial S} > 0$ for $S < S_i^*$, and $\frac{\partial g(a_i,S)}{\partial S} < 0$ for $S > S_i^*$ where $S_i^*$ is the instruction level that maximizes achievement gains for student $i$.

These assumptions together imply that the higher the students' current achievement levels, the higher are their optimal levels of instruction. We formalize this in the following Lemma.

***Lemma:*** The optimal instruction level for a student, $S^*(a)$, is increasing in the student's achievement level: $\frac{\partial S(a)}{\partial a} > 0$. In other words, if $a_i > a_j$, then $S_i^* > S_j^*$ where $S_i^* = S^*(a_i)$ and $S_j^* = S^*(a_j)$ for simplicity of notation.

***Proof:*** Let $S_i^*$ be the curriculum level that maximizes learning for student i: $\frac{\partial g(a_i,S_i^*)}{\partial S} = 0$. Without loss of generality, assume that $a_i > a_j$ where $a_j$ is the current achievement level for student *j*. Given Assumption 3b, $\frac{\partial^2 g(a,S)}{\partial S \partial a} > 0$, student *j*'s achievement gain is decreasing in $S$ at $S_i^*$ : $\frac{\partial g(a_j,S_i^*)}{\partial S} < \frac{\partial g(a_i,S_i^*)}{\partial S} = 0$. Furthermore, given Assumption 3c, $\frac{\partial^2 g(a,S)}{\partial S^2} < 0$, the instruction level must be decreased for the first order condition $\frac{\partial g(a_j,S_j^*)}{\partial S} = 0$ to be satisfied. In other words, the optimal instruction level for student $j$, $S_j^*$, will be lower than the optimal instruction level for student $i$, $S_i^*$.

In the absence of fully personalized education programs in schools, a common instruction level must be set for a given group of students. Let $S$ be such common instruction level provided to all students in a classroom. In setting this instruction level, schools may have various objectives and focus on the lower or upper tail of the current classroom achievement distribution. In our analysis, we assume that schools aim to maximize the total achievement gains and set the instruction level accordingly. This assumption does not affect the essence of our results.

The total achievement gains are given by

$$G(S) = \int_{a_{min}}^{a_{max}} g(a,S) f(a) da$$

where $f(a)$ denotes the density function of $a$ on [ $a_{min}, a_{max}$].

The optimal instruction level, S*, is chosen to maximize $G(S)$, and hence it solves the first order condition:

$$\frac{\partial G}{\partial S} = \int_{a_{min}}^{a_{max}} \frac{\partial g(a,S^*)}{\partial S} f(a) da = 0, \qquad (1)$$

Note that the second order condition is satisfied due to Assumption 1c:

$$\frac{\partial^2 G}{\partial S^2} = \int_{a_{min}}^{a_{max}} \frac{\partial^2 g(a, S^*)}{\partial S^2} f(a) da < 0.$$

The resulting total achievement gain is

$$G^* = \int_{a_{min}}^{a_{max}} g(a, S^*) f(a) da.$$

**Gifted Intervention**

Now, assume that a school offers a gifted program for the $(1 - \pi)$ proportion students who had the highest achievement levels. Our aim is to analyze how this affects the achievement gains for the left-behind students as well as the achievement gains for the selected students. We also explore if the impact of the program on students depends on the interaction between the students' current achievement level and the group they are assigned to. We assume that the school adjusts the rigor of the gifted program to the achievement levels of their advanced learners.[6]

Let $\hat{a}$ be the cut-off achievement level so that $F(\hat{a}) = \pi$ where $F$ is the cumulative distribution function for the achievement level. Given $\pi$, and hence $\hat{a}$, the school sets its gifted instruction level, $S^*_{1-\pi}$ to maximize the achievement gains of their advanced learners. The optimal instruction level for the gifted program, $S^*_{1-\pi}$, solves:

$$\frac{\partial G^*_{1-\pi}}{\partial S} = \int_{\hat{a}}^{a_{max}} \frac{\partial g(a, S^*_{1-\pi})}{\partial S} f(a) da = 0,$$

and the total achievement gains for the selected students with the gifted program is:

$$G^*_{1-\pi} = \int_{\hat{a}}^{a_{max}} g(a, S^*_{1-\pi}) f(a) da.$$

We assume that in a responsive education system, teachers will adjust the instruction level as the distribution of student achievement changes. Thus, the instruction level for regular students will also adapt after introducing the gifted program. Let $S^*_\pi$ and $G^*_\pi$ denote the optimal instruction level for the regular students and the maximized total achievement gain for this group. The optimal instruction level for the regular students, $S^*_\pi$, solves:

$$\frac{\partial G^*_\pi}{\partial S} = \int_{a_{min}}^{\hat{a}} \frac{\partial g(a, S^*_\pi)}{\partial S} f(a) da = 0,$$

and the resulting total achievement gains for regular students are:

$$G^*_\pi = \int_{a_{min}}^{\hat{a}} g(a, S^*_\pi) f(a) da.$$

[6] The gifted program that we analyze in the paper adjusted the gifted instruction level to the student achievement levels.

Next, we want to compare schools with and without gifted programming, total aggregate learning gains $G(S^*)$ versus $G_\pi(S_\pi^*) + G_{1-\pi}(S_{1-\pi}^*)$, advanced learners' aggregate gains $G_\pi(S^*)$ to $G_\pi(S_\pi^*)$ and regular students' aggregate learning $G_{1-\pi}(S^*)$ versus $G_{1-\pi}(S_{1-\pi}^*)$.

We can rewrite the first-order-condition for the optimum common instruction level (Equation 1) as follows:

$$\int_{a_{min}}^{\hat{a}} \frac{\partial g(a, S^*)}{\partial S} f(a)da + \int_{\hat{a}}^{a_{max}} \frac{\partial g(a, S^*)}{\partial S} f(a)da = 0.$$

Given Assumption 1b, this implies

$$\int_{a_{min}}^{\hat{a}} \frac{\partial g(a, S^*)}{\partial S} f(a)da < 0,$$

$$\int_{\hat{a}}^{a_{max}} \frac{\partial g(a, S^*)}{\partial S} f(a)da > 0.$$

Together with Assumption 1c, this means that for regular students, the common instruction level is higher than the group's optimal instruction level, and for the gifted students the common instruction level is lower than the group's optimal instruction level. Therefore, we have the following ranking:

$$S_{1-\pi}^* > S^* > S_\pi^*.$$

The gifted program would be implemented to provide a higher level of instruction for the gifted students, and any adjustment for the regular students will lower the instruction level.

Since at the common instruction level $\int_{\hat{a}}^{a_{max}} \frac{\partial g(a,S^*)}{\partial S} f(a)da > 0$, given Assumption 1c, increasing the instruction level will increase achievement gains for selected students, materializing the objective of the program. However, same dynamics will also benefit regular students. At the common instructional level, $\int_{a_{min}}^{\hat{a}} \frac{\partial g(a,S^*)}{\partial S} f(a)da < 0$, and given Assumption 1c, reducing the instructional level increase achievement gains for regular students.

> ***Proposition 1***: The gifted programming will increase total achievement gains for both selected students, $G_\pi^*$, and regular students, $G_{1-\pi}^*$. The overall achievement gains, $G_\pi^* + G_{1-\pi}^*$, will therefore increase.

Note that this result does not depend on a specific cut-off point. Regardless of the percentage of selected students, 10% or 50%, adjustments in the instruction level will be beneficial for both groups of students. However, our result depends on the implicit assumption that educators respond to changes in the student body appropriately without taking resources away from one group of students to give it to others.

Proposition 1 states that the gifted programming will increase total achievement gains for both groups of students, but it does not suggest that all students will benefit. For some students, the adjustment in the instruction level due to the split may reduce achievement gains.

Let $\tilde{a}_\pi$ be the achievement level of the student who learns the same in the regular classroom with the gifted programming as under no gifted programming. Similarly, let $\tilde{a}_{1-\pi}$ be the achievement level of the student who learns the same in the gifted classroom as under no gifted programming.

$$g(\tilde{a}_\pi, S_\pi^*) = g(\tilde{a}_\pi, S^*)$$

$$g(\tilde{a}_{1-\pi}, S_{1-\pi}^*) = g(\tilde{a}_{1-\pi}, S^*).$$

Given Assumption 1b, $\frac{\partial^2 g(a,S)}{\partial S \partial a} > 0$, these imply that $S_\pi^* < S^*(\tilde{a}_\pi) < S^* < S^*(\tilde{a}_{1-\pi}) < S_{1-\pi}^*$.

Now consider students with $a_{1-\pi}^+ > \tilde{a}_{1-\pi}$ and those with $a_{1-\pi}^- > \tilde{a}_{1-\pi}$. Given Assumption 1b,

$$g(a_{1-\pi}^+, S_{1-\pi}^*) - g(a_{1-\pi}^+, S^*) > g(\tilde{a}_{1-\pi}, S_{1-\pi}^*) - g(\tilde{a}_{1-\pi}, S^*) = 0,$$

and

$$g(a_{1-\pi}^-, S_{1-\pi}^*) - g(a_{1-\pi}^-, S^*) < g(\tilde{a}_{1-\pi}, S_{1-\pi}^*) - g(\tilde{a}_{1-\pi}, S^*) = 0.$$

In words, the change in achievement gains in response to an increase in instruction level is higher (i.e., positive) for students with higher achievement levels than the student who is indifferent between the two instruction levels. Conversely, the change in achievement gains in response to the same change is lower (i.e., negative) for students with lower achievement levels. Thus, students with higher achievement levels will gain from being selected to the program while students with lower achievement levels may lose from the resulting increase in the instruction level. More specifically, if the cutoff ability is less than $\tilde{a}_{1-\pi}$, then students with lower abilities in the selected group will lose.

Similarly, for students with $a_\pi^+ > \tilde{a}_\pi$ and those with $a_\pi^- < \tilde{a}_\pi$ we can write

$$0 = g(\tilde{a}_\pi, S_\pi^*) - g(\tilde{a}_\pi, S^*) > g(a_\pi^+, S_\pi^*) - g(a_\pi^+, S^*),$$

and

$$g(a_\pi^-, S_\pi^*) - g(a_\pi^-, S^*) > g(\tilde{a}_\pi, S_\pi^*) - g(\tilde{a}_\pi, S^*) = 0.$$

In words, students with lower achievement levels will have higher achievement gains from the lower instruction levels after the departure of the gifted while students with relatively higher achievement levels may lose from this change. More specifically, if the cutoff ability is greater than $\tilde{a}_\pi$, then students with higher abilities in the remaining group will lose.

*__Proposition 2:__* The gifted program will improve achievement gains for relatively lower achieving students in the left-behind group and relatively higher achieving selected students in the gifted program. However, it may reduce achievement gains for relatively higher achieving left-behind students and/or relatively lower achieving selected students.

To summarize, a gifted program targeting students with higher achievement levels has the potential to raise achievement gains not only in the targeted group but also in the left-behind group. Left-behind is not left behind! Students in the middle who were better served in the absence of a gifted intervention, may lose from this intervention in one or both groups.

The next section discusses our estimation strategy.

## III. Econometric Model

In this section we develop the econometric strategy to estimate the effect of a school offering a gifted program on academic outcomes. We specify the academic achievement of a student at a given time as a function of current and prior experiences as well as the student's innate academic ability and inclination. The following equation presents a standard linear academic achievement model of student *i* at time *t*:

$$a_{i,t} = \lambda a_{i,t-1} + \mathrm{X}_i'(\beta_0 + \beta_t) + \alpha_i + \alpha_{if} + \alpha_{is} + u_{i,t}$$

where, $a_{i,t}$ and $a_{i,t-1}$ are student *i's* achievement level at time *t* and *t-1,* respectively, $X_i$ is a vector of observable characteristics for student *i,* $\alpha_i, \alpha_{if}, \alpha_{is}$ are fixed effects of the student, the student's family, and the student's school, respectively, and $u_{i,t}$ groups the combined unobservable factors. Factors such as academic intelligence, parental education, and quality of teachers can be sources of fixed effects.

Next, we introduce a gifted program in treated schools between time *t* and *t+1.* The academic achievement model of student *i* at time *t+1* becomes:

$$a_{i,t+1} = \gamma TREAT_i + \lambda a_{i,t} + \mathrm{X}_i'(\beta_0 + \beta_{t+1}) + \alpha_i + \alpha_{if} + \alpha_{is} + u_{i,t+1}$$

where, $TREAT_i$ is an indicator variable taking a value of one if the student is in a school offering the gifted program and zero otherwise.

Taking the first difference eliminates all time-invariant effects and yields:

$$a_{i,t+1} - a_{i,t} = \lambda(a_{i,t} - a_{i,t-1}) + \gamma TREAT_i + \mathrm{X}_i'(\beta_{t+1} - \beta_t) + u_{i,t+1} - u_{i,t},$$

which can be rewritten as

$$g_{i,t+1} = \lambda g_{i,t} + \gamma TREAT_i + \mathrm{X}_i'\mathrm{B}_{t+1} + v_{i,t+1}, \qquad (2)$$

where $g_{i,t+1}$ is student *i*'s academic achievement gains between period *t* and *t+1*, $\mathrm{B}_{t+1} = \beta_{t+1} - \beta_t$ and $v_{i,t+1} = u_{i,t+1} - u_{i,t}$. This specification allows us to control for pre-growth and eliminate all fixed effects with time invariant impacts.

Equation (2) is the general equation we use to estimate treatment effects; it asks whether students in treated schools improved more than in non-treated schools, relative to what would have been expected based on their characteristics and previous achievement levels. The theoretical model predicts that $\gamma > 0$.

We also predict that the gifted program will have a positive impact both for students who were selected to the program as a group and for those who remained in the regular classroom as a group. In order to test these predictions, we first divide students into two groups: one receiving the gifted program and the other continuing with the regular classroom. We then estimate the impact of the program in each group separately using the following specifications:

$$g_{i,t+1} = \lambda g_{i,t} + \gamma^{gifted} TREAT_i + \mathrm{X}_i' \mathrm{B}_{t+1} + v_{i,t+1}, \qquad (3)$$

$$g_{i,t+1} = \lambda g_{i,t} + \gamma^{remain} TREAT_i + \mathrm{X}_i' \mathrm{B}_{t+1} + v_{i,t+1}, (4)$$

where we expect that $\gamma^{gifted} > 0$ and $\gamma^{remain} > 0$.

Our theoretical model also suggests that not all students will gain from the intervention. In particular, whether a student will gain in either group depends on students' previous achievement levels. Among the advanced learners, the higher achievers will be more likely to benefit from the split. Conversely, among the remaining in the class group, the lower achievers are more likely to benefit. Moreover, we predict that students closer to the cutoff score to qualify for the program (lower achievers among the advanced learner group) may not benefit from the program, and those close to the cutoff score who do not qualify (high achievers among the remain in the class group) are likely to lose.

To check these predictions, we divide students within each group into subgroups based on their previous achievement levels. We run three separate regressions in the selected group, one for each subgroup $l \in \{low, medium, high\}$:

$$g_{i,t+1} = \lambda g_{i,t} + \gamma^{gifted,l} TREAT_i + \mathrm{X}_i' \mathrm{B}_{t+1} + v_{i,t+1}, \qquad (5)$$

where expect that $\gamma^{gifted,high} > 0$ and $\gamma^{gifted,medium} > 0$, but we are not sure about the sign of $\gamma^{gifted,low}$.

Likewise, we run separate regressions in the remain-in-the-class group, one for each subgroup $l \in \{low, medium, high\}$:

$$g_{i,t+1} = \lambda g_{i,t} + \gamma^{remain,l} TREAT_i + \mathrm{X}_i' \mathrm{B}_{t+1} + v_{i,t+1}, \qquad (6)$$

where expect that $\gamma^{remain,low} > 0$ and $\gamma^{remain,medium} > 0$, but we expect the sign of $\gamma^{remain,high}$ to be negative.

Before presenting the empirical results, we describe the gifted program, the empirical design, and the data in the next section.

## IV. Program and Experimental Design

The data comes from a randomized control trial (RCT) funded by a U.S. Department of Education grant. The goal of the RCT is to evaluate the effects of a gifted program on student academic achievement in typically underserved schools. The program consists of pulling out gifted students during their daily language arts class and providing them with a more advanced level of instruction. We designed our experiment and collected data in collaboration with our partner district and the program providers. The first of three cohorts started during the 2015-2016 school year and finished in the winter of the 2016-2017 school year. In this paper, we use the first (of two) semesters of data for the first (of three) cohorts.[7]

### i. The Gifted Education Program

The grant aims to provide gifted education programs to relatively higher achieving students in schools with large proportions of under-privileged populations. The gifted program selects students to raise their academic achievement and to "create a community of students who are confident and proud to be smart." It offers an advanced online course during the second semester of sixth grade. On every school day, selected students leave their language arts classroom to instead take their online gifted course in a designated area with a learning coach present.[8] The coach is a member of the school staff but is not required to be a certified teacher. The online classroom has a certified teacher who directs students, adjusts curriculum, and gives personalized feedback. Students thus receive a self-directed yet supervised and challenging online writing and reading program.

The students who are not selected were not targeted by the grant but are receiving an indirect treatment. Their classrooms become smaller in size and have a narrower range of prior achievement. The language arts teacher can readjust the instruction to the new range of ability. This is specific to the U.S. school context. Teachers craft their own curriculum during dedicated preparation times and are not tied to a rigid curriculum. Hence, the pull-out nature of the program, combined with a culture of responsive instruction, provide an externality on non-targeted students.

### ii. The Evaluation Design

The main research questions of the grant is: "Does providing gifted education to relatively high achieving students in underserved schools improve their academic achievement?" We chose

[7] The subsequent data sets will probably not be relevant for this analysis because of the "response to intervention" nature of the experiment. Results from this paper are already influencing the implementation of the program. Moreover, as explained later in the paper the standardized test was a new test, making it a good outcome measure.

[8] The program includes a learning coach in each site, which is not typically required.

a partner district with many such schools. We then worked closely with the district to collect data, choose schools, and design an experiment to identify and quantify the causal effects of the program.

### a. Selection of Schools within the District and Randomization Method

In implementing the program, we had two main constraints. First, the budget allowed for a maximum of ten middle schools in the district running the program. Second, we had to conduct a randomization of the treatment at the school level, not at the student level within schools. In setting up the experiment for the evaluation, we had to ensure a large enough and balanced sample. Hence, we targeted larger schools for sample size considerations, and we paired schools on similar observable characteristics at baseline for balance. Respecting those constraints, the description of the matching process is presented in appendix A.

After many iterations, nine school pairs remained in the sample. We then randomly assigned one school within each pair to the treatment (receiving the gifted programming) or to the control group. However, because one of the pairs was not adequately matched, we analyze impact using sixteen schools arranged in eight matched pairs.[9]

### b. Selection of Gifted Students within Each School

Once schools were chosen, we invited students into the gifted language arts program first based on their academic performance and then based on their teachers' inputs. We measured students' academic performance by computing their school percentile test score ranks for the spring of fifth grade and for the fall of sixth grade. We used two pre-tests instead of one in order to decrease the probability of excluding a student due to a bad day on a single test[10]. Students who scored at the seventy-fifth percentile or above in *either* test entered the list of potential candidates to invite into the gifted program. Then, sixth grade language arts teachers, in both treated and control schools, received the list of potential candidates along with a questionnaire to assess skills such as motivation and curiosity. Teachers had a few weeks to observe students in order to respond to the questionnaire. Based on these assessments, teachers then decided which students would benefit the most from the gifted program[11].

Next, we invited chosen students in treated schools to the gifted program. These students were required to obtain parental permission to receive the advanced online curriculum. Since the parental permission requirements cause attrition only in treated schools, we conduct an *intent-to-treat* analysis in the estimation of treatment effects.

[9] Qualitative results using 9 pairs hold but using 8 pairs is truer to the experimental nature of this data collection.
[10] Not all students had both tests available but we used both when available.
[11] For ethical reasons, teachers were allowed to pick from outside the list if they felt strongly that a student should be invited.

### c. Outcomes

The outcome of interest is learning gains measured by the difference between the post-test and pre-test scores in language arts adjusting for time elapsed between tests. The test is computer adaptive and measures basic competence skills taught in middle school. All schools in the district are required to use the test for all students which leads to great test coverage. The program was timed so that the pre-test and post-test windows respectively fell right before and right after the implementation. Furthermore, the test was administered for the first time in the first semester of our first cohort. Therefore, there was no teaching-to-test in regular classrooms putting them at equal footing with the students who were selected to the gifted program. In consequent semesters, the test became less valid because schools started to use it to tailor students' instruction. The gifted program, however, continued its own set of curriculum sensitive tests to measure progress and pays no attention to students' test scores in the district's standardized test.

Our dataset contains test scores for three consecutive periods within the school year (fall, winter, and spring). Each score is normalized based on the fall test scores mean and standard deviation of the control group.[12] Table 1 shows the pre-growth and pre-test distributions. We can see that the randomization was successful as all the standardized differences between the two groups much less than 20.

Note that, the gifted program providers have no incentives to "teach to the standardized test" like the rest of the school does. The gifted program has its own internal curriculum-sensitive testing structure to measure its own progress. Hence, gains in reading observed by students in the program are probably a lower bound because the gifted program has a strong focus on developing writing skills, which are not measured by the test. Curriculum sensitive writing pretests and posttests were supposed to be available for all top scorers in both treated and control schools, but many schools did not test their students.

### iii. Descriptive Statistics and Business Rules

There are 1,191 sixth graders who were registered for the whole year in one of the 16 schools. Teachers selected 224 students to invite into the gifted program, 115 in the control schools and 109 in the treated schools.

Table 1 and 2 below show the descriptive statistics of students in the treatment and control schools for the first semester of the program.

[12] All test scores are normalized for treated and control groups; for each score, we subtract the mean of the pretest scores of the control group and divide by the standard deviation. Hence, the mean and standard deviation of the control schools in the pre-test reading score has a mean of zero and a standard deviation of one.

**Table 1. Treated versus Control Pre-Tests and Pre-Growth Mean Differences**

| | Control | Treated | Standard Difference |
|---|---|---|---|
| Fall reading test | 0.000<br>N = 541 | -0.024<br>N = 535 | 3.431025 |
| Winter reading test | 0.190<br>N = 545 | 0.190<br>N = 530 | -0.04101 |
| Fall to Winter Growth | 0.186<br>N = 532 | 0.214<br>N = 521 | -5.9948 |

**Table 2. Treated versus Control Demographic Proportion Differences**

| | Selected Gifted | | | Remain in Class | | | All Students | | |
|---|---|---|---|---|---|---|---|---|---|
| | Treated N=109 | Control N=115 | Standard Difference | Treated N=491 | Control N=476 | Standard Difference | Treated N=600 | Control N=591 | Standard Difference |
| FRL | 0.688 | 0.730 | 9.33 | 0.792 | 0.765 | -6.65 | 0.773 | 0.758 | -3.61 |
| SpEd | 0.009 | 0.043 | 21.54 | 0.212 | 0.269 | 13.39 | 0.175 | 0.225 | 12.25 |
| ELL | | | | 0.177 | 0.113 | -18.18 | 0.150 | 0.091 | -18.06 |
| Female | 0.486 | 0.582 | 19.42 | 0.436 | 0.471 | 6.99 | 0.445 | 0.492 | 9.51 |
| Black | 0.248 | 0.261 | 3.03 | 0.363 | 0.294 | -14.60 | 0.342 | 0.288 | -11.68 |
| Hispanic | 0.486 | 0.435 | -10.33 | 0.477 | 0.511 | 6.78 | 0.478 | 0.496 | 3.50 |
| White | 0.202 | 0.226 | 5.93 | 0.096 | 0.158 | 18.69 | 0.115 | 0.171 | 16.02 |
| Other Race | 0.064 | 0.078 | 5.48 | 0.096 | 0.025 | -12.42 | 0.018 | 0.027 | -8.45 |

Table 2 shows that the randomization was successful on the observable demographic characteristics. Across the table, all variables but one have a standardized difference less than 20. The only one with a significant difference is the special education indicator in the advanced learner group, which has very few such students. However, we will include all observable variables in the regression because most are known to be strongly correlated to achievement gains causing unnecessary noise and potential omitted variable bias. Concerning the potential bias caused by possible imbalanced proportions of unobservable factors that can affect growth, we rely on the fact that the randomization worked well at balancing the most important variables (pretest scores and test growth the period preceding the experiment). The small standardized differences across the treated and control groups in what we can observe leads us to believe that the two populations are similar with a high probability. Moreover, our estimation strategy, taking the first differences in outcome, rid of all time-invariant factors.

## V. Results

We use three variations for each of the equations (2) to (6) to estimate treatment effects. The first variation (model 1) includes only an intercept and a treatment indicator, the second (model 2) adds pre-growth as a control, and the third (model 3) adds relevant demographic controls. Including more controls in the third model improves on the efficiency of the estimator and reduces the bias due to imbalances in the observable characteristics across treated and control units.

Our theoretical model claims that treatment effects are heterogenous within each of the two groups (selected and remain-in-the-class). Thus, we run regressions for each of the 10 groups as previously defined:

- All student as a whole group,
- the group of students selected into the gifted program,
- the relatively high achievers among the gifted (with a fall pretest percentile rank greater or equal to 90),
- the relatively medium achievers among the gifted (with a fall pretest percentile rank greater or equal to 80 but strictly less than 90),
- the relatively low achievers among the gifted (with a fall pretest percentile rank strictly less than 80),
- the remain-in-the-class group
- the high achievers among the remain-in-the-class (with a fall pretest percentile rank greater or equal to 75)[13],
- the relatively high achievers among the remain-in-the-class (with a fall pretest percentile rank greater or equal to 50 but strictly less than 75),
- the relatively medium achievers among the remain-in-the-class (with a fall pretest percentile rank greater or equal to 25 but strictly less than 50),
- the relatively low achievers among the remain-in-the-class (with a fall pretest strictly less than 25).

[13] This group of students was in the list to be invited by the teachers because they were in the top 25th percentile but did not end up being selected to the program. This group is included for completeness but is not as large as the other three groups within the remain-in-the-class group.

**Table 3. Treatment Effects on the Test Score Growth Regression Results**

| | Model 1 | Model 1 | Model 2 | Model 3 |
|---|---|---|---|---|
| | Intercept | Treatment Effect | Treatment Effect | Treatment Effect |
| All Students (N = 1,030) | 0.115<br>0.041 | 0.126848<br>0.053 | 0.132274<br>0.043 | 0.145896<br>0.007 |
| All Selected Students (N = 219) | 0.168719<br>0 | 0.122764<br>0.022 | 0.145221<br>0.013 | 0.153216<br>0.007 |
| Selected (D10) (N = 70) | 0.066884<br>0.304 | 0.338913<br>0.012 | 0.237881<br>0.05 | 0.285101<br>0.058 |
| Selected (D9) (N = 69) | 0.353432<br>0 | -0.05336<br>0.669 | 0.105846<br>0.396 | 0.12453<br>0.371 |
| Selected (<=D8) (N = 80) | 0.117698<br>0.033 | 0.066154<br>0.385 | 0.045475<br>0.541 | 0.035204<br>0.727 |
| All Students Remaining in Class (N = 811) | 0.100158<br>0.114 | 0.128161<br>0.077 | 0.129041<br>0.078 | 0.146082<br>0.02 |
| Remaining (Q1) (N = 242) | 0.087073<br>0.117 | 0.112973<br>0.11 | 0.135516<br>0.062 | 0.166988<br>0.047 |
| Remaining (Q2) (N = 255) | 0.058001<br>0.46 | 0.151104<br>0.13 | 0.145872<br>0.108 | 0.156098<br>0.079 |
| Remaining (Q3) (N = 212) | 0.106196<br>0.136 | 0.120038<br>0.245 | 0.10695<br>0.298 | 0.123284<br>0.157 |
| Remaining (Q4) (N = 78) | 0.368407<br>0 | 0.020785<br>0.85 | 0.07359<br>0.618 | 0.038915<br>0.79 |
| Pre-growth control | No | No | Yes | Yes |
| Demographic controls | No | No | No | Yes |

Note: Each cell of the table presents the estimates of the treatment effects from different OL.S. regressions with clustered standard errors at the school level estimated using a bootstrap method. The dependent variable in all regressions is the difference in normalized reading test scores between two time periods. We adjusted the growth taking the time elapsed between test into account. We divide growth by the date range between the post and the pretests and then multiply by the average date range of the whole sample. Each coefficient estimate has its corresponding p-value below.

### i. Direct effects on gifted students receiving the program

The theory predicts a positive overall impact on the learning gains of the gifted group. Table 3 shows that students selected into the program experienced extra growth in test scores ranging from 12.3 percent to 15.3 percent of a standard deviation. This result is statistically significant at least at the 5% level, with p-values ranging from 0.022 to 0.07 across models. The theory also predicts that the greatest gainers will be the students receiving the program who were the top scorers within that group. Hence, the highest scorers (defined as scorers in the top decile of their school) benefit greatly from the program by on average quadrupling their growth when compared to similar students in the control schools. The p-values in the three models for that group range from 0.012 to 0.058. These results suggest that providing the program to gifted students in historically underserved schools can lead to great learning gains despite starting points well below national gifted threshold scores and can lead to a narrowing of the excellence gap.

Our model also predicts that the gifted program is not particularly effective for students who are not at the top the test score distribution within the gifted group. The curriculum might be too challenging and the students who are not at the top of that group might lose interest despite downward adjustments reported by the program providers. The empirical results confirm this prediction. Students scoring on the lower end of those identified as gifted (with a test score percentile less than 80) do not seem to benefit from the program: the estimates of the treatment effects are much lower and lose statistical significance. Furthermore, predictions of the model are reinforced by exhibiting the opposite patterns we expected among the students who remained in regular classrooms.

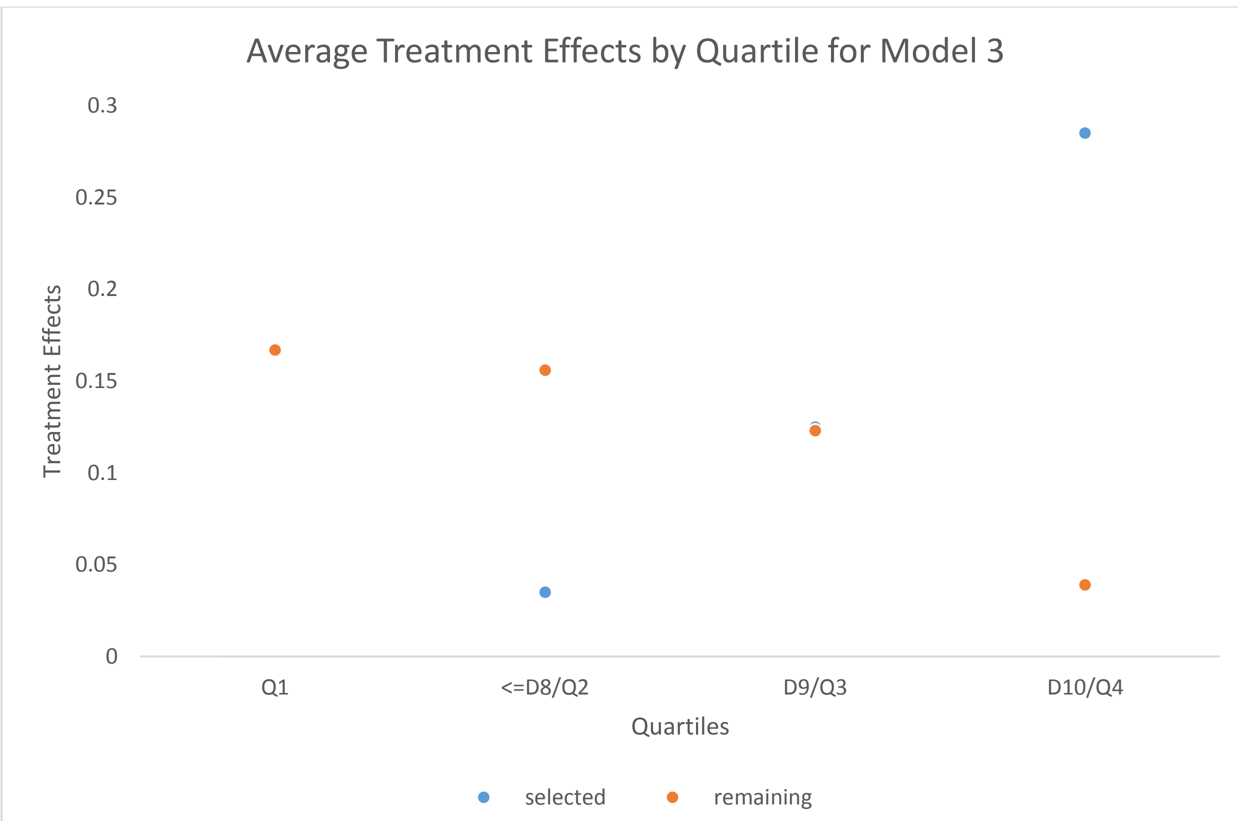
### ii. Indirect effects on students remaining in the classroom

Even if the students remaining in the regular classroom are not receiving a direct treatment by being invited in the gifted program, they are receiving an indirect treatment by remaining in a classroom with a narrower score range. The model predicts an overall gain for the "left-behind" group with heterogeneous impact. With a smaller spread in students' starting points, teachers can shift their instruction level downward to increase learning. In that group, the new level of instruction moves closer to the students in the lower part of the distribution but moves away from the students in the upper end of the distribution. Hence, for this group, we predict the greatest gains for the lower scorers and potential losses for the higher scorers.

Again, the empirical results confirm these predictions – students in treatment schools who remain in their regular classrooms experience between 12.8 percent and 14.6 percent of a standard deviation of extra growth relative to their control peers. This result is highly significant in the third model with a p-value of 0.02 and statistically significant at the 5% level in the first two models. Further, students in the first quartile had an extra growth ranging from 11.3 and 16.7 percent of a standard deviation, which is significant at the 5 percent level in the third model. Likewise, students in the second quartile experienced an extra growth of 15.6 percent of a standard deviation in model 3, significant only at the 10 percent level. Effects on students in the third quartile lose in significance and in size with an extra growth of 12.3 percent of a standard deviation in model 3, and a p-value of 0.079. Lastly, the top scorers remaining in the class have practically no gains, with estimates close to zero and very high p-values in all three models.

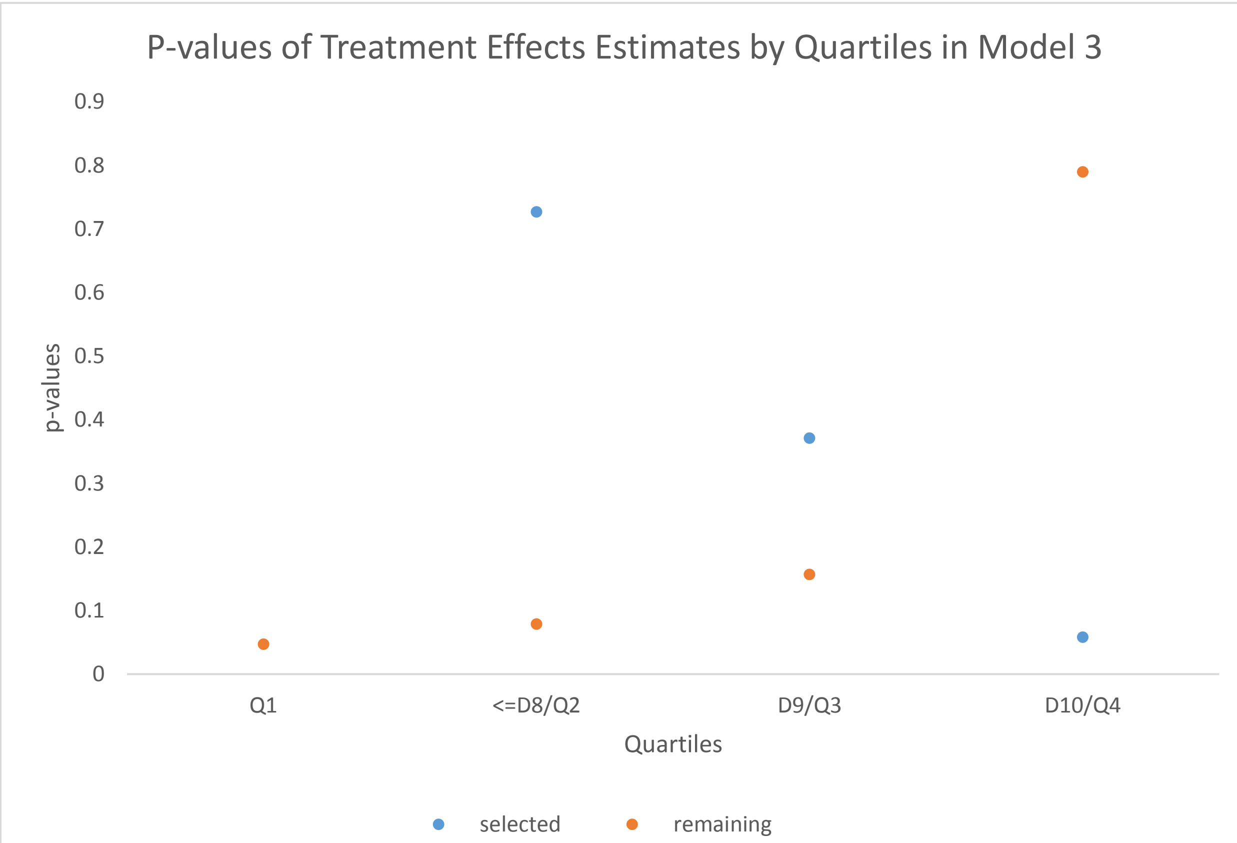
Figure 2 and 3 illustrate our results. The first figure depicts average treatment effects for each sub-group within the selected group and within the remain-in-the-class group. The figure clearly shows that in the selected group, average gains increase as the prior achievement increases, and in the remain-in-the-class group, the opposite happens. This is exactly what our model predicted. According our theory, this is due to a shift in curriculum in opposite direction across groups. The adjustment is downward for the remaining in the class group, thus getting closer to the lower ability students but further from the students in the upper category. Conversely, the adjustment is upward in the selected gifted group, getting closer to the highly gifted but further from the less gifted students.

**Figure 2. Treatment effects estimates for model 3 by pretests categories for selected gifted and remaining in class groups.**




In Figure 3, we depict the p-values associated with the treatment estimators in each sub-group within the selected group and remain-in-the-class group. Reinforcing our theoretical claims, we see that the p-values get smaller, thus showing increasing significance of the treatment effects for the students who get closer to the new curriculum levels in both groups.

**Figure 3. Treatment effects p-values for model 3 by pretests categories for selected gifted and remaining in class groups.**




Given the systematic confirmation of the theoretical predictions, both our heterogeneous students model where teachers maximize learning by choosing instruction levels given a range of ability and our data yield powerful results. For robustness check, we tested all the claims with a prior to treatment growth period and as expected, coefficients are all close to zero and none are statistically significant. Additionally, to ensure that no single pair of school is driving the results, we ran additional regressions dropping a pair at a time and drew the same conclusion for each sub-sample. Both checks increase our confidence in having a trustworthy randomization.

In sum, we provide evidence that teaching-at-the-right-level can be an effective strategy to maximize learning gains of students of all ability levels in a system where the instruction is responsive to classroom makeups. Our U.S. data source is proving a good setting for our empirical test because the educational system is very quick to respond to a new program, teachers modify curriculum continuously and are not tied to a rigid curriculum.

## VI. Conclusion

Our paper suggests that providing gifted programming to higher ability students in schools with high proportions of disadvantaged students can improve their academic achievement, thus potentially narrowing the excellence gap. At the same time, the gifted program helps students at the bottom of the ability distribution and has the potential to narrow the overall achievement gap. Indeed, though it intends to provide resources for gifted students in historically underserved populations, the gifted program is proving to benefit other students, weakening positively monotonic peer effects claims. U.S. datasets are more likely to be prone to show the effects predicted by the theoretical model because teachers routinely adjust their instruction levels to adapt to a new spread of ability. Group formation is fluid and less rigid than in some countries.

Although we document significant gains from the gifted programming, not only for higher-ability but also for lower-ability students, we would like emphasize that the program we study is not separation of students by schools or even by classrooms. It is a relatively low dose intervention, applied for a limited period of time, and thus, social interactions between students of all ability levels are not likely to be disrupted. Our results give a nod to limited, course-based ability grouping, but they do not necessarily extend to grouping students into separate classrooms for the entire day, or into separate schools. What we argue and find is the importance of the match between the student's ability level and the curriculum level, not that students of higher and lower abilities will be better off spending their time solely with other students like them.

## References

Ammermueller, Andreas and Jörn-Steffen Pischke, 2009, Peer Effects in European Primary Schools: Evidence from PIRLS, *Journal of Labor Economics* 27(3):315-348 · July 2009.

Angrist, Joshua, and Kevin Lang, 2004, "Does School Integration Generate Peer Effects? Evidence from Boston's Metco Program," *The American Economic Review*, Vol. 94, No. 5 (Dec., 2004), pp. 1613-1634.

Arnott, Richard, and John Rowse, 1987, "Peer group effects and educational attainment," *Journal of Public Economics* 32, pp. 287-305.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukherji, Marc Shotland, and Michael Walton, 2016, "Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of "Teaching at the Right Level" in India," NBER Working Paper No. 22746

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

Burke, Mary A. and Tim R. Sass, 2013, "Classroom Peer Effects and Student Achievement," *Journal of Labor Economics*, Vol. 31, No. 1 (January 2013), pp. 51-82.

Carrell, Scott E., Richard L. Fullerton, and James E. West, 2008, "Does Your Cohort Matter? Measuring Peer Effects in College Achievement," NBER Working paper No. 14032.

Cicala, Steve, Roland Fryer, and Jong Spenkuch, 2011, "A Roy Model of Social Interactions," NBER Working Paper No. 16880.

Cartwright, Nancy, and Angus Deaton, 2017, "Understanding and Misunderstanding Randomized Control Trials," NBER working paper No. 22595.

Duflo, Esther, Pascaline Dupas, and Michael Kremer; 2011, "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya," *The American Economic Review*, Volume 101, No. 5, August 2011, pp. 1739-74.

Epple, Dennis, and Richard E. Romano, 1998, "Competition between Private and Public Schools, Vouchers, and Peer-Group Effects," *The American Economic Review*, Vol. 88, No. 1 (Mar., 1998), pp. 33-62.

Epple, Dennis and Richard Romano, 2011, "Peer Effects in Education: A Survey of the Theory and Evidence," *Handbook of Social Economics*, Volume 1B, Chapter 20.

Epple, Dennis, Elizabeth Newlon, and Richard Romano, 2002, "Ability tracking, school competition, and the distribution of educational benefits," Journal of Public Economics, 2002, vol. 83, issue 1, pages 1-48. (The version I have is NBER Working Paper 7854, 2000).

Epple, Dennis, Richard Romano, and Holger Sieg, 2003, “Peer effects, financial aid and selection of students into colleges and universities: an empirical analysis,” *Journal of Applied Econometrics Special Issue: Empirical Analysis of Social Interactions*, 18(5), pages 501–525.

Finn Jr., Shester E. and Brandon L. Wright, “Failing our Brightest Kids, the Global Challenge of Educating High-Ability Students,” Harvard Education Press.

Gibbons, Stephen, and Shqiponja Telhaj, 2016, “Peer Effects: Evidence from Secondary School Transition in England,” *Oxford Bulletin of Economics and Statistics*, Volume 78, Issue 4, August 2016, pages 548–575.

Glewwe, Paul, Michael Kremer, and Sylvie Moulin, 2009, “Many children left behind? Textbooks and test scores in Kenya,” *American Economic Journal: Applied Economics*, 1 (1), 112–135.

Hoxby, Caroline, 2000, “Peer Effects in the Classroom: Learning from Gender and Race Variation” *NBER Working Paper* No. 7867, Issued in August 2000.

Hoxby, Caroline M., and Gretchen Weingarth, 2006, “Taking Race Out of the Equation: School Reassignment and the Structure of Peer Effects,” Unpublished manuscript; Presented at the 2006 AEA Annual Meetings.

Imberman, Scott, Adriana D. Kugler, and Bruce Sacerdote, 2012, “Katrina's Children: Evidence on the Structure of Peer Effects from Hurricane Evacuees,” *The American Economic Review*, Vol. 102, No. 5 (August 2012), pp. 2048-2082.

Jackson, C. Kirabo, 2010, “Do Students Benefit from Attending Better Schools? Evidence from Rule-based Student Assignments in Trinidad and Tobago,” *The Economic Journal* Volume 120, Issue 549, December 2010, pages 1399–1429.

Jacob, Brian A. and Lars Lefgren, 2004, “Remedial Education and Student Achievement: A Regression-Discontinuity Analysis,” *The Review of Economics and Statistics*, 86(1), February 2004, pp. 226-244.

Kaganovich, Michael and Xuejuan Su, 2016, “College Curriculum, Diverging Selectivity, and Enrollment Expansion,” CESifo Working Paper Series No. 6122.

Koppensteiner, Martin F., 2014, “Automatic grade promotion and student performance: Evidence from Brazil,” *Journal of Development Economics*, 2014, 107, 277–290.

Lavy, Victor, Olmo Silva, and Felix Weinhardt, 2012a, “The Good, the Bad, and the Average,” *Journal of Labor Economics,* Vol. 30, No. 2 (April 2012), pp. 367-414.

Lavy, Victor, M. Daniele Paserman, Analia Schlosser, 2012b, “Inside the Black Box of Ability Peer Effects: Evidence from Variation in High and Low Achievers in the Classroom,” *The Economic Journal*, Volume 122, Issue 559, March 2012, Pages 208–237.

Lazear, Edward, 2001, “Educational Production,” *Quarterly Journal of Economics* 116(3), 777-803.

Lefgren, Lars, 2004, “Educational Peer Effects and the Chicago Public Schools,” *Journal of Urban Economics* 56, pp. 169-191.

Levy, Victor, Olmo Silva, and Felix Weinhardt, 2012, “The Good, The Bad, and The Average: Evidence from Variation in Low Achievers in the Classroom,” *Journal of Labor Economics*.

Levy, Victor , D. Passerman and A. Scglosser, 2012, “Inside the Black of Box of Ability Peer Effects: Evidence from Variation in the Proportion of Low Achievers in the Classroom,” *Economic Journal.*

Manacorda, Marco, 2012, “The Cost of Grade Retention,” *Review of Economics and Statistics*, 2012, 94 (2), 596–606.

Muralidharan, Karthik, Abhijeet Singh, Alejandro J. Ganimian, (2016), “Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India,” NBER Working Paper No. 22923.

Pritchett, Lant and Amanda Beatty, 2015, “Slow down You’re Going Too Fast: Matching Curricula to Student Skill Levels,” *International Journal of Educational Development*, 40, 276–288.

Saavedra, Juan, Emma Näslund-Hadley, and Mariana Alfonso, 2017, “Targeted Remedial Education: Experimental Evidence from Peru,” NBER Working Paper No. 23050.

Sacerdote, Bruce; 2011, Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far?; *Handbook of the Economics of Education*, Elsevier, edition 1, volume 3, number 3.

Vardardottir, Arna, 2013, “Peer effects and academic achievement: a regression discontinuity approach,” *Economics of Education Review* Volume 36, October 2013, Pages 108–121.

Appendix A. Selection of schools within the district.

First, we used publicly available grade level school enrollment and demographic data from the 2014-15 school year to isolate schools that contained grades 6-8, and then added up all of the students in each school to get a total enrollment count. Second, we merged 2013-14 school performance information (the most recent at the time) with each school's enrollment data. Next, we separated the schools by the grades they contained.

To match the schools, we first sorted by total enrollment. From there, we matched the schools based on demographics and accountability measures. The enrollment data included percentages of students by race. The school performance data included percentages of economically disadvantaged students, students with disabilities, and students with limited English proficiency in each school, as well as each school's overall accountability score and rating.

We also needed to determine whether each school had a sufficient sample size of $6^{th}$ and $7^{th}$ graders. Using its universal screener (a computer adaptive test, at the time), the district typically identifies the top 10% of students as gifted/talented; therefore, we determined that any school with fewer than 50 students in both $6^{th}$ *and* $7^{th}$ grades (i.e. likely with fewer than 5 potential gifted/talented students) would be removed from the list of matched pairs. This initial process resulted in a set of 17 matched pairs (34 schools in total): three pairs of 6-8 schools, one pair of 6-12 schools, and 13 pairs of K-8 schools. One of the K-8 pairs was comprised of charter schools.

From there, we consulted with the district and the program staff regarding the viability of the schools we initially included. Those conversations had the following effects on the school selection:

- One 6-8 school we had selected had closed, and one school in the 6-12 pair was removed because it was a dual immersion school and thus had no reasonable comparison school. Initially, however, this only resulted in the loss of one pair, as the 6-8 school that lost its pair due to closure was paired with the middle grades in the remaining 6-12 school.
- Three schools (two of which were paired together) were receiving a grant that would have taken precedence over our program and may have confounded our program's effects with the other initiative's effects. Removing these schools resulted in elimination of two more pairs.
- District staff recommended matching two bilingual schools that were not initially paired (though this pair was ultimately removed at a later date). The bilingual schools were paired, and their original mates were then matched with one another.

At this point, 14 matched pairs remained (28 schools in total). From there, to get to our desired number of pairs, we removed four pairs of schools. The pair of bilingual schools was removed because its schools were not alike on achievement measures. Two additional pairs were removed because we did not believe there would be a sufficient number of gifted and talented students based on our analysis of test scores at those schools. A fourth pair was removed for similar reasons but kept as an alternate.

However, shortly before programming was to begin, district staff informed us that three of the schools on our list of 10 pairs were offering gifted and talented services through a new curriculum. The effects of this change on our school selection were as follows:

- One 6-8 middle school and the remaining 6-12 school were removed. The remaining four 6-8 schools were then re-matched with each other.
- One K-8 school was also eliminated, but we were able to pair its mate with another K-8 school that previously had been removed.
- We included the alternate K-8 pair referenced in the previous paragraph, which ultimately provided racial and ethnic diversity to our sample and returned our total number of pairs to 10.

Finally, the two schools in the charter school pair chose not to participate. Therefore, we ended up with 9 pairs (18 schools): two pairs of schools containing grades 6-8, and seven pairs of schools with grades K-8.