High-Dosage Tutoring and Reading Achievement: Evidence from New York City[·]

Roland G. Fryer, Jr.
Harvard University and NBER

Meghan Howard-Noveck
The Education Innovation Laboratory

June 2018

forthcoming in Journal of Labor Economics

Abstract

This study examines the impact on student achievement of high-dosage reading tutoring for middle school students in New York City public schools, using a school-level randomized field experiment. Across three years, schools offered at least 130 hours of 4-on-1 tutoring based on a guided reading model. We demonstrate that, at the mean, tutoring has a positive and significant effect on school attendance, a positive, but insignificant, effect on English Language Arts (ELA) state test scores and no effect on math state test scores. There is important heterogeneity by race. For black students, our treatment increased attendance by 2.0 percentage points (control mean 92.4 percent) and ELA scores by 0.09 standard deviations per year – two times larger than the effect of KIPP Charter Middle Schools on reading achievement. We argue that the increased effectiveness of tutoring for black students is best explained by the average tutor characteristics at the schools they attend.

[·] We are grateful to Lawrence Katz for helpful comments and suggestions. Financial Support from the Education Innovation Laboratory, the Ford Foundation, New York City Council, and The Robin Hood Foundation is gratefully acknowledged. ExpandEd Schools was an invaluable implementation partner. Adriano Fernandes, Blake Heller, Alex McNaughton, Adam Pfander, and Hannah Ruebeck provided terrific project management and research assistance. Correspondence can be addressed to the authors by mail: Department of Economics, Harvard University, 1805 Cambridge Street, Cambridge, MA, 02138 (Fryer) or Education Innovation Laboratory, 1280 Massachusetts Avenue, Cambridge MA, 02138 (Howard-Noveck); or by email: rolandfryer@edlabs.harvard.edu, mhoward@edlabs.harvard.edu. The usual caveat applies.

American fifteen year olds rank 20th out of 35 OECD countries in reading achievement according to the Program for International Student Assessment. The performance of black and Hispanic students on this assessment is just above the overall student performance of the worst two performing countries – Mexico and Turkey (OECD 2016). National Association of Education Progress scores paint a similarly bleak picture of American reading achievement – 34% of all eighth graders are at least "proficient" in reading; 16% (21%) of black (Hispanic) eighth graders score at this level (US DOE 2015).

While many interventions show promise in improving adolescents' math achievement, few interventions demonstrate significant impacts on reading achievement after elementary school.[1] For instance, high-performing charter middle and high schools consistently increase math scores by 2-5 times as much as they increase reading scores (Abdulkadiroglu et al. 2011; Angrist et al. 2012; Dobbie and Fryer 2011).[2] Fryer (2014) shows a similar result using charter school best practices in traditional public schools. The average effect of reading interventions targeting students with an average age greater than 12 is 0.04 standard deviations (hereafter σ).[3]

In an effort to increase reading achievement among adolescents – particularly underachieving minority students – we conducted a randomized experiment with approximately 1,700 students across 60 traditional New York City public schools. Treatment schools provided 2.5 hours of daily after-school programming, which included 45 to 60 minutes of 4-on-1 reading tutoring for a subset of students. They also participated in the NYC Middle School Quality Initiative (MSQI) – a school-wide literacy program that provided professional development to teachers and administrators, literacy coaching, access to literacy curricula and software, and a daily period of differentiated literacy support. Control schools participated in MSQI, but did not have tutoring or other after-school programming. A secondary set of control schools is used, when needed, as a "pure" control. These schools did not administer the assessment that was used to determine which subset of students would participate in tutoring in the treatment schools, and thus, cannot be used

[1] There are very successful reading interventions for younger children. This includes early childhood interventions such as Head Start, Breakthrough to Literacy, and Ready, Set, Leap, which increase reading achievement by 0.19σ, 0.55σ, and 0.51σ, respectively (Puma et al. 2010, Layzer et al. 2007). School-based interventions in elementary schools, such as Success for all and Reading Recovery, are similarly effective – increasing reading scores on average by 0.3σ and 0.6σ, respectively (Borman et al. 2007; May et al. 2013).

[2] A potentially illuminating exception is the SEED urban boarding school, which increases reading achievement by 0.21σ and math achievement by 0.23σ, per year (Curto and Fryer 2014).

[3] Calculated using data from Fryer (2017a).

for our main evaluation of the impact of tutoring. They can, however, be used to evaluate the independent impact of MSQI.

The tutoring intervention was implemented as part of the after-school programming for a targeted subset of students in eligible grades each year. Some schools conducted a portion of their tutoring within the school day. With guidance from literacy specialists, we targeted students who were on the cusp of falling significantly behind – i.e. those who demonstrated at least basic fluency but had below-grade-level comprehension skills. A team of teachers and middle school literacy experts developed a detailed tutoring curriculum (approximately 50 pages of curricular material per book for 150 books) centered on high-interest chapter books that were appropriate for the range of reading levels that were eligible for tutoring. These curricula are available from the authors upon request for both practitioners and researchers.

The results of our experiment are interesting, and in some cases, quite surprising. Unless otherwise noted, we report Intent-to-Treat (ITT) effects with standard errors clustered at the school level. On project administration data that serve as a "proof of treatment," there are large treatment effects. Students in treatment schools had a 38 percentage point (or 1.3σ) increase in attendance at tutoring. On average, treatment students attended tutoring on 38 percent of school days; control students attended our tutoring program on 0.02 percent of days. This translates into a treatment effect equal to 67 days of tutoring per year where students read, on average, 1.8 books, or 403 pages per year. All of these effects are highly significant.

These changes translated into significant positive gains in student attendance at school and positive, but insignificant gains on English Language Arts (ELA) state test scores at the mean. The average impact of treatment on attendance is 1.2 (0.3) percentage points (control mean = 91.9 percent). The average impact on the New York state ELA assessment is 0.05σ (0.04) per year. The impact of treatment on the New York state math assessment is -0.002σ (0.06) per year.

We explore heterogeneity of treatment effects across various student, neighborhood, tutor, and school characteristics. Surprisingly, there are no significant differences by student economic disadvantage, English proficiency, language spoken at home, or neighborhood characteristics. Non-special education students may have benefitted more than students with special learning needs. Differences by student baseline reading level suggest that tutoring may have had a larger impact for lower-ability students, but our ability to detect differences is limited due to sample size. Tutor characteristics do not seem to explain differences in the effectiveness of tutoring in increasing ELA

achievement. Student attendance at school increased more in schools with a higher percentage of black tutors, a lower percentage of Hispanic tutors, and in schools with below-median average tutor interview scores. Consistent with the individual student subsamples, ELA scores and attendance at school increased more in schools with an above-median percent black student population and a below-median percent special education student population.

Perhaps the most interesting partition of the data is by students' race. The treatment effect on black students is 0.09σ (0.03) in ELA and 0.10σ (0.06) in math. The treatment effect on Hispanic students is 0.01σ (0.04) in ELA and -0.08σ (0.06) in math. The differences – 0.08σ in ELA and 0.18σ in math – are statistically significant.[4] The impact on reading for black students is approximately twice the impact of attending Promise Academy in the Harlem Children's Zone, having a Teach For America teacher, or the average intervention designed to boost reading achievement for students over age twelve (Dobbie and Fryer 2011, Tuttle et al. 2013, Fryer 2017a). The impact on attendance is 2 percentage points for black students and 0.8 percentage points for Hispanic students.

We test the robustness of our main results in four ways: attrition out of sample, further investigation of school-level heterogeneity via school-level regressions, finite sample inference, and correcting for multiple hypothesis testing. The mean impact of treatment on attendance passes all tests. The effects on attendance for black students survive all four tests and the positive effects on ELA scores for black students maintain significance in three of the four tests. School level regressions yield qualitatively similar estimates of the impact on ELA scores for black students, but are measured with significant error. We cannot reject the null hypothesis that the black coefficient estimated by individual or school-level regressions is the same.

The paper concludes with a speculative discussion of why high-dosage tutoring is more effective for black students than Hispanic students. Accounting for language spoken at home, birth country, whether a student is an English Language Learner, attrition, neighborhood characteristics, and students' pre-treatment reading achievement quintiles reduces the black-Hispanic difference by 32%. Additionally accounting for tutor characteristics reduces the coefficient by 109%, though large standard errors still allow for sizable differences in the effect of treatment in the 95% confidence interval.

[4] The effect in ELA for white students is -0.07σ (0.12) and 0.19σ (0.11) for Asian students; however there are fewer than a hundred white and Asian students in the sample making it unwise to generalize from these estimates.

The basic economics of this experiment is trivial: tutoring provides more signals and a way to individualize instruction. In many regards, it's the opposite of the experiment in Fryer (2018) where teachers were departmentalized to exploit comparative advantage in subject-area expertise. The model that emerges in that paper is one of dial-setting – teachers who have the same students all day are able to better understand the "state" their students are in relative to teachers who see many more students for an hour a day. Tutoring allows for intense individualization.

The field experiment most closely related to our demonstration project is an in-school literacy support program that provides instruction to help struggling ninth-grade readers develop the skills and strategies used by proficient readers, improve their reading comprehension, and increase their self-motivation to read (Somers et al. 2010).[5] This intervention takes the place of elective courses, is additive to regular ELA courses, and is taught by a trained ELA or social studies teacher. The classes were designed for groups of 10 to 15 students that meet for 225 minutes per week throughout the school year.[6] The experiment, which has several elements in common with MSQI, was conducted in 34 schools across 10 school districts – a total of 5,500 students. This literacy intervention increases scores on state reading tests by 0.11σ (0.04) and significantly increases students' reading comprehension, grade point averages, and number of credits earned in core classes. An important difference is that our approach does not crowd out other subjects. Additionally, our intervention is designed to improve reading proficiency in middle school with the goal that students can become proficient readers before they enter high school.

A second intervention – at MATCH charter high school in Boston – that is closely related to our experiment combines small-group (2-4 students) tutorials in math and English with an extended day. Kraft (2015) shows that this high-dosage tutoring in both math and ELA increases ELA scores by 0.15σ to 0.25σ per year, but has no marginal effect on math scores (on top of the large gains in math that MATCH students were making before the advent of the extended day program).

Finally, small-group tutoring is also a key feature of several attempts to "turn around" failing public schools, and it has been demonstrated that intensive small-group instruction in mathematics can lead to especially large gains in student achievement in math (Fryer 2014). There is also evidence

[5] There are several other interventions targeting elementary school students that have been shown to increase reading achievement (e.g. Borman et al. (2009) – see Fryer (2017a) for a complete review).

[6] These results are for an evaluation of the *Reading Apprenticeship Academic Literacy* and *Xtreme Reading* literacy programs. For more information on these specific courses, see www.wested.org or sim.kucrl.org/products/details/xtreme-reading.

that short-term, intensive small group instruction during school breaks can be particularly effective as part of "turnaround" efforts (Schueler et al. 2017).

The paper is structured as follows: Section II provides information on participating schools and program details. Section III describes our data and research design. Section IV presents estimates of the impact of the experiment on student achievement and school attendance. Section V provides a set of standard robustness checks of our results. Section VI discusses possible explanations for some surprising facts generated by our experiment and the final section concludes. There are three online appendices. Appendix A is an implementation guide. Appendix B describes how the variables were constructed in our analysis. Appendix C describes a cost-benefit analysis for the tutoring experiment.

II. Background and Program Details

The New York City Department of Education (NYC DOE) is the largest school district in the United States and one of the largest school districts in the world, serving 1.1 million students in 1,429 schools. Over seventy percent of NYC DOE students are black or Hispanic, fifteen percent are English Language Learners, and over seventy percent are eligible for free or reduced-price lunch.

Table 1 provides a bird's eye view of the experiment. Appendix A, an implementation guide, provides further details. To begin the field experiment, we followed standard protocol. First, we garnered support from the district chancellor and other key district personnel. The district then solicited interest in the project from approximately 150 middle school principals across the district and then provided a list of 129 schools that were eligible for random assignment from this list of interested schools. To control costs and maximize the probability of success for the program, experimental schools were selected for the randomization based upon school size and subject to a minimum school environment grade on the NYC DOE school survey. The sixty smallest interested schools with school environment grades of D or higher formed the experimental group.

The list of sixty schools were randomly assigned to one of three groups – the treatment group consisted of twenty schools that were assigned to be a part of NYC DOE's Middle School Quality Initiative (MSQI) and added 2.5 hours of after-school programming which included 4-on-1 tutoring and the main control group consisted of twenty schools that were assigned to be a part of NYC DOE's MSQI. A second control group serves the purpose of a "pure" control – a way of understanding the impact of MSQI on outcomes.

After treatment and control schools were chosen, treatment school principals attended a meeting in June of 2013. During this meeting, the general outline of the project was described and principals were given a forum to ask any questions they had about participation in the project. The treatment was implemented over the course of three school years, serving 6th graders in 2013-14; 6th and 7th graders in 2014-15; and 6th, 7th, and 8th graders in 2015-16. Due to the obvious selection concerns, our analysis is restricted to students who were in 6th grade in 2013-14, and follows that cohort for three years.

The treatment group participated in a comprehensive set of interventions beginning in the 2013-14 school year – a school-wide literacy initiative, after-school programming, and high-dosage reading tutoring for a targeted group of students. The school-wide literacy program was designed to cultivate a "literacy culture" by providing ongoing literacy coaching, access to literacy curricula and software, and programming support in implementing a daily period of differentiated literacy support (a strategic reading period or "SRP").

The after-school programming was implemented by ExpandED Schools. Each treatment school partnered with a community-based organization (CBO) to extend the length of the school day and expand the types of learning experiences available to students. These programs added approximately 2.5 hours to the school day, offering students a mixture of academic and non-academic activities, as well as an additional meal. The types of activities offered varied by site, depending upon a school's particular CBO partner, the skills of the CBO employees, and the involvement of teachers and school administrators in the after-school programming. Examples of typical activities include dance, sports, science labs, homework help, robotics, photography, and debate club. Students could attend the after-school program if they were enrolled in the grade being offered tutoring, irrespective of whether or not they qualified for tutoring. If a student was not eligible for tutoring, they could participate in the programming described above. If they were assigned to tutoring, students were not allowed to opt out of tutoring to participate in other after-school activities (though attending after-school was still voluntary). If tutoring occurred during the school day, it was treated like any other (mandatory) class.[7,8] For the rest of the time spent in after-

[7] One might expect differential effects of the tutoring program depending on whether it occurred during school or after school. One might particularly anticipate higher attendance rates at tutoring that occurred during the school day. On the other hand, tutoring that occurred during the school day was the most common to be cannibalized by schools if they needed to utilize class time (for an assembly, etc.) and students may have been more likely to skip tutoring classes than other classes, since it was not taught by a normal teacher and students did not receive a grade or course credit. Unfortunately, we do not have usable data on which students were assigned to in- or after-school tutoring.

school but not spent in tutoring, students could participate in the activities implemented by ExpandED Schools described above.

High-dosage tutoring was implemented as a part of the after-school programming. Tutoring served a subset of students in eligible grades each year. Students were assigned to be tutored based upon their diagnostic score on the Degrees of Reading Power (DRP) assessment (taken in September of 2013), a multiple choice reading comprehension assessment, and were grouped homogeneously by their DRP scores. The eligibility range was established to identify 6th grade students who demonstrated at least basic middle-elementary level fluency, but who demonstrated below grade-level comprehension skills.[9] Assignment was done so that the lowest performing eligible students received priority when there were more eligible students than available tutoring seats at a school.[10] In the first year, the percent of sixth graders assigned to tutoring in a given school ranged from 17 percent to 64 percent, and in the average school, 46 percent of students were assigned to tutoring.

Each tutoring session was scheduled for between 45 and 60 minutes, depending on each school's schedule. Each tutoring group consisted of four students, grouped according to DRP scores in order to allow for effectively targeted book selection and instruction. The high-dosage tutoring was modeled after the guided reading framework described in Fountas and Pinnell (2001). The curriculum centered on high-interest chapter books – fiction and non-fiction – appropriate for the range of reading levels eligible for tutoring participation. A team of teachers and middle school literacy experts developed a detailed curriculum for tutors to use in the tutorial and worked with NYC DOE to curate a library of high-interest chapter books for each grade level of participating students.

---

[8] We do not have information on what students were doing if they were not attending the after-school program. Anecdotally, we were told that students who didn't stay for after-school programming were going home, or generally, had no close substitutes.

[9] In 2014-15, 7th grade students who were eligible for tutoring in year one, who scored below 64 on the year two beginning of year DRP assessment, and who remained enrolled in a treatment school remained eligible to participate in high-dosage tutoring. The same process was in place for the 2015-16 school year, with the upper cut-off increasing to 67 on the DRP for 8th grade students. Students whose initial year one DRP scores qualified them for high-dosage tutoring could "graduate out" of the program for one school year by scoring above 64 at the start of year two, or above a 67 at the start of year three. This was very uncommon – only 2.7% percent of students did so in year two and 4.7% did so in year three.

[10] Eligible students who were not assigned to tutoring were assigned to a waitlist, ordered by diagnostic DRP score (from low to high), to fill open seats as the year progressed. If an assigned student stopped participating in the tutoring program, the empty seat was filled by an eligible student from the waitlist. Students who may have gotten into tutoring off of the waitlist are not included in the analysis.

All tutors were required to meet certain baseline requirements before they were hired. Tutors were required to hold at least a bachelor's degree, pass a high school reading assessment, and pass a background check. Prior to being placed at a given school, principals or their chosen designee interviewed each tutor candidate to determine their fit on a particular campus. In year one, at full capacity there were 110 tutors working in 20 schools; in years two and three, there were 145 tutors at 19 schools.[11] More details on tutor staffing by school can be found in Appendix Table 1.

Before each school year, tutors participated in an intensive week-long training to implement the guided reading model designed for the tutoring program. The summer training focused on the guided reading instructional model, lesson planning using the curriculum, managing student behavior, building relationships with students, and establishing systems and procedures to run effective tutorials.

In year one, tutors and the tutoring program were supervised by seven Regional Tutoring Coordinators (RTCs). These seven RTCs managed tutor staffing, coordinated tutor logistics, and provided instructional coaching for tutors at 2 or 3 schools each. In years two and three, the supervisory structure changed slightly. Each treatment school was assigned an ExpandED Schools Program Manager, who coordinated with the CBOs to oversee the after-school programming at the school and maintained supervisory responsibilities over the tutoring program, in order to better integrate the components of treatment. RTCs worked closely with our research team to ensure fidelity of implementation to the program's research design and instructional model. RTCs served as instructional leaders for tutors, providing regular informal feedback, instructional coaching, and material resources. Tutors also attended professional development seminars during the school year that focused on the guided reading instructional model, lesson planning using the curriculum, managing student behavior, building relationships with students, and establishing systems and procedures to run effective tutorials. Half-day professional development seminars were offered 3-5 times annually, typically during school holidays.

The cost of the experiment was $1.76 million per year, or approximately $2,500 per student per year.

III. Data, Research Design, and Econometrics

Data

[11] One school dropped out of the tutoring program; students attending that school are included in all years of analysis to maintain the integrity of the research design.

We use student-level administrative data on 1.1 million students enrolled in NYC DOE schools, each year, from the 2010-2011 to 2015-2016 school year. The student-level data include information on student race, gender, free- and reduced-price lunch eligibility, and attendance for grades K-12, and state math and ELA test scores for students in grades 3-8. To supplement NYC DOE data, we also collected administrative data throughout the project, described below.

The main outcome variable is an achievement test unique to New York state. The state ELA and math tests, developed by McGraw-Hill, are high-stake exams administered to students in the third through the eighth grade. Students in third, fifth, and seventh grades must score at level 2 or above (out of 4) on both math and ELA tests to advance to the next grade without attending summer school. Material covered in the math test is divided among five strands: (1) number sense and operations; (2) algebra; (3) geometry, (4) measurement, and (5) statistics and probability. The ELA test is designed to assess students on three learning standards: (1) information and understanding, (2) literary response and expression, and (3) critical analysis and evaluation. Both tests include multiple-choice, short-response, and extended response questions. The ELA test also includes a brief editing task and is divided into reading, listening, and writing sections.

All public-school students are required to take the math and ELA tests unless they are medically excused or have a severe disability. Students with moderate disabilities or limited English proficiency must take both tests, but they may be granted special accommodations (additional time, translation services, and so on) at the discretion of school or state administrators. In the analysis, test scores are normalized to have a mean of zero and a standard deviation of one for each grade and year across the entire New York City sample.

We use a parsimonious set of controls to aid in precision and to correct for any potential imbalance between treatment and control groups. The most important controls are achievement test scores from three years prior to treatment and their squares, which we include in all regressions. Pre-treatment years' test scores are available for most students who were in the district in the previous year. We also include indicator variables that takes on the value of one if a student is missing a test score from a given pre-treatment year and zero otherwise.

Other individual-level controls include a mutually exclusive and collectively exhaustive set of race dummies, as well as indicators for gender, free lunch eligibility, special education status, and English Language Learner status. All controls are interacted with whether a student is black, Hispanic, or another race. See Appendix B for further details on the construction of all variables.

To supplement NYC DOE's administrative data, we collected a large set of project implementation data. This includes data on tutor characteristics collected from resumes and site visits, student attendance at tutoring, and the books read by each student in tutoring. This data was collected in all three years of treatment.

Research Design

To partition the set of interested schools into treatment and control, we used a matched-triple randomization procedure similar to those recommended by Abadie and Imbens (2011). Recall, 60 schools were included in the randomization, from which we constructed 20 matched triples.

To increase the likelihood that our control and treatment groups were balanced on a variable that was correlated with our outcomes of interest and following the recommendations in Abadie and Imbens (2011), we used schools previous years' average proficiency rates in math and ELA to construct our matched triples.[12] First, the full set of sixty schools was partitioned into "matched triples" using each school's sixth grade enrollment and their average proficiency rate in math and ELA. Specifically, schools were first grouped by size into 9 clusters, by rounding sixth grade enrollment to the nearest 50. Then, schools within each cluster were ordered by the average of their math proficiency rate and ELA proficiency rate. Using this ordering, schools were then grouped into triples within each size cluster such that the three schools with the lowest proficiency rate in each size cluster became one matched triple, etc. When there was a size-cluster with a number of schools that was not divisible by three, similarly-achieving schools were pulled from the next largest size-cluster. From each "matched triple" we randomly selected one school to be treatment, one to be main control, and one to be supplementary control.

Columns (1) and (2) of Table 2 provide descriptive statistics for both participating and non-participating schools.[13] Column (3) provides p-values for each individual variable. This is estimated by regressing school- and student-level characteristics on an indicator for being in the experimental sample. Panel A includes variables measured at the school-level – the unit of analysis in our random

[12] Ideally, one would want to use scale scores on the state test for all students. NYC DOE typically releases these data in December. We had to choose between more recent but more aggregate data or less recent microdata. We chose more recent aggregate data.

[13] Two schools in the control group failed to administer the DRP assessment used to assign students to tutoring groups. Those matched pairs are therefore eliminated from the sample for analysis, since we cannot determine which students belong in the comparison group of students who would have been tutored in control schools. As a robustness check, we predict DRP scores for students in these matched pairs and include them in the analysis – see Appendix Table 6. The qualitative results are unchanged.

assignment. Overall, the participating versus non-participating sample is unbalanced at both the school and individual student-level (p-value the joint F-tests both 0.000). Schools in the experimental sample have a lower percentage of white students, a higher percentage of Hispanic students, English Language Learners, and economically disadvantaged students, and lower average test scores in math and ELA. This is consistent with the district's proposed selection of schools to participate in the experiment.

Columns (4)-(7) provide identical information for schools randomly assigned to the treatment, main control, and supplementary control groups. No individual characteristic is statistically different across groups and the schools are balanced overall (p-value on the joint F-test is 0.954). The students in the 6th grade cohort (the group used to randomize), and the group of students who fall in the range of DRP scores that qualify them for tutoring are also balanced overall (p-values from the joint F-tests are 0.147 and 0.405, respectively) and no individual characteristic is statistically different between groups. This makes inference relatively straightforward.

Econometric Specifications

Let Z_i indicate whether student i was enrolled in a school selected for treatment at the beginning of the first treatment year, let X_i denote a vector of control variables consisting of the demographic variables in Table 2, let $f(\cdot)$ represent a polynomial including three years of prior test scores and their squares. All these variables are measured pre-treatment. Ψ_m is a matched-pair fixed effect, and η_t is a year fixed effect.

We can then estimate the Intent-to-Treat (ITT) effect τ_{ITT} using the twenty treatment and twenty control schools in our experimental sample via the following regression model:

$$(1)\ Y_{i,s,m,t} = a + \tau_{ITT} \cdot Z_i + f(Y_{i,t-1}, Y_{i,t-2}, Y_{i,t-3}) + \beta X_i + \Psi_m + \eta_t + \varepsilon_{i,s,m,t}$$

Equation (1) identifies the impact of being offered a chance to attend a treatment school, τ_{ITT}, where students in the matched-pair schools correspond to the counterfactual state that would have occurred for the students in treatment schools had their school not been randomly selected.

We focus on a fixed population of students. A student is considered treated if she was in a treatment school before October 31st in the first year of treatment, and if her score on the DRP lies within the range used to determine the tutoring groups in her school. A student is similarly assigned to the control group if her score on the DRP lies within the range used to determine the tutoring

groups in the treatment school in her school's matched triple. All student mobility (in or out of schools or in and out of the tutoring program) after treatment assignment is ignored. Note: more than 95% of students in our sample remain in the same school between September and June in each year of the treatment. Further, at the end of the second year of treatment, 87% of students in our sample are in the same school that they were in at the start of the experiment, and at the end of the third year of treatment, 80% of students are in the same school that they were in at the start of the experiment (retention is similar between treatment and control). We follow the initial experimental cohort of 6th graders for three years. Standard errors in all specifications are clustered at the school level.

Under several assumptions (e.g. that treatment assignment is random, control schools are not allowed to participate in the program and treatment assignment only affects outcomes through program participation), we can also estimate the causal impact of *attending* a treatment school or *participating* in tutoring. This parameter is commonly known as the Local Average Treatment Effect (LATE).

We estimate a LATE parameter through a two-stage least squares regression of student achievement on fraction of years a student is present in tutoring, using random assignment as an instrumental variable for the first stage regression. The LATE parameter measures the average effect of attending a treatment school on students who are treated as a result of their school being randomly selected.

IV. Analysis

Direct Outcomes

We begin our analysis by estimating treatment effects on a variety of outcomes that are directly related to the quality of implementation of tutoring by schools. Table 3 contains these estimates.

Perhaps the most straightforward and obvious way to provide "proof of treatment" is to estimate whether students attended the tutoring sessions outlined in Section II. Over three years, treatment students' attendance rate in this tutoring program was 38 percentage points higher than

control students' (control mean = 0.02 percent).[14] Treatment students attended an average of 67 days of tutoring *per year* – a total of 150-200 hours over the three-year experiment in which they read, on average, a total of 5.4 books or 1209 total pages (1.8 books or 403 pages per year) and discussed, in detail, those books with their tutor and peers. All of these estimates are highly significant.

To get a sense of how large the treatment effect on this set of direct outcomes is, consider that the average treated student read approximately 67,000 words per year in tutoring. The average sixth grader in America reads 425,000 words per year in school (Renaissance Annual Report 2012). Thus, tutoring increased the number of words read in a year for an average sixth grader by approximately 16%.

Effects on direct outcomes were largest in year one and smaller in years two and three. In year one, students in tutoring read, on average, 600 pages, whereas in year two and three students read, on average, 300 pages. Similarly, students attended approximately 80 days of tutoring in year one and 60 days per year in years two and three. Note well: of the students assigned to tutoring in the first year of the program, 45% attended at least one day of tutoring in all three years of the program, 25% attended at least one day of tutoring in two years of the program, 23% attended at least one day in only the first year of the program, and 6% attended zero days in zero years of the program.[15]

State Test Scores

Panel A of Table 4 presents a series of ITT estimates of the effect of the tutoring treatment on ELA and math state test scores. Columns (1) through (3) present estimates for years one through three separately, and Column (4) displays results pooled over all three experimental years. All specifications control for matched pair fixed effects; three years of baseline test scores and their squares; indicators for whether a student is economically disadvantaged, an English Language Learner, or received Special Education services; and missing indicators in all variables. All control variables are interacted with whether a student is black, Hispanic, or another race. Results are presented in standard deviation units. Standard errors, clustered at the school level, are displayed below each estimate.

[14] Unfortunately, we do not know if students in the control group were attending private tutoring programs. If they were doing so, the programs were not structured within the public school system.

[15] Given our research design and the fact that we ignore student mobility between schools, or in and out of the tutoring program, average attendance at tutoring will weakly decrease mechanically each year.

The impact of being offered the opportunity to participate in a tutoring intervention on New York state test scores is 0.05σ (0.04) per year in English Language Arts (ELA) and -0.002σ (0.06) per year in math. The ELA score can be further decomposed into its reading and writing subscores. The impact of being offered a chance to participate in a tutoring intervention on pooled reading subscores is 0.02σ (0.03) per year and 0.08σ (0.06) per year for writing. Consistent with the patterns in direct outcomes, the effects on ELA scores are largest in the first year.

We also use assignment to treatment to instrument for students' attendance rate at tutoring. These results suggest that the effect of *attending* the high-dosage tutoring program for one year is to increase ELA scores by 0.16σ (0.08) in the first year or 0.12σ (0.09) per year when pooled over all three years.

School Attendance

Panel B of Table 4 presents ITT estimates of the effect of treatment on student attendance at school. Tutoring significantly increases student attendance by 1.2 percentage points per year, relative to a control mean of 92 percent, and the effect is highly significant in each year and pooled across years. We also instrument for students' attendance rate at tutoring with assignment to treatment. *Attending* the tutoring program for one year increases attendance at school by 3 (0.7) percentage points per year.[16]

Conceptually, tutoring (or after-school programming) could increase students' attendance at school via at least two mechanisms – either students wanted to attend after-school programming/tutoring and knew that they could not do so if they were absent from school, or students felt more confident at school because of the tutoring and therefore attended more days of school.

We can partially separate the effect of after-school programming from the effect of tutoring coupled with after-school programming by comparing students who were eligible for after-school programming but **not** eligible for tutoring (students in the same grade as students who were offered tutoring but whose DRP scores were too low or too high to qualify for tutoring) with similar

[16] One might wonder how much of the effect on ELA and math scores we should attribute to the direct impacts of tutoring versus indirect impacts mediated by higher overall rates of attendance at school. Goodman (2014) presents evidence from snow days that suggests that attendance at school may increase math scores but not ELA scores. This provides some speculative evidence that the positive effects on ELA scores are direct effects of tutoring rather than secondary effects of increased school attendance.

students in the control group who only were exposed to MSQI. This provides the treatment effect of the tutoring component of the after-school program – relative to the other elements such as sports or photography – on attendance rates. A key assumption in this approach is that the effect of after-school programming is the same for students who were eligible for tutoring and those who were not eligible for tutoring (tutoring assignment is formulaic but not random).

The results of this approach are displayed in Appendix Table 7. Interestingly, the effects on school attendance are driven by participation in the tutoring program, not the after-school program without tutoring. In other words, these results suggest that students are more likely to come to school because of our after-school tutoring program than the other services offered after school such as science labs or dance.

Heterogeneous Treatment Effects

Table 5 explores the heterogeneity of our treatment effects on test scores and school attendance across a wide variety of subsamples of the data. Splitting the sample by whether or not students are economically disadvantaged or are English Language Learners, whether they speak English or Spanish at home, or their neighborhood characteristics (e.g. median income or number of police stops) yield either insignificant or inconsistent results.

Students without special learning needs gained 0.06σ (0.04) in ELA while students with special learning needs lost 0.05σ (0.04) in ELA (p-value on the difference is 0.06), but there is no differential effect on attendance. Students in all quartiles of the pre-treatment ELA state test or DRP assessment score distribution had statistically similar treatment effects on ELA, but magnitude of coefficients decreases with student pre-treatment test scores. The effects on school attendance were larger for students with lower pre-treatment ELA state test or DRP assessment scores.

Partitioning the sample by tutor characteristics, such as schools' average tutor quality score, percent of tutors with a degree in English or education, percent of tutors with teaching or tutoring experience, percent of tutors with a graduate degree, or the percent of tutors who are black, Hispanic, or white yields no significant results for ELA. The effects on attendance were significantly larger in schools with below-median average tutor interview scores and in schools with more black tutors or fewer Hispanic tutors.

The most robust partition of the data is by student race. Black students gain 0.09σ (0.03) in ELA and 0.10σ (0.06) in math. Hispanics gain 0.01σ (0.04) in ELA and lose 0.08σ (0.06) in math.[17] The p-value on the difference between black and Hispanic students is 0.05 in ELA and 0.01 in math. Consistent with these results, treatment also increases students' attendance at school significantly more for black students (2.0 (0.4) percentage points) than it did for Hispanic students (0.8 (0.4) percentage points).[18]

To put these magnitudes in context, Dobbie and Fryer (2011) report that the impact of attending the Promise Academy Middle School in the Harlem Children's Zone for one year on ELA achievement is 0.05σ (0.03). Tuttle et al. (2013) document a similar impact for KIPP schools. The impact of having a Teach For America teacher is 0.03σ (0.04) on ELA scores (Glazerman et al. 2006). Implementing charter school best practices in traditional public schools had no effect on ELA scores in secondary schools, and training principals in better management practices increases ELA scores by 0.05σ (0.02) (Fryer 2014; Fryer 2017b). Thus, high-dosage tutoring has a similar impact on reading achievement to other well-known interventions.

Consistent with the above results, schools that have a below-median percent of students in special education and schools with an above-median percent of black students have larger effects on ELA scores. Schools with a below-median percent ELL students and above-median percent of black students have larger effects on attendance.

The Impact of MSQI on State Test Scores and Attendance

We conclude our main statistical analysis by estimating the impact of our secondary treatment arm – the MSQI literacy curriculum. Recall, both treatment and control schools received the MSQI literacy curriculum which provided professional development to teachers and administrators, literacy coaching, access to literacy curricula and software, and support to implement a daily period of differentiated literacy support.

Our set of supplementary control schools – those that did not implement MSQI – provide a way to test the independent effect of MSQI. While these pure control schools cannot be used to

[17] There are fewer than a hundred students in the sample of tutored students who are white, Asian, or other race. Results are therefore only presented for black and Hispanic students.

[18] Appendix Table 8 shows that there is no statistically significant or economically meaningful difference in the effect of treatment on direct outcomes (attendance at tutoring, books read in tutoring, or pages read in tutoring) for black students versus Hispanic students.

identify the impact of tutoring coupled with MSQI (as they did not administer the DRP and therefore one cannot identify the counterfactual group of control students in the correct diagnostic score range), we can compare outcomes for students in the main control schools (that implemented MSQI) to students in the supplementary control schools to estimate the effect of MSQI for the entire sixth grade cohort, employing the same specifications used to evaluate the tutoring program.

The effect of MSQI is -0.02σ (0.03) per year on state ELA scores and -1.2 percentage points (0.002) per year on attendance. All other coefficients are also statistically zero [not shown in tabular form].

V. Robustness Checks

In this section we briefly explore the robustness of our key results – that high-dosage tutoring increases student attendance at the mean and is highly effective for black students on both the attendance and student achievement margins – under potential threats to the interpretation of the data.

Attrition and Bounding

Our estimates thus far have been based on the sample of students who take the NY state test at the end of each year of treatment. If treatment affects selection into this sample, our results may be biased.

To test this potential concern, Appendix Table 2 provides estimates of the effect of treatment on attrition for the overall sample and by student race. There is no detectable effect of treatment on attrition in the overall sample, nor is there a significant difference in the attrition rate for black and Hispanic students.

Appendix Table 3 includes bounds on the main estimates that account for differential attrition. As described in Lee (2009), we calculate lower bounds by dropping the highest-achieving treatment students, or lowest-achieving control students, until attrition is equal between treatment and control. This process occurs independently for each outcome. We re-run the main specification, including all of the same controls, on this new sample to estimate the worst-case scenario treatment effect – i.e., the treatment effect if all of the excess treatment (excess control) respondents were the "best" ("worst") respondents on each measure.

The ELA effect for black students is still highly significant in the first year and averaged over all years – the pooled coefficient is 0.07σ (0.03). The effect on attendance for black students also

remains highly significant in all years. The positive effect on attendance at the mean remains highly significant in all years.

School Level Heterogeneity

In general, controlling for matched pair fixed effects should yield consistent standard errors (Abadie and Imbens 2011), but this may not correct for school-level heterogeneity in finite samples. This heterogeneity is uncorrelated with treatment due to random assignment, but could affect inference (Moulton 1986, 1990). Therefore, all results presented above cluster standard errors at the school level.

A second, more conservative way to address this issue is to estimate school-level regressions to evaluate the impact of treatment. Appendix Table 4 displays these estimates for our experimental sample. The results are qualitatively the same, but estimated with such imprecision as to render the ELA effects for black students insignificant. Notice, however, one cannot reject the hypothesis that the school-level regression results and the individual regression results are statistically the same. The effects on attendance continue to be significant both at the mean and for black students when estimated at the school level.

Permutation Tests

A third robustness check is to understand how our small number of clusters (36) impact inference. Cameron, Gelbach, and Miller (2008) advise concern in designs that have fewer than 30 clusters. We pass this standard, but 36 clusters is still cause for concern that standard asymptotics do not apply.

Appendix Figure 1 provides exact p-values calculated via permutation tests for the key results (Fisher 1935, Rosenbaum 1988). To conduct the permutation test, we re-randomize the sample 10,000 times between matched pairs at the school level, identical to the original random assignment design, and calculate a simulated treatment effect. The exact p-value is the proportion of simulated treatment effects that are larger than the actual observed treatment effect (in absolute value).

The effect on ELA scores for black students remains marginally significant and the effects on attendance, both for black students and at the mean, remain highly significant.

Multiple Hypothesis Testing

We have run many regressions with various outcomes to measure treatment effects. One might worry that we are simply detecting false positives due to multiple hypothesis testing. Using a standard Bonferroni correction (the most conservative correction), effects on both ELA and attendance remain significant for black students. The effect on attendance at the mean also maintains significance. Appendix Tables 5A and 5B contain these estimates.

VI. Discussion and Speculation

Our field experiment generated an unexpected set of facts. At the mean, tutoring had a positive and statistically significant effect on school attendance for all students and for black students and a positive, but an insignificant impact on ELA state test scores. Interestingly, the effects at the mean are largely driven by black students, who made considerable gains, while Hispanic students gleaned no measurable benefit from tutoring. The effect on attendance for black students survive all of our robustness checks; the effect on black students' ELA scores pass three of four tests.

In this section, we take the point estimates literally and provide a (necessarily) more speculative discussion of why high-dosage tutoring is more effective for black students relative to Hispanic students. To be clear, the empirical tests to come would not have been a part of any pre-analysis plan as we did not expect these results a priori. And, importantly, we are limited in what we can test due to data limitations.

To understand racial differences in treatment effectiveness, we estimate empirical models of the following form:

$$(2)\ Y_{i,s,m,t} = a + \chi_{ITT} \cdot Z_i + \phi_{ITT} \cdot Z_i \cdot HISP + f(Y_{i,t-1}, Y_{i,t-2}, Y_{i,t-3}) + \beta X_i + \Psi_m + \eta_t + \varepsilon_{i,s,m,t}$$

where Z_i is an indicator for treatment and *HISP* is an indicator for being Hispanic. χ_{ITT} represents the treatment effect for black students and ϕ_{ITT} represents the difference in treatment effects for Hispanic students relative to black students. To test a particular hypothesis – say, that language spoken at home explains racial differences – we add controls that proxy for that hypothesis and investigate the change in ϕ_{ITT}.

Table 6 estimates equation (2) and presents estimates of ϕ_{ITT}, additively accounting for several potential hypotheses to explain observed racial differences in treatment effects. The first row contains the estimated racial difference using only the controls in the main analysis. Each following row adds control variables that proxy for each of the following potential explanations for the observed racial differences: 1) language and native country; 2) neighborhood characteristics; 3) diagnostic test scores; 4) previous reading achievement quartiles; 5) tutor characteristics; and 6) student attendance at school. All controls are interacted with student race.

In the main analysis, the coefficient on the difference for Hispanic students relative to black students is -0.08σ (0.04). Adding controls for whether a student speaks English, Spanish or another language at home and whether or not a student was born in the US, census-tract level measures of neighborhood quality, students' diagnostic test scores, or students' pre-treatment reading achievement quartiles reduces the difference coefficient by 32%.

A final set of hypotheses for the racial differences in the effects of tutoring are concerned with potential differences in tutoring quality, student attendance, or engagement with the tutors or the books read. Controlling for the average tutor characteristics of a students' school – the percent of tutors in a school with an English or education degree, the average tutor quality score from the screening process, the percent of tutors with teaching or tutoring experience, the percent of tutors with a graduate degree, and the percent of tutors who are black, Hispanic, and white – reduces the racial difference in the effect of tutoring to 0.007σ (0.08), a 109% reduction. Further controlling for students' attendance at school leaves the coefficient virtually unchanged.[19] The same set of controls explains 58% of the racial differences in the effect of treatment on attendance at school. Taken together, these data suggest that the main drivers of the difference in treatment effects for Hispanic and black students is differences in the average characteristics of tutors at their school.[20]

---

[19] Adding controls for students' participation in tutoring, the number of books read that were tagged as particularly "black interest" or "Hispanic interest," or individual tutor characteristics *increases* the difference in the effect of treatment for black and Hispanic students. Whether books were of particular "black interest" or "Hispanic interest" was determined by the race of the protagonist(s), the race of the author, the image on the cover, or the historical relevance of the plot. Including these controls in addition to those described above yields a final difference in treatment coefficients of 0.01σ (0.09). One potential explanation for why school-level tutor characteristics explain the observed racial differences while individual tutor characteristics do not is that at the school level these measures might proxy for school implementation or commitment to the tutoring program, since principals were part of the tutor selection process.

[20] We are unable to isolate whether racial differences in the effects of tutoring are driven by differences in the accumulation of specific reading skills without more detailed subscores on the state ELA exam (e.g. word reading accuracy, fluency, background knowledge, phonetics, vocabulary, etc.). Unfortunately, the DRP measures only one of these skills – students' comprehension of text passages so we can only test one mechanism. Overall, the effects

VII. Conclusion

Reading achievement is lagging in the United States – particularly for minority students. There are well-known methods of increasing math achievement – e.g., charter school best practices, high-dosage tutoring, among others (Fryer 2017a) – but our understanding of how to increase reading achievement after elementary school is limited.

In an effort to increase reading achievement among middle school students, we conducted a randomized field experiment with roughly 1700 students across 60 NYC public schools. On several direct outcomes – attendance at tutoring, or the number of books or pages read at tutoring – there were large treatment effects. These changes resulted in a positive effect on students' school attendance and a positive, but insignificant, effect on ELA test scores at the mean. The effect on ELA scores is driven largely by black students who experienced a relatively large treatment effect while Hispanic students garnered no marginal benefit from tutoring. Indeed, the expected IRR for black students is 18%; for all students it is 8%. Both pass a simple cost-benefit analysis (see Appendix C for details). Moreover, to the extent that the types of individual relationships that form between tutors and students encourage students to stay in school longer or result in higher educational attainment (as suggested by the positive impacts on attendance), benefits may be significantly understated.

As school districts across the country grapple with how to increase student achievement among adolescents, high-dosage tutoring may be a viable policy solution for both math and reading. Taking the effects of tutoring secondary school students in math from Fryer (2014), and the impacts estimated above, implies treatment effects of 0.09σ in math and 0.05σ in ELA, though the effects on ELA are estimated imprecisely at the mean. These effects are strikingly similar to the impacts of attending a charter school in NYC (Hoxby and Murarka 2009). Importantly, tutoring may be a more politically palatable way to increase achievement in states constrained by legislation that caps the number of charter schools.

---

of treatment on students' DRP scores in the fall of 2014 and 2015 are positive in 2014 and negative in 2015, though insignificant in both years. Recall that DRP scores range from 14-100 and that the range of scores that made a student eligible for tutoring were scores from 40-60, with specific cutoffs varying by school. In 2014, the treatment effect for black students is 1.34 points (0.61) and -0.026 points (0.88) for Hispanic students. The p-value on the difference is 0.188. In 2015, the treatment effect for black students is -1.22 points (1.26) and -1.07 points (1.04) for Hispanic students. The p-value on the difference is 0.923.

References

Abadie, Alberto and Guido W. Imbens (2011). "Bias-Correcting Matching Estimators for Average Treatment Effects," *Journal of Business & Economic Statistics* 29(1): 1-11.

Borman, Geoffrey, Robert Slavin, Alan Cheung, Anne Chamberlain, Nancy Madden, and Bette Chambers (2007), "Final Reading Outcomes of the National Randomized Field Trial of Success for All." *American Education Research Journal*, 44(3): 701-731.

Borman, Geoffrey, James G. Benson, and Laura Overman (2009). "A Randomized Field Trial of the Fast ForWord Language Computer-Based Training Program." *Education Evaluation and Policy Analysis* 31(1): 82-106.

Cameron, Colin A., Jonah B. Gelbach, and Douglas L. Miller (2008). "Bootstrap-Based Improvements for Inference with Clustered Errors." *The Review of Economics and Statistics* 90(3): 414-427.

Curto, Vilsa E. and Roland G. Fryer (2014). "The Potential of Urban Boarding Schools for the Poor: Evidence from SEED," *Journal of Labor Economics* 32(1): 65-93.

Dobbie, Will and Roland G. Fryer (2011). "Are High Quality Schools Enough to Increase Achievement Among the Poor? Evidence From the Harlem Children's Zone," *American Economic Journal: Applied Economics* 3(3): 158-187.

Fisher, Ronald A. (1935), The Design of Experiments, Edinburgh: Oliver and Boyd, Ltd, 1951 (6e).

Fountas, Irene C., and Gay Su Pinnell (2001). *Guiding Readers and Writers, Grades 3-6: Teaching Comprehension, Genre, and Content Literacy*. Connecticut: Heinemann.

Fryer, Roland G. (2014), "Injecting Charter School Best Practices Into Traditional Public Schools: Evidence from Field Experiments," *The Quarterly Journal of Economics* 129(3): 1355-1407.

Fryer, Roland G. (2017a). "The Production of Human Capital in Developed Countries: Evidence from 196 Randomized Field Experiments." In: Handbook of Field Experiments.

Fryer, Roland G. (2017b). "Management and Student Achievement: Evidence from a Randomized Field Experiment." *NBER Working Paper 23437*.

Glazerman Steven, Daniel Mayer, and Paul Decker. (2006). "Alternative Routes to Teaching: The Impacts of Teach For America on Student Achievement and Other Outcomes." *Journal of Policy Analysis and Management* 25(1): 75-96.

Hoxby, Caroline M. and Sonali Murarka. (2009). "Charter Schools in New York City: Who Enrolls and How They Affect Their Students' Achievement." *NBER Working Paper 14852*.

Kraft, Matthew A. (2015). "How to Make Additional Time Matter: Integrating Individualized Tutorials into an Extended Day," *Education Finance and Policy,* 10(1): 81-116.

Layzer, Jean, Carolyn Layzer, Barbara Goodson, and Cristofer Price (2007). "Evaluation of Child Care Subsidy Strategies: Findings From Project Upgrade in Miami-Dade County." Cambridge, MA: Abt Associates.

Lee, David S. (2009). "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *Review of Economic Studies*, 76(3): 1071-1102.

May, Henry, Abigail Gray, Jessica Gillespie, Philip Sirinides, Cecile Sam, Heather Goldsworthy, Michael Armijo, and Manrata Tognatta (2013). "Evaluation of the i3 Scale-up of Reading Recovery." Philadelphia, PA: CPRE.

Moulton, Brent R. (1986). Random group effects and the precision of regression estimates. *Journal of Econometrics*, 32(3): 385-397.

Moulton, Brent R. (1990). An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *The Review of Economics and Statistics*, 334-338.

OECD (2016). PISA 2015 Results (Volume 1): Excellence and Equity in Education, PISA, OECD Publishing, Paris. http://dx.doi.org/10.1787/9789264266490-en.

Puma, Michael, Stephen Bell, Ronna Cook, and Camilla Heid (2010). "Head Start Impact Study Final Report." Washington, D.C.: U.S. Department of Health and Human Services, Administration for Children and Families.

Renaissance Learning Annual Report (2012). "What Kids Are Reading: The Book-Reading Habits of Students in American Schools." Wisconsin Rapids, WI. www.renlearn.com.

Romano, Joseph P., Azeem M. Shaikh, and Michael Wolf (2010). "Hypothesis testing in econometrics." *Annual Review of Economics*, 2: 75-104.

Rosenbaum, Paul R. (1988). "Permutation Tests for Match Pairs with Adjustments for Covariates," *Journal of the Royal Statistical Society*. Series C (Applied Statistics) 37(3): 401-411.

Schueler, Beth E., Joshua S. Goodman, and David J. Deming (2017). "Can States Take Over and Turn Around School Districts? Evidence from Lawrence, Massachusetts," *Educational Evaluation and Policy Analysis* 39(2): 311-332.

Tuttle, Christina, Brian Gill, Philip Gleason, Virginia Knechtel, Ira Nichols-Barrer, and Alexandra Resch (2013). "KIPP Middle Schools: Impacts on Achievement and Other Outcomes." Final Report. Mathematica Policy Research, Princeton, NJ.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (2015). *2015 Reading Assessment.*

Table 1: Description of Treatment

Schools	NYC DOE provided a list of 129 interested and eligible schools, of which 20 schools were randomly selected into treatment, 20 into a "main" control, and 20 into a "supplementary" control.
School Years	2013-2014 to 2015-2016
Treatment Students	858 students entering 6th grade in 2013: 47% black, 44% Hispanic, 86% economically disadvantaged
Control Students	913 students entering 6th grade in 2013: 41% black, 45% Hispanic, 81% economically disadvantaged
Student Assignment to Tutoring	Students were eligible for tutoring if their scores on the Degrees of Reading Power assessment fell between 40 and 60. Students with the lowest DRP scores in this range were assigned to tutoring until the tutoring cohort reached the maximum capacity at each school. Students were grouped by DRP score into groups of four to allow for targeted book selection and instruction.
Tutoring Program	Tutoring was scheduled for between 45 and 60 minutes depending on schools' schedules. The tutoring curriculum was designed by teachers and middle school literacy experts, and modeled after the guided reading framework described in Fountas and Pinnell (2001). A 60-minute tutoring session would consist of 10 minutes discussing an introductory question and vocabulary review, 40 minutes of 1-on-1 reading with the tutor and independent reading, and 10 minutes of wrap-up and discussion. Tutoring was held during after-school programming in year one; in years two and three most schools moved tutoring to be during the school day.
Outcomes of Interest	*Direct Outcomes:* Number of books and pages read in tutoring, attendance rate at tutoring *Student Achievement:* Student attendance at school, state test scores in mathematics and ELA
Testing Windows	*2014:* English Language Arts: 4/1-4/3; Mathematics: 4/30-5/1 *2015:* English Language Arts: 4/14-4/16; Mathematics: 4/22-4/24 *2016:* English Language Arts: 4/5-4/7; Mathematics: 4/13-4/15

Table 2: Pre-Treatment Summary Statistics

	Non-Exp	Exp	*p-value*	Supp. Control	Main Control	Treatment	*p-value*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: School Characteristics							
Percent Female	0.479	0.482	0.839	0.492	0.500	0.453	0.239
Percent Black	0.365	0.408	0.287	0.381	0.361	0.483	0.423
Percent Hispanic	0.390	0.483	0.018	0.497	0.519	0.433	0.675
Percent White	0.124	0.039	0.000	0.017	0.056	0.045	0.093
Percent Asian	0.096	0.060	0.043	0.096	0.052	0.031	0.345
Percent Other Race	0.026	0.009	0.000	0.008	0.012	0.008	0.464
Percent English Language Learners	0.116	0.193	0.000	0.229	0.198	0.151	0.301
Percent Special Education Services	0.237	0.214	0.050	0.209	0.210	0.223	0.752
Percent Economically Disadvantaged	0.739	0.852	0.000	0.883	0.836	0.835	0.077
Mean Attendance Rate 12-13	0.918	0.922	0.326	0.922	0.922	0.922	0.992
Mean Math Score 12-13 (σ)	-0.150	-0.421	0.000	-0.428	-0.416	-0.418	0.992
Mean ELA Score 12-13 (σ)	-0.143	-0.353	0.000	-0.401	-0.318	-0.341	0.621
Mean Math Score 11-12 (σ)	-0.154	-0.384	0.000	-0.389	-0.395	-0.369	0.952
Mean ELA Reading Score 11-12 (σ)	-0.148	-0.350	0.000	-0.406	-0.317	-0.328	0.541
Number of Schools	615	54		18	18	18	
p-value from joint F-test			*0.000*				*0.954*
Panel B: Student Characteristics							
6th Grade Cohort							
Female	0.493	0.478	0.272	0.485	0.487	0.462	0.713
Black	0.280	0.372	0.044	0.323	0.358	0.442	0.553
Hispanic	0.398	0.482	0.081	0.533	0.472	0.439	0.682
White	0.154	0.047	0.000	0.015	0.065	0.060	0.065
Asian	0.157	0.088	0.010	0.120	0.088	0.053	0.430
Other Race	0.011	0.012	0.899	0.009	0.018	0.007	0.346
English Language Learner	0.148	0.218	0.002	0.238	0.231	0.183	0.489
Special Education Services	0.185	0.205	0.032	0.196	0.212	0.207	0.722
Economically Disadvantaged	0.726	0.835	0.000	0.856	0.812	0.839	0.540
Attendance Rate 12-13	0.948	0.935	0.000	0.932	0.936	0.935	0.762
Math Score 12-13 (σ)	0.037	-0.383	0.000	-0.366	-0.398	-0.384	0.968
ELA Score 12-13 (σ)	0.030	-0.311	0.000	-0.339	-0.293	-0.302	0.920
Math Score 11-12 (σ)	0.046	-0.347	0.000	-0.341	-0.375	-0.322	0.857
ELA Reading Score 11-12 (σ)	0.037	-0.317	0.000	-0.344	-0.321	-0.284	0.876
Number of 6th Graders	66470	5988		2008	2109	1871	
p-value from joint F-test			*0.000*				*0.147*
Sample Eligible for Tutoring							
Female	—	—	—	—	0.503	0.479	0.467
Black	—	—	—	—	0.411	0.475	0.528
Hispanic	—	—	—	—	0.448	0.445	0.980
White	—	—	—	—	0.065	0.041	0.495
Asian	—	—	—	—	0.063	0.030	0.234
Other Race	—	—	—	—	0.013	0.010	0.501
English Language Learner	—	—	—	—	0.199	0.174	0.453
Special Education Services	—	—	—	—	0.211	0.192	0.447
Economically Disadvantaged	—	—	—	—	0.811	0.865	0.163
Attendance Rate 12-13	—	—	—	—	0.937	0.937	0.894
Math Score 12-13 (σ)	—	—	—	—	-0.423	-0.455	0.584
ELA Score 12-13 (σ)	—	—	—	—	-0.379	-0.407	0.557
Math Score 11-12 (σ)	—	—	—	—	-0.402	-0.426	0.667
ELA Reading Score 11-12 (σ)	—	—	—	—	-0.369	-0.382	0.795
Number of 6th Graders Eligible for Tutoring	—	—		—	891	837	
p-value from joint F-test							*0.405*

Notes: This table reports school- and student-level pre-treatment summary statistics. In Panel A, schools are only included if they have at least 3 6th grade students enrolled in 2013-14. In Panel B, students are only included in the sample if they have at least one valid test score outcome variable in 2013-14 and are enrolled in grade 6 in 2013-14. Column (1) reports the mean of the non-experimental group. Column (2) reports the mean of the experimental group. Column (3) reports the p-value on the null hypothesis of equal means in the experimental and non-experimental groups. To calculate the p-value, we divide the means by the square root of the average variance of teh non-experimental and experimental groups, following Winship and Morgan (2007). We then regress the adjusted variable on a dummy indicator denoting whether the observation was experimental or not and bootstrap the standard errors with 500 replications (Efron, 1979). Columns (4)-(7) report similar values for the two treatment groups and the control group, where the p-value in Column (7) is on the null hypothesis of equal means across the three groups. Standard errors are also clustered at the school level for Panel B. All demographic and test score measures are culled from administrative data collected pre-treatment. See the Online Appendix for details on variable construction. Student test scores and teacher effect measures are standardized to have a mean of zero and standard deviation one over the district sample by grade and by subject, respectively. Note that the schools that were dropped from the main analysis due to their main control schools not administering the DRP are included in Column (1) only.

Table 3: Effects on Direct Outcomes (ITT)

	2013-14	2014-15	2015-16	Pooled
	(1)	(2)	(3)	(4)
Attendance Rate at Tutoring	0.472***	0.337***	0.333***	0.381***
	(0.025)	(0.023)	(0.037)	(0.024)
	1,769	1,691	1,630	5,090
Number of Days Attended Tutoring	82.186***	59.042***	59.630***	67.011***
	(4.394)	(3.987)	(6.652)	(4.162)
	1,769	1,691	1,630	5,090
Number of Books Read	2.537***	1.433***	1.448***	1.813***
	(0.228)	(0.165)	(0.176)	(0.169)
	1,769	1,691	1,630	5,090
Number of Pages Read	600.733***	286.397***	313.783***	402.660***
	(45.238)	(31.509)	(36.610)	(29.723)
	1,769	1,691	1,630	5,090
Number of Words Read	99035***	42421***	57421***	66610***
	(7561)	(4901)	(7283)	(5163)
	1,769	1,691	1,630	5,090

Notes: This table reports ITT results of the tutoring program. The sample is students in treatment and main control schools with DRP scores within the range that would qualify them to receive tutoring services. Students who qualify for tutoring or after-school programs in years 2014-15 and 2015-16 who are not in the experimental cohort are not included in any sample in those years. Students are assigned to the school that they are enrolled in by October 31st in the first year of treatment. All specifications control for matched-pair fixed effects and the student-level demographics summarized in Table 2 plus three years of baseline reading and math scores and their squares. All controls are interacted with indicators for whether a student is Hispanic, black, or other race. Standard errors, reported in parentheses, are clustered at the school level. Significance at the 1%, 5%, and 10% levels indicated by ***, **, and *, respectively.

Table 4: Effects on Student Achievement and Attendance (ITT)

	2013-14	2014-15	2015-16	Pooled
	(1)	(2)	(3)	(4)
Panel A: Effects on State Test Scores				
Math	-0.073	-0.022	0.099	-0.002
	(0.056)	(0.050)	(0.122)	(0.056)
	1,715	1,658	1,454	4,827
ELA	0.075*	0.020	0.045	0.045
	(0.040)	(0.039)	(0.052)	(0.036)
	1,718	1,663	1,578	4,959
Reading Subscore	0.041	0.012	0.013	0.020
	(0.031)	(0.034)	(0.047)	(0.029)
	1,718	1,663	1,578	4,959
Writing Subscore	0.148*	0.000	0.090	0.078
	(0.073)	(0.085)	(0.080)	(0.064)
	1,718	1,663	1,578	4,959
Panel B: Effects on Student Attendance at School				
Attendance	0.012***	0.010***	0.014***	0.012***
	(0.004)	(0.003)	(0.003)	(0.003)
	1,759	1,682	1,605	5,046
Control Mean	0.921	0.921	0.914	0.919

Notes: This table reports ITT results of the tutoring program. The sample is students in treatment and main control schools with DRP scores within the range that would qualify them to receive tutoring services. Students who qualify for tutoring or after-school programs in years 2014-15 and 2015-16 who are not in the experimental cohort are not included in any sample in those years. Students are assigned to the school that they are enrolled in by October 31st in the first year of treatment. Testing variables are drawn from district test score files and are standardized to have a mean of zero and a standard deviation of one within each year and grade among students with valid test scores. All specifications control for matched-pair fixed effects and the student-level demographics summarized in Table 2 plus three years of baseline reading and math scores and their squares. When the outcome is student attendance, controls also include student attendance in the year prior to treatment. All controls are interacted with indicators for whether a student is Hispanic, black, or other race. Standard errors, reported in parentheses, are clustered at the school level. Significance at the 1%, 5%, and 10% levels indicated by ***, **, and *, respectively.

Table 5: Effects on Student State Test Scores and Attendance by Subsample (ITT)

	Math	*p-value*	ELA	*p-value*	Attendance	*p-value*
	(1)	(2)	(3)	(4)	(5)	(6)
Full Sample	-0.002		0.045		0.012***	
	(0.056)		(0.036)		(0.003)	
	4,827		4,959		5,046	
Panel A: Demographics						
Black	0.101*	0.012	0.089**	0.053	0.020***	0.019
	(0.056)		(0.033)		(0.004)	
	2,091		2,168		2,209	
Hispanic	-0.079		0.010		0.008**	
	(0.057)		(0.037)		(0.004)	
	2,201		2,244		2,277	
Econ. Disadvantaged - Yes	-0.014	0.096	0.041	0.951	0.012***	0.364
	(0.055)		(0.035)		(0.003)	
	4,030		4,149		4,219	
Econ. Disadvantaged - No	0.117		0.037		0.017***	
	(0.075)		(0.060)		(0.005)	
	797		810		827	
Special Ed. - Yes	-0.015	0.638	-0.046	0.055	0.005	0.262
	(0.056)		(0.036)		(0.006)	
	992		1,009		1,026	
Special Ed. - No	0.007		0.059		0.012***	
	(0.057)		(0.041)		(0.003)	
	3,835		3,950		4,020	
ELL - Yes	0.033	0.538	0.065	0.747	0.004	0.111
	(0.080)		(0.050)		(0.004)	
	920		918		939	
ELL - No	-0.010		0.047		0.014***	
	(0.057)		(0.039)		(0.003)	
	3,907		4,041		4,107	
English Spoken at Home	0.029	0.201	0.072*	0.385	0.016***	0.068
	(0.053)		(0.039)		(0.004)	
	2,864		2,960		3,027	
Spanish Spoken at Home	-0.053		0.031		0.007*	
	(0.067)		(0.045)		(0.003)	
	1,461		1,486		1,499	
Born in the US	0.001	0.596	0.062	0.011	0.014***	0.093
	(0.056)		(0.042)		(0.003)	
	4,014		4,130		4,197	
Not Born in the US	-0.042		-0.152**		0.004	
	(0.077)		(0.060)		(0.004)	
	807		823		843	

Panel B: Reading Ability

DRP Quartile 1	0.056	0.078	0.050	0.607	0.018***	0.000
	(0.088)		(0.051)		(0.005)	
	1,419		1,448		1,480	
DRP Quartile 2	0.001		0.043		0.024***	
	(0.056)		(0.054)		(0.005)	
	1,231		1,259		1,277	
DRP Quartile 3	-0.023		0.024		0.021***	
	(0.067)		(0.051)		(0.006)	
	1,049		1,073		1,095	
DRP Quartile 4	-0.117*		-0.026		-0.007	
	(0.063)		(0.046)		(0.005)	
	1,128		1,179		1,194	
State ELA Quartile 1	0.053	0.000	0.077	0.303	0.017**	0.000
	(0.055)		(0.047)		(0.006)	
	1,597		1,626		1,658	
State ELA Quartile 2	-0.041		0.067*		0.018***	
	(0.061)		(0.038)		(0.004)	
	2,042		2,091		2,115	
State ELA Quartile 3	-0.067		-0.003		-0.011**	
	(0.067)		(0.055)		(0.005)	
	867		911		925	
State ELA Quartile 4	0.696***		0.119***		0.131***	
	(0.023)		(0.026)		(0.004)	
	133		143		143	
Panel C: Neighborhood Characteristics						
Above-Med Household Income	0.084	0.044	0.086	0.226	0.015***	0.165
	(0.075)		(0.055)		(0.004)	
	2,364		2,433		2,481	
Below-Med Household Income	-0.108*		0.004		0.007*	
	(0.054)		(0.036)		(0.004)	
	2,463		2,526		2,565	
Above-Med Num Stop Frisk	-0.000	0.935	0.049	0.938	0.016***	0.305
	(0.108)		(0.052)		(0.004)	
	2,178		2,238		2,282	
Below-Med Num Stop Frisk	-0.010		0.043		0.010**	
	(0.052)		(0.050)		(0.004)	
	2,649		2,721		2,764	
Above-Med Single Parent Households	-0.120**	0.019	-0.002	0.230	0.014***	0.417
	(0.051)		(0.033)		(0.005)	
	2,687		2,747		2,805	
Below-Med Single Parent Households	0.111		0.082		0.009***	
	(0.079)		(0.061)		(0.003)	
	2,140		2,212		2,241	
Panel D: Tutor Characteristics						
Above-Med Ave. Tutor Interview Score	-0.023	0.522	0.062	0.807	0.009**	0.015
	(0.089)		(0.051)		(0.004)	
	2,380		2,456		2,496	
Below-Med Ave. Tutor Interview Score	0.048		0.076*		0.018***	
	(0.063)		(0.038)		(0.003)	
	2,138		2,180		2,228	
Above-Med % Tutors with Eng./Ed.	0.059	0.472	0.073*	0.532	0.008**	0.063

	(0.107)		(0.041)		(0.003)	
	2,173		2,270		2,316	
Below-Med % Tutors with Eng./Ed.	-0.025		0.052		0.017***	
	(0.053)		(0.040)		(0.004)	
	2,345		2,366		2,408	
Above-Med % Tutors with Teaching Exp.	-0.119**	0.000	0.054	0.731	0.014***	0.605
	(0.053)		(0.034)		(0.003)	
	2,557		2,602		2,650	
Below-Med % Tutors with Teaching Exp.	0.183***		0.071		0.012***	
	(0.061)		(0.052)		(0.004)	
	1,961		2,034		2,074	
Above-Med % Tutors with Tutoring Exp.	0.035	0.733	0.043	0.464	0.012**	0.823
	(0.092)		(0.041)		(0.005)	
	2,295		2,372		2,410	
Below-Med % Tutors with Tutoring Exp.	-0.008		0.086*		0.013***	
	(0.074)		(0.046)		(0.003)	
	2,223		2,264		2,314	
Above-Med % Tutors with Grad. Degree	0.039	0.801	0.050	0.754	0.015***	0.780
	(0.125)		(0.060)		(0.005)	
	1,632		1,704		1,738	
Below-Med % Tutors with Grad. Degree	0.002		0.069**		0.013***	
	(0.060)		(0.033)		(0.003)	
	2,886		2,932		2,986	
Above-Med % Tutors Black	0.137**	0.003	0.075*	0.257	0.017***	0.046
	(0.066)		(0.043)		(0.004)	
	2,266		2,347		2,397	
Below-Med % Tutors Black	-0.147**		0.022		0.008**	
	(0.062)		(0.042)		(0.003)	
	2,561		2,612		2,649	
Above-Med % Tutors Hispanic	-0.109	0.051	-0.072	0.003	0.003	0.011
	(0.066)		(0.048)		(0.004)	
	1,685		1,734		1,768	
Below-Med % Tutors Hispanic	0.041		0.092***		0.017***	
	(0.059)		(0.034)		(0.004)	
	3,142		3,225		3,278	
Above-Med % Tutors White	-0.003	0.933	0.063	0.577	0.010***	0.342
	(0.084)		(0.052)		(0.003)	
	2,569		2,659		2,698	
Below-Med % Tutors White	0.006		0.029		0.015***	
	(0.065)		(0.038)		(0.005)	
	2,258		2,300		2,348	
Panel E: School Characteristics						
Above-Med % ELL	-0.032	0.212	0.040	0.808	0.006**	0.001
	(0.075)		(0.048)		(0.003)	
	3,533		3,630		3,685	
Below-Med % ELL	0.074*		0.057		0.032***	
	(0.038)		(0.047)		(0.006)	
	1,294		1,329		1,361	
Above-Med % Special Education	-0.016	0.688	-0.007	0.072	0.009**	0.535
	(0.077)		(0.039)		(0.004)	
	2,281		2,349		2,384	

Below-Med % Special Education	0.027		0.111**		0.012***	
	(0.072)		(0.050)		(0.004)	
	2,546		2,610		2,662	
Above-Med % Free/Reduced Lunch	-0.065	0.406	-0.006	0.345	0.008	0.363
	(0.066)		(0.048)		(0.006)	
	1,576		1,597		1,633	
Below-Med % Free/Reduced Lunch	0.018		0.058		0.014***	
	(0.074)		(0.048)		(0.004)	
	1,576		1,597		1,633	
Above-Med % Black	0.163***	0.000	0.130***	0.005	0.021***	0.015
	(0.047)		(0.039)		(0.006)	
	2,260		2,336		2,385	
Below-Med % Black	-0.174***		-0.039		0.004	
	(0.062)		(0.042)		(0.003)	
	2,567		2,623		2,661	
Above-Med % Hispanic	-0.098*	0.071	0.015	0.341	0.008**	0.264
	(0.053)		(0.036)		(0.004)	
	2,545		2,609		2,652	
Below-Med % Hispanic	0.077		0.080		0.015***	
	(0.078)		(0.057)		(0.005)	
	2,282		2,350		2,394	
Above-Med % of 6th Graders Tutored	-0.009	0.786	0.014	0.381	0.017***	0.144
	(0.082)		(0.046)		(0.005)	
	2,114		2,174		2,207	
Below-Med % of 6th Graders Tutored	0.020		0.076		0.008**	
	(0.068)		(0.052)		(0.004)	
	2,713		2,785		2,839	

Notes: This table reports ITT results of the tutoring program from the pooled specification in Column (4) of Table 4. The sample is students in treatment and main control schools with DRP scores within the range that would qualify them to receive tutoring services. Students who qualify for tutoring or after-school programs in years 2014-15 and 2015-16 who are not in the experimental cohort are not included in any sample in those years. Students are assigned to the school that they are enrolled in by October 31st in the first year of treatment. Testing variables are drawn from district test score files and are standardized to have a mean of zero and a standard deviation of one within each year and grade among students with valid test scores. All specifications control for matched-pair fixed effects and the student-level demographics summarized in Table 2 plus three years of baseline reading and math scores and their squares. When attendance is the outcome variable, controls include students' attendance rate in the year prior to treatment. All controls are interacted with indicators for whether a student is Hispanic, black, or other race. Standard errors, reported in parentheses, are clustered at the school level. Significance at the 1%, 5%, and 10% levels indicated by ***, **, and *, respectively.

Table 6: Understanding Racial Differences in Treatment Effects

	ELA Effect		Attendance Effect	
	Difference in TE	% Increase from Row 1	Difference in TE	% Increase from Row 1
Main Controls	-0.078*	—	-0.012**	—
	(0.039)		(0.005)	
+ Language and Native Country	-0.076*	3%	-0.011**	8%
	(0.038)		(0.005)	
+ Neighborhood Characteristics	-0.065	17%	-0.012**	0%
	(0.040)		(0.005)	
+ Diagnostic Test Scores	-0.055	29%	-0.013**	-8%
	(0.036)		(0.005)	
+ ELA State Test and DRP Quartile FE	-0.053	32%	-0.013**	-8%
	(0.036)		(0.005)	
+ School Average Tutor Characteristics	0.007	109%	-0.005	58%
	(0.087)		(0.008)	
+ Student Attendance at School	0.011	114%	—	
	(0.085)			

Notes: This table reports differences in the ITT effect for Hispanic and black students, controlling for additional student, school, and tutor characteristics. In Column (1) the dependent variable is state ELA test scores; in Column (3) it is student attendance rates at school. Row one controls for matched-pair fixed effects and the student-level demographics summarized in Table 2 plus three years of baseline reading and math scores and their squares. When the outcome is student attendance, controls also include student attendance in the year prior to treatment. Row two additionally controls for whether a student speaks English, Spanish, or another language at home and whether a student was born in the US. Row three additionally controls for the the median household income, poverty rate, percent of single parent households, number of stop and frisks, and percent of stop and frisks that are tagged as being in a high-crime area of the census tract that each school is in. Row four additionally controls for students' diagnostic score on the Degrees of Reading Power assessment that was used to assign students to tutoring. Row five additionally controls for fixed effects for students pre-treatment ELA test score quartile and DRP quartile. Row six additionally controls for tutor characteristics including the percent of tutors with an English or Education degree, the average tutor quality score from the interview process, the percent of tutors with teaching experience, the percent of tutors with tutoring experience, the percent of tutors with any graduate degree, and the percent of tutors who are black, Hispanic, and white. Row seven additionally controls for students' attendance at school. All controls are interacted with indicators for whether a student is Hispanic, black, or other race. Missings are included for all control variables. Standard errors, reported in parentheses, are clustered at the school level. Significance at the 1%, 5%, and 10% levels indicated by ***, **, and *, respectively.