Northwestern | INSTITUTE FOR POLICY RESEARCH

IPR Working Paper Series

WP-21-12

# Not Too Late: Improving Academic Outcomes Among Adolescents

Jonathan Guryan
Northwestern University and IPR

Jens Ludwig
University of Chicago

Monica Bhatt
University of Chicago

Philip Cook
Duke University

Jonathan Davis
University of Oregon

Kenneth Dodge
Duke University

IPR Working Paper Series

WP-21-12

**George Farkas**
University of California, Irvine

**Roland Fryer Jr.**
Harvard University

**Susan Mayer**
University of Chicago

**Harold Pollack**
University of Chicago

**Laurence Steinberg**
Temple University

Version: March 8, 2021

**DRAFT**

*Please do not quote or distribute without permission.*

# Abstract

There is growing concern that it is too difficult or costly to substantially improve the academic skills of children who are behind in school once they reach adolescence. But perhaps what we have tried in the past relies on the wrong interventions, failing to account for challenges like the increased variability in academic needs during adolescence, or heightened difficulty of classroom management. This study tests the effects of one intervention that tries to solve both problems by simplifying the teaching task: individualized, intensive, in-school tutoring. A key innovation by the non-profit the researchers study (Saga Education) is to identify how to deliver "high-impact tutoring" at relatively low cost ($3,500 to $4,300 per participant per year). Their first randomized controlled trial (RCT) of Saga's tutoring model with 2,633 9th and 10th grade students in Chicago public schools found participation increased math test scores by 0.16 standard deviations (SDs) and increased grades in math and non-math courses. The authors replicated these results in a separate RCT with 2,710 students and found even larger math test score impacts—0.37 SD—and similar grade impacts. These effects persist into future years, although estimates for high school graduation are imprecise. The treatment effects do not appear to be the result of a generic "mentoring effect" or of changes in social-emotional skills, but instead seem to be caused by changes in the instructional "technology" that students received. The estimated benefit-cost ratio is comparable to many successful model early-childhood programs.

*This paper was made possible by the generous support of the Abbvie, Bank of America, Laura and John Arnold, Paul M. Angell, Edna McConnell Clark, Crown, Lloyd A. Fry, Joyce, Logan, MacArthur, Polk, Pritzker-Pucker, Smith Richardson, and Spencer foundations, the Chicago Center for Youth Violence Prevention, EquiTrust, JPAL-North America, the Illinois Criminal Justice Information Authority, the city of Chicago, grant number 2012-JU-FX-0019 from the Office of Juvenile Justice and Delinquency Prevention, Office of Justice Programs, US Department of Justice, and award number 1P01HD076816 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health. The authors are also grateful for operating grants to the University of Chicago Crime Lab from Susan and Tom Dunn, Ira Handler, and the MacArthur, McCormick and Pritzker-Pucker foundations. For vital assistance in making this work possible, the authors thank Chad Adams, Roseanna Ander, Barbara Byrd-Bennett, Trayvon Braxton, Valerie Chang, Akeshia Craven, Rukiya Curvey-Johnson, Gretchen Cusick, Aarti Dhupelia, Tony DiVittorio, Mayor Rahm Emanuel, Wendy Fine, Jaureese Gaines, Michael Goldstein, Antonio Gutierrez, Craig Howard, Tim Jackson, Barbara Kelley, Ed Klunk, Timothy Knowles, Tim Lavery, Stig Leschly, Jonathan Lewin, Daniel Lopez, Julia Quinn, Arnaldo Rivera, Janey Rountree, Alan Safran, Leonetta Sanders, Maitreyi Sistla, Julia Stasch, Sara Stoelinga, Elizabeth Swanson, Robert Tracy, Karen Van Ausdal, and John Wolf, as well as the staffs of the Chicago Public Schools system, Match Education, and Saga Education. Thanks to Jeffrey Broom, Sarah Dickson, Kylie Klein, Stacy Norris, and Jared Sell for their help in accessing the data we analyze here, and to Nathan Hess, Stephanie Kirmer, Daniel Kowalski, Khoa Nguyen, Kelsey Reid, Michael Rosenbaum, Catherine Schwarz, Noah Sebek, Robert Webber, Nathan Weil, and David Welgus for their invaluable contributions to the data analysis. For useful suggestion the authors thank seminar participants at New York University and the Institute for Research on Poverty at the University of Wisconsin. Points of view or opinions in this document are those of the authors and do not necessarily represent the views of the US Department of Justice, National Institutes of Health or any other funder. Any errors are of course their own. This RCT was registered in the American Economic Association Registry for randomized control trials under trial number AEARCTR-0000041.*

## I. INTRODUCTION

Education is one of the American public's top priorities, but both the data and public opinion suggest there is room for improvement—perhaps especially for adolescents.[1] The failure to improve schooling outcomes at a pace to keep up with technological change has harmed economic growth, leading to a dramatic rise in the returns to schooling over time (Goldin and Katz, 2008). Long-term trends in test scores seem more encouraging for relatively younger than older students (Hanushek et al., 2020). Furthermore, disparities in achievement test scores between students who are academically succeeding versus struggling seem to grow rather than shrink as children progress through school (Cascio and Staiger, 2012). Studies of educational interventions for teens tend to yield much more disappointing results than of interventions for young children.[2] And parents of public high school students are notably less satisfied with the schools their children attend than are parents of elementary-school children.[3]

The relatively disappointing academic outcomes and interventions for teens relative to younger children has raised the concern that perhaps it is too late to substantially increase human capital and remediate disparities once children reach adolescence, due to causes that are fundamental and unavoidable. For example, developmental plasticity declines from early to

[1] See for example https://www.pewresearch.org/politics/2019/01/24/publics-2019-priorities-economy-health-care-education-and-security-all-near-top-of-list/ and https://news.gallup.com/poll/1612/education.aspx

[2] For young children, examples include Perry Preschool (Schweinhart et al., 2005; Belfield et al., 2006; Heckman et al., 2010) and Carolina Abecedarian (Campbell et al., 2002; Barnett & Masse, 2007). A number of studies of state-sponsored universal pre-K programs have used a research design that exploits a regression discontinuity in age (see for example Gormley & Phillips, 2008). For a discussion of some concerns about that design, and a review of the evidence on the federal Head Start program, see Ludwig and Phillips (2007) and Gibbs, Ludwig and Miller (2013). For excellent reviews of the literature for older children see Carneiro & Heckman (2003) or Heckman (2012).

[3] Data from the 2012 National Household Education Survey shows that in public school districts where the district assigns schools (as opposed to for instance choice districts), 66% of parents with children 5-10 are very satisfied with the school their child attends compared to 47% of those with children 14-18 (Cheng and Peterson 2017).

middle childhood, although may increase again during adolescence (Shonkoff & Phillips, 2000; Steinberg, 2014; Bonnie & Backes, 2019). Motivation becomes more challenging as students age, as reflected by increased prevalence of disruptive behaviors.[4] The growing variability of academic achievement as children age means the variability of what students need within a given classroom is also heightened for older children (Cascio and Staiger, 2012). No wonder teachers report that the two most challenging parts of their job are individualizing instruction and classroom management.[5] Thus Carneiro and Heckman's (2003, p. 90) pessimistic conclusion: "The return to [human capital] investment in the young is apparently quite high; the return to investment in the old and less able is quite low."

The hypothesis tested in this paper is that the limited success of previous efforts to improve academic outcomes for adolescents is not because success is impossible, but rather because of a reliance in the past on the wrong intervention approach, or "technology."[6] In some sense an instructional technology that works with students of any age has long been known: intensive tutoring. This method of instruction dates back at least to the 15th century at Oxford. Randomized experiments comparing tutoring to classroom instruction confirm the former to be

[4] Disciplinary actions in school increase with age (https://nces.ed.gov/programs/raceindicators/indicator_RDA.asp). This is not just a function of changing disciplinary standards by schools since we also see increases in absences (https://www2.ed.gov/datastory/chronicabsenteeism.html) and arrests, including for serious offenses where presumably the discretion of police to arrest a suspect or not is relatively limited.

[5] For example in the School and Staffing Survey (SASS), 43% of new elementary school teachers and 47% of new secondary school teachers say they felt not at all prepared or only somewhat prepared to deal with classroom management; 41% of new elementary school teachers and 44% of new secondary-school teachers said they were unprepared or only somewhat prepared to differentiate instruction. (From original author tabulations of SASS data).

[6] Examples include generally disappointing studies of the Job Corps (Long et al., 1981), Catholic schools, which may increase graduation but do not seem to increase achievement test scores (Grogger and Neal, 2000, Altonji, Elder and Taber, 2005), and accountability or school voucher programs (Rouse and Barrow, 2006). Studies of increased school funding typically document the effects of changes across all grades, which makes it somewhat difficult to determine the impacts on adolescents specifically (see for example the review in Jackson, 2018).

"the best learning conditions we can devise" (Bloom, 1984, p. 4). The challenge to widespread implementation has not been a pedagogical problem so much as an economic one—that is, cost.

The key insight behind the specific educational intervention we study here, developed by Saga Education, is that being a successful tutor requires fewer specialized skills and less on-the-job-learning than does classroom teaching.[7] This insight expands the applicant pool, and also means that the people hired—tutors—can be successful right away (rather than the usual three-year learning curve for classroom teachers[8]). Saga therefore uses a very different HR model from regular education: It hires people to tutor for just one year for a modest stipend, as a sort of public service. Rather than boost teacher quality, which has been a key focus of much education policy, Saga simplifies what instructors are asked to do by changing the teaching task itself.

We carried out two separate large-scale randomized controlled trials (RCTs) with thousands of teens who essentially formed a representative sample of students enrolled in some of Chicago's (and hence America's) most economically disadvantaged neighborhood schools. We focused on math because failing core math classes is a driver of dropout in Chicago (Allensworth and Easton, 2005), because math is so important for success in school and later earnings (Duncan et al., 2007), and because math skills for school-age children may be more responsive to school-based interventions than reading skills (e.g. Fryer, 2014, 2017).

In the first RCT ("study 1"), our research team randomly assigned 2,633 male youths in summer 2013 who were rising 9th or 10th graders to a treatment group that was offered Saga

[7] At the time we initially started working with Saga tutoring they were part of the Match charter school organization of Boston, Massachusetts. In 2015, executives from Match Education spun off from Match to form SAGA Innovations, a national non-profit that aims to bring this tutoring model into traditional public school systems across the country. SAGA Innovations changed their name to Saga Education in 2019.

[8] See e.g. Rivkin, Hanushek and Kain, 2005; Boyd, Grossman, Lankford, Loeb and Wykoff, 2006; Clotfelter, Ladd and Vigdor, 2007 and 2010; Goldhaber, 2007; and Kane and Staiger, 2008.

tutoring or a control group that was not. Students were drawn from 12 CPS high schools mostly on the west and south sides of Chicago. After one year of program participation the intention-to-treat effect (ITT) on standardized math achievement test scores is 0.08 standard deviations (SD), and the treatment-on-the-treated (TOT) effect is 0.16 SD. This gain is about what the average high school student learns in a year, so the intervention (roughly speaking) doubles the annual math test score gain.[9] These gains in test scores do not appear to be the result of tutors narrowly teaching to the test; we also see a TOT effect on grades in regular classroom math courses equal to 0.56 points (on a 0 to 4-point scale), and a decline in math course failures of 49%.

Motivated in part to see if these results could be replicated at this large scale, in the summer of 2014 we randomized a separate sample of 2,710 9th and 10th graders ("study 2"). The TOT effect on math scores after one year is roughly twice as large as in study 1 (0.37 SD). We also find sizable positive effects on math grades in study 2 similar to the findings for study 1.

What is the mechanism behind these effects? Given that past efforts to improve meta-cognition and social-emotional skills for adolescents and young adults have been more encouraging than efforts to improve academic skills (for example Blattman et al., 2017, Heller et al., 2017), one might wonder whether tutoring—which puts youth in close contact with a pro-social adult every day—is just operating through a generic "mentoring effect." But survey data show no detectable effects on whether youth report feeling close to or supported by adults in school or out of school, or on grit, conscientiousness, or locus of control. Nor do we find evidence consistent with the idea of classroom management being the key mechanism in the sense that the impacts are no larger in schools where disciplinary incidents in classrooms at

[9] Reardon (2011) finds in the NAEP that the average high school student's scores in reading and math increase by around 0.6 to 0.7 standard deviations over 4 years, or approximately 0.15-0.18 standard deviations per year.

baseline are more common. We do find gains in whether students like math and care about their math grades—mediators closely tied to math instruction and achievement specifically.

We also present suggestive findings consistent with, although certainly not definitive proof for, better 'personalization' of instruction as a relevant mechanism. Some studies find classroom teachers orient material towards the top third of the classroom (Bloom, 1984),[10] which would suggest students at the bottom of the baseline achievement distribution might benefit most from personalized instruction with tutoring. But identification of this pathway is complicated by signs of potential 'floor effects' in the achievement tests administered by CPS. To explore floor effects (the possibility that students learn more math, but their skills are still below the easiest items included on a given test) we estimate heterogeneous treatment effects using the machine learning approach of Athey, Tibshirani and Wager (2019). About one-quarter of the sample have large gains in math grades, but no gain in math test scores, consistent with floor effects on the tests. (This could simply reflect increased effort, but we see no signs of that in survey data.)

We then exploit the fact that we administered our own math tests as well, for which we have item-level data. When we rank-order items by difficulty and estimate effects for specific items we see suggestive (but imprecisely estimated) indications that students with below-median baseline scores gain most on the easiest items, while above-median students gain most on harder items. This suggestive result is consistent with Duflo et al.'s (2011) finding that personalization by tracking improved learning outcomes for students in *both* the top *and* bottom halves of the achievement distribution.[11]

[10] Previous research suggests there can be mismatches between the developmental needs of youth and their social environments, also called "stage-environment fit" (see Hunt, 1975; Eccles et al. 1993). The same sort of mismatch may occur for youths' academic needs as well (see for example, Engel, Claessens and Finch 2012).

[11] See Figlio & Page, 2002 for a similar finding among initially lower-performing students in the US.

One way to read this study is as a program evaluation of an intervention that substantially improves learning of adolescents, at a cost of $3,500 to $4,300 per year per student.[12] (By way of comparison average CPS spending per pupil is about $17,000 per year.)[13] The ultimate cost effectiveness of the program depends on whether impacts persist and can help students who are behind get back up to grade level to engage effectively with regular classroom instruction, which affects whether students need this support consistently or just on a one-time basis. When we look at 11th grade outcomes, a year or two after tutoring, we see persistent gains in math test scores of 0.22 SD (pooling study 1 and 2) and improved math grades of 0.18 GPA points. The point estimate for graduation is positive, 1.9 percentage points, but imprecisely estimated; we cannot rule out declines of up to 4.8 points or gains as large as 8.6 points. Our findings echo Banerjee et al. (2007)'s study in India of gains from substituting a simpler teaching task for teacher skill.[14]

A second way to read this study is as a test of whether it is really too late to substantially improve academic outcomes of children who have reached adolescence. We find it is possible to perhaps double, or even triple, how much math students learn in a year, for something like one-fifth of the average cost of schooling per pupil per year. These sizable gains are consistent with studies of 'no-excuses' charter schools for high school students, which, interestingly, also often

[12] At the time of this intervention (2013-2015) the per-pupil cost of Saga was approximately $3,800 with a defensible range of $3,500 to $4,300, as stated. Subsequently Saga has dropped its charge to districts from $3,800 per-pupil (2013-2015) to $3,100 per-pupil (2015-2019) to now $1,800 per-pupil as of the time of release of this paper by obtaining an AmeriCorps subsidy of $15,000 per fellow and using a blended-learning model, in which the student:tutor ratio is 4:1 in lieu of 2:1 and students spend half their time on a learning platform, e.g. ALEKS.

[13] This is "operating expenditure" excluding summer school, adult education, capital expenditures and debt. https://www.illinoisreportcard.com/district.aspx?source=environment&source2=perstudentspending&Districtid=15016299025

[14] Banerjee et al. (2007) found that assigning third and fourth graders in India who are far behind to receive instruction in remedial skills for two hours per day in a classroom of 15-20 students (who are thus fairly homogenous in academic level) increased test scores by around 0.60 SD. The instructors for these classes were women from the local community who were trained for just a short period of time and paid only $10-15 per month. The effects of a computer-assisted program that also individualized instruction were found to increase test scores by up to 0.47 SD after the second year of intervention, although impacts from both strategies were short-lived.

include intensive or ‘high-impact’ tutoring as an element (Fryer, 2014; Fryer 2015; Dobbie & Fryer, 2019; Tuttle, et al., 2015; Angrist et al., 2016; Abdulkadiroglu, et al., 2017). Extrapolating these test score impacts to earnings gains based on the best estimates in the literature implies that the benefit-cost ratio of the program is comparable to both exemplar early childhood model programs, like the Abecedarian Project and the Perry Preschool Program, as well as larger-scale efforts to improve outcomes for younger children such as the Tennessee Star class size reduction experiment.

## II. THE INTERVENTION

We selected Saga Education’s tutoring model to study in part because of its low cost relative to the intensity of the intervention. One major innovation is the recognition that the instructional “technology” of tutoring is quite different from that of a classroom, and so the set of skills and experiences required to be a successful tutor are plausibly different. Compared to regular classroom instruction, one-on-one (or two-on-one) tutoring greatly simplifies the instructional task, expanding the set of people who can be successful instructors. Teachers learn how to individualize instruction or handle classroom management usually through extensive pre-service pedagogical training as well as on-the-job learning over the first several years of classroom teaching (Rockoff, 2004; Clotfelter et al., 2010; Henry et al., 2011). But the importance of that prior training and on-the-job learning is plausibly less for tutoring.

This insight enables Saga to expand the applicant pool to anyone who both possesses strong math skills and are willing to devote one year to public service—for example, recent college graduates, retirees or career-switchers—but do not necessarily have extensive prior training or experience as teachers. A school in which classroom teachers cycle in and out after just one year would face significant challenges. As with other public service programs the tutors were willing to serve at relatively low wages ($16,000 plus benefits for the nine-month academic year during the study period, $20,000 plus benefits today). As with Banerjee et al. (2007), Saga substitutes a very different teaching method for many dimensions of what previous studies call "teacher skill" or "quality" (such as teaching experience or extensive pedagogical training). This substitution makes high-dosage tutoring more feasible from a cost perspective.

The tutors were mostly recent college graduates hired because they had both strong math skills (according to Saga's screening assessment) and strong interpersonal skills (as revealed by interviews that also involved delivering a mock tutoring session). Hired tutors have higher SAT scores than what other studies have reported for big-city public school classroom teachers (Jacob et al., 2018).[15] But Saga tutors neither had formal teacher training nor were licensed Illinois teachers. The Saga intervention is similar in spirit to how charters often hire teachers with less experience and formal teaching credentials than normal public schools in exchange for smaller class sizes (Lake, Bowen, & Demeritt, 2012). Saga hired 139 total tutors across both study years out of an estimated pool of approximately 1,200 applicants. Roughly half of tutors hired were Black or Latinx, and approximately 50% were female.[16] We do not know how many tutors it would be possible to hire before effectiveness declines, but over this range there is little difference between tutors at the top vs. bottom of Saga's ranked hiring list (Davis et al., 2017).

For our study 1, described further below, Saga delivered this tutoring intervention in CPS starting in the 2013-14 academic year (AY). Each tutor participated in approximately 100 hours

[15] A comparison of Saga Tutors SAT scores from AY2016-17 and AY2017-18 in Chicago reveals that average math SAT scores were 707 compared with the average math SAT scores for new NYC teachers of 613 reported in Rockoff, Jacob, Kane, and Staiger (2008). We see similar patterns for English scores.

[16] Of the 54 Saga tutors hired for study 1 in AY2013-14, 18 were Black and 8 were Latinx. Nineteen tutors spoke fluent Spanish, and every school with a high proportion of Spanish speakers had multiple bilingual tutors. Of the 85 Saga tutors hired in AY2014-15, 23 were Black, 9 were Latinx and about a quarter spoke fluent Spanish. There was also a significant increase in the percentage of tutors with advanced degrees for the second year, with nearly 20% of tutors possessing an advanced degree, in comparison to 7% in AY2013-14.

of training prior to the start of the school year (full-time for four weeks during the summer). Students—as part of their regular class schedule—were assigned to participate in a tutoring session for one class period every day in addition to their regular math class. For study 2 participants, tutoring typically replaced an elective course such as art or physical education. For study 1 9th -graders—i.e., the majority of the study 1 sample—tutoring replaced a second hour of Algebra ("double-dose"). This comparison is important to keep in mind when interpreting results from study 1 versus study 2. The total Saga contact hours could be up to 140 per year. Each tutor worked with two students at a time during each session, focusing on Saga's Algebra curriculum, but teaching foundational mathematics skills where needed to access these Algebraic concepts. Study 1 provided students with up to two years of intervention, whereas study 2 ran for one year.

Each Saga class period in general followed a set routine. A student would first do four to five minutes of warm-up problems before receiving 40 minutes of tutoring on material tailored to that student. Finally, the student would complete one to three problems designed to assess understanding of the material covered during the class period. So, about half of each session focused on remediating skill deficits, for which Saga developed its own curriculum, and the other half was tied to what students were learning in their regular math classrooms.[17] Saga used frequent internal formative assessments of student progress to individualize instruction.[18] In addition, Saga site directors worked with mathematics teachers on a weekly basis to understand what standards were being taught in mainstream math classes so the Saga tutorial covered

[17] For the Fryer study, the tutoring curriculum was reverse-engineered from Texas state standards by Fryer's EdLabs. For this project, an Illinois-certified math teacher helped develop curricula based on Illinois state standards.

[18] These include one- to three-question mini-assessments of the day's lesson, which allows the tutor to revise the next day's lesson based on the prior day's learning; tests before and after each of the seven to 10 "course units" Saga has divided the year up into, which show tutors how much review time to allow for the first two-three weeks of the next unit); 80 item quarterly proficiency assessments, which tutors use to target areas to work on until the student scores at least 90%; and *site-specific norm-referenced tests* (which will involve interim assessments of the tests CPS uses).

complementary concepts. Saga tutors also discussed general study skills with students as part of the formal program (such as structuring how to approach a difficult problem by breaking it down), as well as through informal discussions. Tutors taught six periods a day, and each school was overseen by a site director who handled behavioral issues in the tutoring room and communication with school staff, and offered daily feedback and professional development.[19]

The control group in our study sample was eligible for all the status quo supports in the CPS high schools in our study. These services include No Child Left Behind (NCLB) funded supplemental educational services (SES) tutoring, which is of much lower dosage than Saga tutoring (and without the same structure, curriculum, or supervision). For example, for study 1 we estimate about 25% of control students in our schools received SES tutoring, which involves 21 hours of writing tutoring per *year* and 20 hours of math tutoring (i.e., a bit over one-half hour *per week* of math tutoring, compared to 45-50 minutes *per day* with Saga); previous non-experimental studies of SES tutoring in Chicago find little detectable effect on math scores.[20] While these schools include a variety of other programs (both treatment and control groups are eligible for these other programs) none of them focus on academic skills the way Saga does.[21]

Fryer (2014) examined the effects of Saga tutoring as part of his larger study introducing five elements of successful ‘no excuses’ charter school practices into public schools in Houston

[19] Each site director has some combination of relevant experience, including math teaching / tutoring, mentoring, program direction, nonprofit management, public speaking, and training of adults, and is trained in the Saga model. Tutors complete a daily report to the site director; here, they note each student’s progress and convey any issues.

[20] The most recent evaluation of SES in Chicago is for 2005-6 (http://sesiq2.wceruw.org/documents/chicago_ses.pdf). During 2005-6 school year, around 24 percent of all eligible Chicago Public School students (about 70% of students who applied for SES services) received SES services.

[21] Eight of the schools also include GEAR UP services, which focus on preparing for college (essay writing and ACT prep). Four schools include Youth Guidance’s Project Prepare, another college-readiness program (see https://www.youth-guidance.org/youth-workforce-development/#projectprepare), four of the schools provide Peace Circles to help youth resolve disputes, 11 of the schools include the Mikva Challenge, which focuses on civic education (http://www.mikvachallenge.org/), and one school includes buildOn, which focuses on involving youth in community service (https://www.buildon.org/our-work/buildon-us/).

and several other districts.[22] Saga tutoring did not appear to have large effects on test scores among 6th grade students beyond the effects of the other charter practices, but did appear to have more pronounced effects on test scores among 9th grade students, on the order of 0.32 SD, although high schools could not be randomly assigned.[23]

## III. DATA, RANDOMIZATION AND STUDY SAMPLES

### A. Data

One way we measure academic performance is from longitudinal student-level records maintained by CPS. They capture basic demographics, enrollment, attendance, grades in each course, and disciplinary actions. These data also include achievement test scores for the exams that CPS administered to 9th and 10th graders in our study years—the 9th grade EXPLORE and 10th grade PLAN tests, which are developed by ACT, Inc. We have these CPS data at baseline for all students in our two study samples. There is some missingness in post-randomization data (Appendix Table 1), which for school attendance equals 4.4% for controls and 6.2% for treatment, and somewhat higher for grades (14.6% vs. 14.8%) and test scores (29.8% vs. 29.5%), presumably in part because of a combination of students dropping out, transferring to suburban or private schools, and missing school on testing days. Attrition rates are similar for study 2. Treatment-control differences in missingness are not statistically significant.

[22] Four of the five charter school reforms—increased instructional time, replacing almost all principals and half of teachers, frequent formative assessments, and a culture of high expectations—were administered in all grades, while tutoring in math was given only to students in selected grades for cost reasons in the main Houston study.

[23] For older students a quasi-experimental design was used that exploited whether students were enrolled in a treatment school during the pre-treatment year or were zoned to be in that school. The contrast in math gains between grades that received tutoring (6th and 9th) versus did not (7th and 10th) was on the order of 0.09 to 0.40 SD. Separately, Fryer and Howard-Noveck (2020) study the effects of high-dosage reading tutoring for middle-school students and find statistically significant gains in attendance but not reading scores overall, although for Black students specifically reading scores increased by 0.09 SD.

From Saga we obtained tutoring attendance records, tutor characteristics, and student scores on Saga's own internal math assessments. To measure effects on criminal behavior, we used juvenile justice and adult criminal justice arrest data from the Chicago Police Department.

Our final data source comes from two waves of in-person surveys carried out for our research team by the Institute for Social Research (ISR) at the University of Michigan. We draw on existing survey questions that have been used in previous studies of youth, including the Moving to Opportunity survey.[24] The first wave of surveys was carried out for the study 1 sample mostly in May-June of study year 1, with some surveys in early fall of study year 2. We selected a sub-sample of 881 youth and surveyed 663 for effective response rates of 88.2% for treatment students and 90.6% for control students.[25] We carried out another wave of in-person surveys with 1,238 youth in the study 1 sample in the fall after the second intervention year (2014-15), with effective response rates of 90.1% for the treatment group and 89.1% for controls.

These surveys also included math achievement tests that we administered to help mitigate missingness in the CPS test data. These additional math tests were based mostly on items from the achievement tests generated for the US Department of Education's NELS:88 8th grade wave. We supplemented the 8th grade NELS:88 math questions with items from the 5th grade wave of the Early Childhood Longitudinal Sample math assessment to broaden the range of math topics the assessment covered to help address possible "floor effects."[26] We return to this point below.

### B. Sample selection and randomization

[24] See https://www.nber.org/mtopublic/

[25] ISR used two-phase sampling: after interviewing 70% of the survey sample frame they selected a random sub-sample for intensive follow-up. Our analyses employ sampling weights to account for this design.

[26] We had the Educational Testing Service (ETS), which designed the tests for the NELS:88, run three-parameter item response theory (IRT) models (Lord, 1980) on math test score results to allow us to create scale scores and also place the students in our study in the NELS metric so that we could compare to the NELS sample.

We invited 30 of the larger high schools in CPS to a briefing about the study. Of those 30, 12 schools, primarily located on the south and west sides of Chicago, signed up to participate in study 1. Over the summer, we used administrative data to identify male students expected to attend each school that fall. Because Saga Education's tutoring capacity varied by school and sometimes by grade, we carried out random assignment conditional on school-by-grade "randomization blocks" and varied the probability of assignment to the treatment condition across randomization blocks.[27] All analyses control for randomization-block fixed effects. We randomized a total of 2,633 students for study 1 starting the summer before the program year, with 2,103 of those randomized showing up in study schools that fall. This represents 86.4% of the 2,434 total 9th and 10th grade male students enrolled in our study schools.

In our study 1 schools we also independently randomized students in the sample to receive the meta-cognitive intervention studied in Heller et al. (2017), Youth Guidance's Becoming a Man (BAM) program. Because that randomization was independent of assignment to Saga tutoring, controlling or not controlling for BAM assignment has no impact on the estimates of Saga effects reported here. Those BAM results are reported in Bhatt et al. (2021).

For study 2, we randomized a sample of 2,645 students in 15 schools (12 of which were also in study 1).[28] Study 2 students included a new cohort of incoming 9th grade male students in 14 schools, a cohort of 9th and 10th grade male students in one school, and a cohort of 9th and 10th grade girls in seven schools.[29] Of those randomized starting in the summer before the program

[27] Most randomization blocks had treatment-assignment probability between 0.45 and 0.60 (range 0.25 to 0.73).

[28] For study 2 we randomized a total of 2,710 students. However, due to student mobility and other factors N=65 students were randomized into study 2 twice, resulting in the 2,645 "unique" students randomized noted here. See Table 1 for more details on this.

[29] The range of treatment-assignment probabilities across blocks in study 2 was 0.44 to 0.80.

year, 1,823 (69%) wound up attending a study 2 school.[30] The number of randomized students who showed up at these study schools comprised 36% of the 5,068 total 9th and 10th grade students enrolled in study 2 schools.

### C. Sample Characteristics

Table 1 provides some context for our study sample. The average test score on the EXPLORE and PLAN tests among all CPS students is close to the national median. CPS administered some of these tests in the spring, but they are normed against U.S. students taking fall tests, so CPS students had more days of school before the test than the national sample used to generate the percentiles. In comparison, the average test scores of students in our study schools are 9 to 12 percentile points lower than the CPS average. Looking within the study schools, the specific students in the study have similar average test scores to the schoolwide averages of the schools they attend. The implication is our sample has scores below the CPS average, in contrast to many studies of ‘no excuses’ charters where applicants tend to have slightly better baseline scores than the host school system (e.g., Angrist et al., 2013).

Table 2 shows that the study 1 sample is split about evenly between Black and Latinx youth. Almost 90% are eligible for free or reduced lunch (FRL). The average GPA the year before our study was 2.11 on a 4-point scale. The study 2 sample is similar on these characteristics but included more Black and fewer Latinx students (and included female students). Randomization appears to have been successful. We carry out an F-test of the null hypothesis that baseline characteristics are jointly the same across treatment and control groups

[30] The study sample includes 629 girls who showed up at the study 2 schools (out of 799 randomly assigned), which equals 26.4% of all 9th and 10th grade female students in these schools, and 1,194 male students who showed up in study 2 schools (out of 1,847 randomly assigned), which accounts for 44.5% of the total 9th and 10th grade male students in our study schools.

by regressing a treatment-group indicator against all variables in Table 2, separately for studies 1 and 2, controlling for randomization blocks. The p-value for study 1 is p=0.832, and for study 2 is p=0.893.

## IV. ANALYSIS PLAN

Because of our randomized experimental design, our analysis plan is straightforward. We estimate both the intent-to-treat (ITT) effect and the effect of treatment-on-the-treated (TOT). The ITT estimate comes from estimating equation (1):

$$(1)\ Y_i = \pi_0 + \pi_1 Z_i + X_i \pi_2 + B_i + \varepsilon_i$$

where $Y_i$ is an outcome for student *i* measured after random assignment, $Z_i$ is an indicator for having been offered Saga tutoring, $B_i$ is a full set of randomization block fixed effects, $\varepsilon_i$ is a random error term, and $X_i$ is a set of baseline controls to improve precision.[31] To ensure the standard errors we calculate are not misleadingly small as an artifact of finite sampling issues (Young, 2019), we also report p-values from a non-parametric permutation test (Efron & Tibshirani, 1993). We randomly re-assigned the treatment indicator $P = 100{,}000$ times, storing the t-test statistic (T) in each replication, then calculating the share of replications where this exceeds the t-test statistic from using actual treatment assignment, $T^*$, or $\frac{1}{P}\sum_{i=1}^{P} I(|T| > |T^*|)$.[32]

While missingness of outcomes is balanced across treatment and control, as a sensitivity analysis we show results that use multiple imputation to fill in missing values. These methods assume outcomes are missing at random (MAR), i.e. values are unrelated to missingness

[31] These include test scores from the previous year and, in some models, also include age and grade fixed effects, free or reduced lunch status indicators, an indicator for having a learning disability, indicators for black and Latinx, and the following academic measures measured in the 2012-13 and 2013-14 baseline school year: GPA, days absent, days out-of-school suspension, days in-school suspension, and number of disciplinary incidents.

[32] For the permutation tests for the effects of treatment on the treated (TOT), described below, we randomly re-assign both the endogenous variable for actual treatment participation (D) and treatment assignment (Z).

conditional on observed covariates. We also estimate a quantile regression on median test scores and impute arbitrarily low scores (zeros) to students missing tests, given baseline data suggest those with missing tests are disproportionately students with low baseline test scores and grades.

To estimate the treatment-on-the-treated (TOT) effect we use random assignment ($Z_i$) as an instrumental variable (IV) for participation ($D_i$), as in equations (2) and (3) (Angrist, Imbens & Rubin, 1996; Bloom, 1984). The first-stage equation is:

$$(2)\; D_i \;=\; \gamma_0 + \;\gamma_1 Z_i + X_i\gamma_2 \;+\; B_i \;+\; \mu_i$$

where $D$ is an indicator for having participated in Saga tutoring (defined as having participated in at least one Saga tutoring session), the $\gamma$'s are parameters to be estimated, $\mu$ is a random error term, and all other variables are defined as above. The relationship of interest is:

$$(3)\; Y_i \;=\; \beta_0 + \;\beta_1 D_i + X_i\beta_2 \;+\; B_i \;+\; \vartheta_i$$

The identifying assumption here is that treatment assignment has no effect on the outcomes of those assigned to treatment who do not participate. Below we discuss what evidence we have about one potential threat to this, the stable unit treatment value assumption (SUTVA). The IV estimate for the parameter $\beta_1$ in equation (3) is essentially a ratio of two ITT estimates—the ITT effect on the outcome of interest in the numerator, and the ITT effect on participation in the denominator.

The final methodological issue has to do with statistical inference in the presence of multiple testing. We group our outcomes into four different "families" that we expect to be affected in a similar way by the intervention: (1) mathematics achievement; (2) achievement in other academic subjects; (3) school behavior; and (4) out-of-school behavior (arrests). We calculate the false discovery rate (FDR) q-value, which is the share of significant estimates within a family that are expected to be false positives (Benjamini and Hochberg 1995).[33]

## V. MAIN RESULTS

### A. Impacts from one year of intervention

The participation rate of study 1 youth in year 1 (defined as receipt of any Saga tutoring at all) was 40.2% for those assigned to treatment and 1.1% for controls, and for study 2 was 36.9% for the treatment group and 7.8% for controls. The most common reasons for non-participation among those assigned to programming were: (1) we randomized students over the summer and then they did not wind up attending the expected study school (this was true of 20.1% of the study 1 sample and 31.1% of the study 2 sample), or (2) the student had a scheduling conflict with a different required class and could not add Saga tutoring as a class in their schedule.[34] It was rare for students to either decline Saga or ask to be rescheduled out if it had been added to their schedule by default.

Table 3 shows the ITT effect in study 1 on math achievement test scores (EXPLORE and PLAN tests) equals 0.082 standard deviations (SD), with a TOT effect of 0.16 SD. One potential concern is that perhaps the tutors are just "teaching to the tests" rather than building broad knowledge. So it is notable that we see changes in math *grades* as well, with a TOT impact of 0.56 points on a 1-4 GPA scale, relative to the control complier mean (CCM) of 1.63; this represents a change of about a C- to a C+.[35] We also estimate a decline in percent of math courses failed of 49% of the CCM (-0.087 / 0.179). Other evidence that the treatment effects are

[33] The results are similar if we use the method from Benjamini, Krieger and Yekutieli (2006).

[34] This was most likely to be true for 10th graders, who tend to have less schedule flexibility than 9th graders, who could replace their second 9th grade algebra "double dose" section with tutoring in study 1.

[35] The College Board lists a 1.7 GPA as C-, 2.0 as C and 2.3 as C+. How to Convert Your GPA to a 4.0 Scale (collegeboard.org)

not the result of tutors teaching narrowly to the primary CPS accountability tests is that we find TOT effects of 0.19 SD on the math tests we had ISR administer on our behalf (and which teachers, tutors and students did not know would happen in advance). These impacts are statistically significant with respect to the p-value, calculated using either analytic standard errors (Table 3) or a permutation test (see appendix Table 2), as well as with respect to the FDR q-value that accounts for the number of tests in this family of outcomes (with the exception of the math test we gave, with q-value = 0.058).

The second panel of Table 3 shows that tutoring seems to have some positive spillovers on outcomes in other subject areas. Reading test scores do not show significant impacts but the TOT effect on grades in non-math courses was 0.20 points (relative to a CCM of 1.72) and percentage of courses failed in non-math classes are cut by 23% (-0.056/0.22). There do not seem to be any detectable spillovers to behavioral outcomes, as shown in the final two panels of Table 3. However, some of the estimated effects on arrests are large relative to the control means, though they are not statistically significant. We return to this below.

Table 4 suggests the learning gains experienced by students in study 1 are not statistical flukes, or the result of some unusually good program implementation that cannot be replicated, since the effects are at least as large in study 2. The TOT effects are a 0.37 SD increase on the EXPLORE and PLAN math tests, a 0.42 point increase in math grades (relative to a CCM of 1.79), and a 44 percent decline (-0.082/0.187) of math courses failed. There are no statistically significant indications of spillovers on non-math courses once we account for multiple testing. We see proportionately large changes in arrests, but they are not statistically significant once we account for the number of hypotheses we are testing within that family of outcomes.

Table 5 reports the results of pooling together the year 1 data from studies 1 and 2 to improve statistical power. In the pooled sample, the TOT estimate is a 0.26 SD increase in math test scores, a 0.52 point increase in math GPA relative to a CCM of 1.67, and a decline of 0.09 in percentage of math courses failed, equal to 47% of the CCM. Pooling the two studies is particularly valuable for detecting effects on outcomes that were not the primary target of the math tutoring intervention. However even with this added power, the proportionately large changes in measures like arrests and out-of-school suspensions are not quite significant.

### B. Effects from two years of intervention

Our study 1 cohort was able to participate in up to two years of the intervention, which raises the question of whether the gains from tutoring each year are cumulative. One challenge is that, for year 2, the experiment cleanly identifies the ITT effect, but cannot by itself tell us how much of the ITT effect at the end of year 2 is due to gains among treatment youth who received tutoring just in year 1 versus in both years 1 and 2. This challenge stems from the fact that we did not randomly assign dosage duration.

However, we can logically bound the possible effects. At one extreme, if there was *no persistence* in effects of year 1 participation on year 2 test scores, the year 2 test score ITT is due entirely to year 2 participation. In this case, the effect of two years of Saga can be estimated as a TOT where participation is defined as receiving Saga tutoring in year 2. This estimate is presented in the column labeled “Effect of Treatment in Year 2 on Treated (TOT)” in Table 6, and is equal to 0.84 SD. At the other extreme, if there is *no fade out* of effects of year 1 participation, the effect of two years of SAGA can be estimated by a TOT where participation is defined as receiving Saga tutoring in either year 1 or year 2, or both years. This estimate is shown in table 6 in the column labeled “Effect of Treatment in Year 1 and/or Year 2 on Treated

(TOT)" and is equal to 0.30 SD. The true effect of two years of Saga should be in between these two bounds. The results suggest the effects are at the least additive in years of participation, and perhaps more-than-additive as models of dynamic complementarities in learning would suggest.

Table 7 provides a different look at the effects of two years of intervention on risky behavior and crime victimization, using data from the second wave of surveys to study 1 youth given in the fall after year 2 of the program. We focus on the ITT, but the TOT can be calculated using the bounds described above. We can see in Panel A that there is a reduction in alcohol use that is statistically significant. We also see suggestive reductions in use of drugs (other than marijuana) and seriously hurting someone in a fight, though these are not statistically significant once we account for the number of hypotheses being tested here. There are no detectable effects on the other measures.

### C. Longer-Term Effects

Do the effects we measure fade out, persist, or grow over time? Table 8 pools together data from studies 1 and 2 (for improved power) and examines impacts for students as measured in what would be each student's 11th grade year if they were not retained in grade. Interpretation of these results could be complicated if there are treatment-control differences in grade retention, but we can rule out effects on this outcome of any larger than plus or minus 2 percentage points in ITT terms. We see TOT effects on math test scores in 11th grade of 0.22 SD, about the same size as the pooled impact measured at the end of one year of programming, and the TOT on math grades is 0.18 GPA points, about 35% of the year one effect.

Table 9 shows that the estimated TOT effect on graduating on time from high school, pooling the study 1 and 2 samples again, is positive 1.9 percentage points relative to a CCM of 77.2%, but this is imprecisely estimated. The standard error of 3.4 percentage points means we cannot rule out a decline in graduation rates as large as -4.8 percentage points or an increase as large as +8.6 points.[36] The point estimate for the effects of tutoring on graduation is close to the effect we might expect just from higher math test scores alone. This comes from multiplying the experimental impact on math test scores (0.26 SD from Table 5) by the coefficient of a non-experimental regression of graduating on time on 9th grade math test scores controlling for student characteristics. That exercise suggests higher test scores would boost graduation by 3.0 percentage points, well within the confidence interval around our estimated effect of the intervention on graduation directly.[37]

### D. Robustness checks and extensions

The estimated effects on math outcomes we measure are robust to a range of estimation decisions; the non-math GPA result (but not classes failed in non-math subjects) is somewhat more sensitive to these choices. Appendix Table 11 shows what happens when we change the set of baseline covariates we control for in our regression, while Appendix Tables 12 and 13 show the results if we drop from the analysis sample students we thought would be in our study schools during the summer months when we carried out random assignment but wound up not showing up at those schools in the fall. Appendix Tables 14 and 15 show the results using multiple imputation for missing outcomes and, for continuous outcomes, quantile regression where missing values are imputed arbitrarily low (in this case, zero) values.

The counterfactual condition in our study is easiest to describe for those students enrolled in 9th grade; for them, the most common alternative to Saga tutoring is the second period of

[36] Results for the study 1 and 2 cohorts separately are in Appendix Tables 8 and 9, while results for the pooled cohorts and the full set of graduation outcomes that can be calculated using the CPS data are in Appendix Table 10.

[37] That regression uses data on N=24,782 students in 9th grade in AY2013-14 for whom we have valid test scores and later graduation outcomes (82% of the total cohort of AY2013-14 9th graders). The coefficient on 9th grade math scores in that regression equals 0.116.

"double dose" algebra provided as part of the regular CPS curriculum in a standard classroom setting. In contrast, as noted above, for 10th graders the counterfactual treatment is whatever elective a student chose not to take. Appendix Table 16 shows results for 9th graders are similar to those from the pooled sample of 9th and 10th graders, consistent with the findings of Nomi and Allensworth (2009) suggesting the effects of CPS double dose algebra are limited. Appendix Table 17 replicates this analysis for the sample of 10th graders pooled from both studies.

The comparison of year 1 impacts of study 1 versus study 2 is complicated somewhat by the fact that study 2 includes female as well as male students. Appendix Table 18 shows that the results for female students are not so different from those of the full study 2 results that pool males and females together. (For completeness Appendix Table 19 presents results for males only, pooling together data from studies 1 and 2.)

The racial / ethnic composition of the study 1 sample is also fairly different from the study 2 sample; 46% of the study 1 sample are Black students, while in study 2, 64% of students are Black. We test the interaction of treatment with student race/ethnicity but do not see differential treatment effects for Black and Latinx students (Appendix Tables 21 and 22 present the pooled ITT and TOT results for this subsample of students, respectively).

Finally, the SUTVA assumption could be violated if tutoring has some spillover effects on control students, for example if control students see benefits of higher-achieving peers. If control students are most affected when exposed to relatively more higher-achieving peers, then our estimated effect of tutoring on student learning outcomes should be inversely related to the share of students assigned to treatment within each of our randomization blocks. Figure 1 plots these randomization-block-specific treatment assignment rates against block-specific TOT effects. We see that the treatment effect actually seems to *increase* (rather than decrease) with larger share of individuals within a block who are randomized to treatment. This is the opposite relationship we would expect if control group students were benefitting from positive spillovers, and so at least under this test we do not see evidence for violation of the SUTVA assumption.

## VI. MECHANISMS

In this section we explore patterns of impacts across schools, across students, and across specific items on the math achievement tests, in order to learn more about mechanisms.

### A. Is this an academic or non-academic intervention?

Given the positive effects this tutoring intervention seems to have, and given that most previous academic interventions for economically disadvantaged teenagers yield disappointing results while many non-academic ones yield encouraging results (for example Heller et al., 2017), one might wonder whether this is actually a *non-academic* intervention instead? That is, perhaps the very small student-to-instructor ratio make Saga tutoring effectively a non-academic, mentoring program instead. One finding already presented cuts against this interpretation. The academic subject matter of the tutoring was specifically math, and we find significant positive effects on math test scores, but no effects on reading test scores. An obvious interpretation of this pattern of results is that the academic subject matter of the tutoring sessions matters; tutoring students in math helps them to learn math. However, we find effects on non-math grades, suggesting there is at least some spillover of the learning or behavioral changes induced by tutoring that affects learning in other areas.

We investigate this hypothesis further by analyzing treatment effects on student responses to the first wave of survey questions we administered, at the end of year 1 for study 1. Panel A of Table 10 shows no statistically significant changes in the number of adults the student reports having available to talk to in their school, the number of adults they think care about them, or their willingness to talk to adults in the school. Panels B, C, and D show there are no statistically significant changes in measures of “grit”, conscientiousness, or (after multiple testing corrections) locus of control, respectively. These results are not different if we look at an index of outcomes in each family instead. One might wonder if tutoring affects learning by changing who the student spends time with. Panel E shows the only significant change in our collection of social-network measures is for Saga participants compared to controls to think it is *less* likely their friends think studying is important; whether this is due to an actual change in friends, or a new appreciation of what it means to take studying seriously, is not clear.

Because the wave 1 surveys were carried out mostly in May or June of the first year of intervention for the study 1 students, perhaps this is just too soon to see very large changes in some of these measures of candidate mechanisms of action. Table 11 shows results from wave 2 of our survey, which was carried out after the second year that cohort 1 received the intervention. Here again we see no detectable impacts on perceptions about supportive adults, grit, conscientiousness, locus of control, growth mindset, or social networks.

What we do see impacts on are measures of how connected, engaged or ambitious students are with respect to their studies, as in Table 12. Panel A shows, in the first wave of our survey, that participants are more likely than controls to say it is important to get good grades in math, and a marginally significant increase in the share who say they like math (FDR q-value = .085). Panel B shows this result does not persist until the fall after the 2nd year of the program.

Taken together, these results seem to suggest that the subject matter focus on math was important to generating the treatment effects on learning, and that tutoring seems unlikely to have boosted learning *solely* through a mentoring effect.

### B. Classroom management

A different hypothesis for how intensive tutoring improves student learning is by making it much easier to handle "classroom management" than it is in a typical classroom setting, so that students then experience fewer disruptions to learning and more time-on-task. In-person observations we conducted of the Saga program suggest students were spending approximately 90% of their time on math. This figure is high relative to other published figures for regular public school classrooms, but we could not carry out similar observations for the specific CPS classrooms that serve as the control condition in this study.

An alternative empirical test comes from examining across-school variation in impacts. Specifically, we would expect the benefits of fewer disruptions to instruction with Saga to be more pronounced in schools where the baseline levels of classroom disruptions are most pronounced, since that is where the gains from reduced disruption are largest. Table 13 measures the level of disruptions within a school as the number of disciplinary incidents per study student in a given school during the study year. This is taken from the CPS administrative school records and calculated for the first year of intervention separately for study 1 and study 2.[38] We find that the difference in impacts on academic outcomes are not statistically significant across high-disruption and low-disruption schools. All else equal, this would not seem to support the hypothesis that improved classroom management is a key driver of the overall Saga impacts.

These results do not seem sensitive to our reliance on CPS disciplinary actions as our measure of school behavioral climate, since we see qualitatively similar results when we use arrests of students in a school instead (see Appendix Table 20). In addition, in the first wave of

[38] Given that randomization block fixed-effects fully explain school fixed-effects, we cannot estimate the main effect for the 'above median per capita disciplinary incidents' dummy variable, so we do not report these estimates in the table. However, we can recover and report the interaction effect and see if impacts differ between these groups.

surveys we administered to students to measure mechanisms, we asked if “disruptions by others get in the way of my learning.” We see no statistically significant impact on this outcome.[39]

### C. Personalization

An alternative hypothesis for these effects is that high-dosage tutoring makes it easier for instructors to personalize instruction relative to the classroom-teaching condition. We would expect this benefit to be largest for those students whose academic level is furthest from the level to which teachers target classroom instruction. Unfortunately, we do not have any direct measure of the level at which CPS math teachers target classroom instruction. In a more stylized setting Bloom (1984) found that teachers may target instruction to the top third of the class distribution. If that were true for CPS math classrooms as well, we would expect to see (all else equal) students at the bottom of the achievement distribution benefit the most from high-dosage tutoring. But testing that hypothesis is complicated by the possibility of ‘floor effects’ with our math test score measures, because floor effects would cause the test results to understate any gains in math skills among those at the bottom of the achievement distribution.

To learn more about this hypothesis, we use machine learning techniques to estimate personalized treatment effects (PTE’s) for every individual in the sample. Intuitively, machine learning procedures use the data to identify the groupings defined by baseline covariates that are as similar as possible with respect to their estimated treatment effects. Rather than considering just a few interactions between covariates and treatment assignment, as would be the normal approach in economics, we can search over all covariates and even complicated functions of the covariates to model the structure of heterogeneity in effects across observations. To reduce the risk of fitting noise rather than true structure, these types of machine learning procedures focus

[39] The regression coefficient on the Z-score version of the outcome is -.057 (se=.089) for the intention-to-treat effect.

on maximizing out-of-sample fit. Whereas standard prediction tasks use machine learning to predict the average outcome for different subgroups defined by partitions of the baseline characteristics space, here we follow the recent literature in economics, statistics and computer science to use machine learning to predict treatment effects.

The specific procedure we use to estimate these PTE's for each student is Athey, Tibshirani and Wager's (2019) Generalized Random Forest (GRF) method. GRF is an adaptive locally weighted matching estimator that yields very flexible estimates of conditional average treatment effects. GRFs basically work in two steps. First, a causal forest is estimated, which can essentially be thought of as an average of many decision trees (Wager and Athey, 2018).[40] Rather than directly estimating treatment effects by averaging predicted treatment effects across the causal trees in the forest (as in Wager and Athey, 2018), GRF generates adaptive weights using the frequency with which observations occur in the same terminal nodes together. Because each tree's splits are determined by a criterion that (approximately) minimizes the expected mean squared error of treatment effect predictions (Athey and Imbens, 2016), observations in the same terminal node should have similar treatment effects. This step yields a separate set of adaptive weights for each observation indicating how similar treatment effects are likely to be between the observation and other observations. Second, PTEs for each observation, $\hat{\tau}_i$, are estimated using separate weighted least squares regressions of the outcome on treatment for each observation using the individualized similarity weights determined in the first step.

[40] Our causal forest includes 100,000 trees. For each outcome, we select the tuning parameters (the fraction of the sample used to build each tree, the fraction of the subsample used for determining splits rather than estimating effects, the number of variables tried at each split, the minimum node size, and a penalty for imbalance in observations across splits) by cross-validation on smaller causal forests with 1,000 trees. Note that GRFs determine splits using an approximation to the Athey and Imbens (2016) criterion, whereas Wager and Athey (2018) use the exact criterion. Both yield heuristic solutions to the optimization problem because splits are determined sequentially using a greedy algorithm rather than jointly (which is currently computationally infeasible).

Before using the PTEs to evaluate mechanisms, we assess how well the predictions correlate with the true underlying heterogeneity. To this end, we estimate the best linear predictor (BLP) of conditional average treatment effects based on the estimated PTEs using the following weighted least squares regression (Chernozhukov et al. 2019):

$$(4) Y_i = \alpha_0 + \alpha_1 \hat{Y}_{0,i} + \alpha_2 \hat{\tau}_i + \alpha_3\left(Z_i - \hat{Z}_i\right) + \alpha_4\left(Z_i - \hat{Z}_i\right)(\hat{\tau}_i - E[\hat{\tau}_i]) + \varepsilon_i,$$

where $\hat{Y}_{0,i}$ is a random forest prediction of the outcome using only control group observations, and $\hat{Z}_i$ is the probability of treatment for observation $i$ (i.e. the block mean of treatment for $i$'s block). The weights are given by $w_i = \left(\hat{Z}_i\left(1 - \hat{Z}_i\right)\right)^{-1}$. If the GRF procedure is successfully predicting variation in treatment effects, we would expect the estimate of $\hat{\alpha}_4$ to be close to 1. In contrast, if the estimated PTEs are essentially noise, the estimate of $\hat{\alpha}_4$ would be close to 0.

The BLP estimates, shown in appendix figure 1, suggest that we may be picking up structure rather than noise with heterogenous treatment effects for our math outcomes, because the $\hat{\alpha}_4$ estimates are close to one. But even with our large sample (relatively speaking from the perspective of social science experiments), we are barely powered if not slightly under-powered even to detect heterogenous treatment effects on math outcomes. We do not focus on non-math outcomes in what follows given the poor fit of our heterogeneous treatment effect models. The fact that our predictions are only correlated with underlying heterogeneity for some outcomes, but not others, highlights the importance of first doing these kinds of performance checks before using the predictions for other purposes.

Figure 2 plots PTEs for math achievement test scores.[41] We can see that there are enormous differences in the estimated average effects for the quartile of students who benefit the

[41] We also recreate this plot for our predicted treatment effects on Math GPA in appendix figure 2.

most (average PTE = 0.21 SD in ITT terms) versus the smallest-effect quartile (average PTE = 0.007 SD). But it's also the case that the *baseline* math scores are strikingly different across these quartiles as well, with those in the top quartile having scores 0.61 SD above our study sample average at baseline and those in the bottom quartile 1.2 SD below average at baseline. This highlights the difficulty of interpreting the low estimated PTEs for this bottom quartile. If Reardon's (2011) estimate is right that high school students gain 0.2 SD on average for each year in school, the bottom quartile is about 6 years behind the average student in our study. This is a group for whom the risks of floor effects with these tests would be most pronounced.

Figure 3 provides us with a different way to see this; for each student in our out-of-sample validation set, we plot their estimated PTE on math test scores on the x-axis and their estimated PTE on math GPA on the y-axis. For data visualization purposes we also do K-means clustering and then show students in each 'cluster' in a separate color, and outline the outer boundary of students in each group. This gives us another way to see that a sizable share of students are estimated to have PTE's that are close to zero or even negative with respect to math test scores. On the other hand, almost all students have positive estimated PTE's on math grades.

For present purposes the most relevant feature of Figure 3 is the fact that around one-quarter of our study sample has estimated math test score gains in the neighborhood of zero or negative, but positive (and sometimes substantially positive) gains in math GPA. In principle this could be due to some gain in school effort or study skills from Saga, if math GPA is more sensitive to such changes than is math test scores. But in the CPS administrative data we do not see those students showing particularly large gains (or any gains for that matter) in measures of effort such as school attendance or school disciplinary infractions. Nor in the survey data do we see any changes in measures like time spent on homework, or share homework completed. The figure is at the very least consistent with, if not definitive proof for, floor effects with math tests.

Note that unfortunately math grades can't help us determine which students benefit the most, even though we can see almost all students derive some non-zero math grade benefit. The nice feature of standardized test scores for measuring treatment heterogeneity is that (setting aside floor effects) they are continuous measures designed to capture the normed, latent distribution of student ability in a given student population in addition to being tied to standards that quantify mastery on subject-specific skills and content. The standardized test scores used by CPS—the EXPLORE / PLAN tests developed by ACT, Inc. —are both norm-referenced tests and criterion-referenced tests, meaning that we are able to compare students' position in achievement compared to their peers, as well as (aspirationally) consistently quantify the amount of learning that occurs from one point in the latent distribution to another throughout this distribution. In contrast, there is no guarantee that a given unit change in grades need mean remotely the same thing at different points in the GPA distribution.

Perhaps our best test then of the personalization-of-instruction hypothesis comes from exploiting the fact that we have access to item-level test score data from the math achievement tests that we administered to study participants ourselves. We first calculate the difficulty of each item (j) on the math test by using the testing summary statistic 'P+ Score', calculated as the share of students in our sample who answer the item correctly (Lord, 1980). We divide the sample into whether students had baseline CPS test scores above versus below study median. Above-median students score higher than below-median students on every item, and in both groups, the share who answer a question correctly increases with the P-score (see appendix figure 3).

Figure 4 provides suggestive visual evidence that reduced ‘academic mismatch’ could play a role. On the x-axis we rank order the individual test-score items by their P-score, and then plot the ITT effect for each item for those students whose baseline CPS math test score was above vs. below the median value for our sample. We then fit a kernel-weighted local polynomial to these individual test score items for both groups of students. We test the null hypothesis that there is no difference in the relationship between ITT effects and question difficulty between students with above and below median test scores using a permutation test (Canay, Romano, and Shaikh 2017). Let A be a vector indicating whether students had above or below median test scores and let X denote the remaining data. $T(X, A)$ is the average absolute difference in ITTs between the two groups across all questions. We estimate the p-value for the null hypothesis that there is no difference between the two groups using:

$$(5)\ \hat{p} = \frac{1}{10{,}000} \sum_{i=1}^{10{,}000} 1\{T(X, g_i A) \geq T(X, A)\},$$

where $\{g_i\}_{i=1}^{10{,}000}$ are 10,000 randomly selected within randomization block permutations on the above/below median classifications (A). While this test is a bit under-powered and we cannot reject the null hypothesis that the effect is the same for most items (p=.53), visually it looks like below-baseline-median students seem to benefit the most on items in the bottom or middle of the difficulty distribution while the above-median may benefit the most on more difficult items.

Recognizing that the confidence intervals are somewhat large around both of our kernel-weighted polynomials, what would the difference in shapes of the two curves imply if we took the results at face value? The fact that below-median-at-baseline students seem to benefit more on relatively easier items from Saga tutoring and above-median students may benefit more on harder items would be consistent with the extra personalization built into the Saga program design, where tutors are instructed to devote some time each session to working with students where they are academically as determined by frequent formative testing. Unfortunately, we cannot document this mechanism directly because we do not have direct measures of the academic level at which tutors devoted their time, nor do we have that for classroom teaching. But at the very least the differential shapes of the below- and above-median curves in Figure 4 suggest student learning is concentrated on the math topics close to their baseline levels of knowledge. The implication is that even if Saga were not personalizing, the item-level results imply that personalizing (concentrating instructional time on where students are academically) would be an effective strategy.[42]

## VII. CONCLUSION

One way to read the results reported here is as a test of a specific intervention strategy for urban public school systems. Fryer (2014) shows that identifying a handful of strategies from ‘no excuses’ charter schools and incorporating them into public schools can improve student achievement. But many of those strategies are complicated to implement with fidelity and may be politically difficult to implement at large scale throughout urban public school systems. The present paper shows that substantial progress can be made by narrowing down even further the set of strategies adopted from ‘no excuses’ charter schools to just a single element that is easier to implement and hence scale: high-dosage tutoring. A key innovation of the specific high-dosage tutoring model we study here is that the provider, Saga, has figured out a way to deliver

[42] That is, if we took the results in Figure 4 at face value (ignoring the low power of the test that the shapes of the two curves are the same) they would imply either that (a) below-baseline-median students were showing their biggest gains on relatively easier items (and vice versa for above-baseline-median students) because Saga tutors concentrated their instructional effort on items tailored to where students are, or (b) Saga tutors tried to cover material across the difficulty distribution, and students just benefited most for the content closest to the student’s own academic level, which implies that efforts to further concentrate instruction on that content (i.e. personalization) would yield even bigger benefits to students.

this at relatively low cost given the intensity of the educational intervention. This stands in contrast with how tutoring services are often provided to school systems like Chicago's, where a tutoring provider hires (say) former teachers at prevailing full-time teacher wages, which then for a given level of spending per pupil means students get just a fraction of the number of instructional hours as they do with the intervention we study here.

One way to think about the viability of this strategy at scale is that if personalization is indeed a key mechanism through which Saga works, then the system goal for tutoring would be to deliver each student enough 'dosage' to eventually get them up to grade level and more productively engage with grade-level classroom instruction. That is, getting every student up to speed improves the degree to which regular classrooms are 'personalized.' Our best estimate for the cost per participant during our study period is roughly $3,800, with a defensible range of $3,500 to $4,300. CPS already receives around $300 million in Title 1 grants each year, and existing research suggests for the average district how these funds are deployed leads to few detectable improvements in student learning (Deke, et al., 2012, Dragoset, et al., 2017). As a thought exercise, that funding, if repurposed to Saga tutoring, would be enough to serve around 80,000 students per year, or 23% of the system's total enrollment of 355,000.[43] Proposals for how to scale nationally are in Ander, Guryan and Ludwig (2016) and Kraft and Falken (2021).

But there is another, perhaps more important way to read this study: As a demonstration that it is not too late or too difficult to substantially change the academic outcomes of children who are struggling academically even once they have reached adolescence. The lesson may be that it is possible to improve academic skills by accounting for the challenges of individualizing

[43] This calculation is based on Saga's program cost during our study period ($3,800).

instruction—among other things—and that these strategies can be effective even when implemented in regular public high schools to broad, representative samples of students.

The benefit-cost ratio of Saga tutoring seems to be comparable in magnitude to that of the most successful early childhood programs, like the Abecedarian Project and the Perry Preschool Program, as well as interventions at scale to improve student outcomes such as the Tennessee STAR experiment.[44] Chetty et al. (2011) show that each one-percentile increase in 8th grade test scores is associated with $150 in additional annual earnings at age 27. We find that, on average, Saga increases student's test scores by six percentile points in study 1 and by 14 percentile points for study 2.[45] Together, these two findings suggest Saga increases participants' adult earnings by $900 each year for study 1 and $2,100 each year for study 2. Assuming these gains persist from ages 25 to 59 and discounting at a 5% rate to age 15, the present discounted value of the gains is about $9,000 for study 1 and $21,000 for study 2. With per-student costs of $3,500 to $4,300, this implies the benefit cost ratio is 2.1 to 2.6 in study 1 and 4.9 to 6.0 in study 2.[46]

By way of comparison the benefit-cost ratios that have been estimated for model early childhood programs are on the order of 1.9 and 2.2 for the Abecedarian Project (Masse and Barnett, 2002), and between 3.9 and 6.8 for the Perry Preschool Program (Heckman et al., 2010), both estimated using a 5% discount rate. Krueger (2003) estimates that a 7-student reduction in

[44] Borman and Hewes (2002) show that the Success for All model yields similar math test score gains per $1,000 as the model programs discussed here (a 0.04 SD improvement in math per $1,000). Saga also performs favorably based on this metric, with math TOT effects per $1,000 of 0.04 to 0.05 in Study 1 and 0.09 to 0.11 in Study 2. Success for All, however, also improves reading scores (a 0.09 SD improvement per $1,000).

[45] The test score percentile data is not available for study 2. Therefore, we estimate the percentile effect by assuming the ratio of the TOT effects in study 1 and 2 is the same standard deviations and percentiles.

[46] Alternatively, Hanushek and Woessman (2008) review several studies that consistently find a one standard deviation increase in test scores is associated with about a 12 percent increase in earnings. Applying this effect size combined with our estimated increase on standardized math test scores to a quadratic wage/salary earnings age trajectory estimated using data on Black and Latinx individuals from Chicago in the 2019 ACS (and discounted to age 15 at a 5% rate) implies slightly smaller benefit-cost ratios of 1.6 and 3.8, respectively.

class size in grades K-3 yields a benefit-cost ratio of about 2 using a 4% discount rate (and so would be below 2 with a 5% discount rate). These studies have long been cited as arguments for investments in early childhood compared to later life stages; evidently, however, adolescence is not too late to also realize large social benefits from human capital investment.

# REFRENCES

Abdulkadiroğlu, A., Angrist, J. D., Narita, Y., & Pathak, P. A. (2017). Research Design Meets Market Design: Using Centralized Assignment for Impact Evaluation. *Econometrica*, *85*(5), 1373–1432.

Allensworth & Easton. (2005). The On-Track Indicator as a Predictor of High School Graduation. *Consortium on Chicago School Research*. https://consortium.uchicago.edu/sites/default/files/2018-10/p78.pdf

Altonji, J. G., Elder, T. E., & Taber, C. R. (2005). Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools. *Journal of Political Economy*, *113*(1), 151–184.

Ander, R., Guryan, J., & Ludwig, J. (2016). *Improving Academic Outcomes for Disadvantaged Students: Scaling Up Individualized Tutorials*. 24.

Angrist, J.D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*(434), 444–455.

Angrist, J.D., Pathak, P. A., & Walters, C. R. (2013). Explaining charter school effectiveness. *American Economic Journal: Applied Economics*, *5*(4), 1–27.

Angrist, Joshua D., Cohodes, S. R., Dynarski, S. M., Pathak, P. A., & Walters, C. R. (2016). Stand and Deliver: Effects of Boston's Charter High Schools on College Preparation, Entry, and Choice. *Journal of Labor Economics*, *34*(2), 275–318. https://doi.org/10.1086/683665

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360. https://doi.org/10.1073/pnas.1510489113.

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *Annals of Statistics*, *47*(2), 1148–1178. https://doi.org/10.1214/18-AOS1709.

Banerjee, A. V., Cole, S., Duflo, E., & Linden, L. (2007). Remedying education: Evidence from two randomized experiments in India. *Quarterly Journal of Economics*, *122*(3), 1235–1264.

Barnett, W. S., & Masse, L. N. (2007). Comparative benefit–cost analysis of the Abecedarian program and its policy implications. *Economics of Education Review*, *26*(1), 113–125. https://doi.org/10.1016/j.econedurev.2005.10.007.

Belfield, C. R., Nores, M., Barnett, S., & Schweinhart, L. (2006). The High/Scope Perry Preschool Program: Cost-Benefit Analysis Using Data from the Age-40 Followup. *Journal of Human Resources*, *XLI*(1), 162–190. https://doi.org/10.3368/jhr.xli.1.162.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Benjamini, Y., Krieger, A. M., & Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, *93*(3), 491–507. https://doi.org/10.1093/biomet/93.3.49.

Bhatt, M., Guryan, J., Ludwig, J., Shah, A., & the Chicago Youth Violence Project Team. (2021). *Scope challenges to social impact* (Working Paper No. 28406). National Bureau of Economic Research. http://www.nber.org/papers/w28406

Blattman, C., Jamison, J. C., & Sheridan, M. (2017). Reducing Crime and Violence: Experimental Evidence from Cognitive Behavioral Therapy in Liberia. *American Economic Review*, *107*(4), 1165–1206. https://doi.org/10.1257/aer.20150503

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-on-one tutoring. *Educational Researcher*, *13*(6), 4–16.

Bloom, H. S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review*, *8*(2), 225–246.

Bonnie, R. J., & Backes, E. P. (Eds.). (2019). *The Promise of Adolescence: Realizing Opportunity for All Youth*. The National Academies Press. https://doi.org/10.17226/25388

Borman, G. D., & Hewes, G. M. (2002). The Long-Term Effects and Cost-Effectiveness of Success for All. *Educational Evaluation and Policy Analysis*, *24*(4), 243–266. https://doi.org/10.3102/01623737024004243

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2006). How changes in entry requirements alter the teacher workforce and affect student achievement. *Education Finance and Policy*, *1*, 176–216.

Campbell, F. A., Ramey, C. T., Pungello, E., Sparling, J., & Miller-Johnson, S. (2002). Early Childhood Education: Young Adult Outcomes From the Abecedarian Project. *Applied Developmental Science*, *6*(1), 42–57. https://doi.org/10.1207/s1532480xads0601_05.

Canay, I. A., Romano, J. P., & Shaikh, A. M. (2017). Randomization Tests Under an Approximate Symmetry Assumption. *Econometrica*, *85*(3), 1013–1030. https://doi.org/10.3982/ECTA13081

Carneiro, P., & Heckman, J. (2003). Human Capital Policy. In James J. Heckman & A. B. Krueger (Eds.), *Inequality in America: What Role for Human Capital Policies?* (pp. 77–240). MIT Press.

Cascio, E. U., & Staiger, D. O. (2012). *Knowledge, tests, and fadeout in educational interventions*. National Bureau of Economic Research.

Cheng, A., & Peterson, P. E. (2017). How satisfied are parents with their children's schools? New evidence from a U.S. department of education survey. *Education Next*, *17*(2), 20–28.

Chernozhukov, V., Demirer, M., Duflo, E., & Fernandez-Val, I. (2019). *Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments*.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star. *Quarterly Journal of Economics*, *126*(4), 1593–1660. https://doi.org/10.1093/qje/qjr041

Clotfelter, C., Ladd, H., & Vigdor, J. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, *26*, 673–682.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2010). Teacher credentials and student achievement in high school: A cross-subject analysis with student fixed effects. *Journal of Human Resources*, *45*, 655–681.

Davis, J. M. V., Guryan, J., Hallberg, K., & Ludwig, J. (2017). *The Economics of Scale-Up* (No. w23925). National Bureau of Economic Research. https://doi.org/10.3386/w23925

Deke, J., Dragoset, L., Bogen, K., & Gill, B. (2012). *Impacts of Title I Supplemental Educational Services on Student Achievement*. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Dobbie, W., & Fryer, R. G. (2019). Charter Schools and Labor Market Outcomes. *Journal of Labor Economics*, *38*(4), 915–957. https://doi.org/10.1086/706534

Dragoset, L., Thomas, J., Herrmann, M., Deke, J., James-Burdumy, S., Graczewski, C., Boyle, A., Upton, R., Tanenbaum, C., & Giffin, J. (2017). *School Improvement Grants: Implementation and Effectiveness (NCEE 2017- 4013*. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

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

Duncan, G., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K., & Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, *43*(6), 1428–1446.

Eccles, J. S., Midgley, C., Wigfield, A., Buchanan, C. M., Reuman, D., Flanagan, C., & Iver, D. M. (1993). Development during adolescence: The impact of stage-environment fit on young adolescents' experiences in schools and in families. *American Psychologist*, *48*(2), 90–101.

Efron, B., & Tibshirani, R. J. (1993). An introduction to the bootstrap. *Monographs on Statistics and Applied Probability*, *57*.

Engel, M., Claessens, A., & Finch, M. A. (2012). Teaching students what they already know? The (Mis)Alignment between mathematics instructional content and student knowledge in kindergarten. *Educational Evaluation and Policy Analysis*, *35*(2), 157–178.

Figlio, D. N., & Page, M. E. (2002). School choice and the distributional effects of ability tracking: Does separation increase inequality? *Journal of Urban Economics*, *51*(3), 497–514. https://doi.org/10.1006/juec.2001.2255.

Fryer, R. G. (2017). Chapter 2 - The Production of Human Capital in Developed Countries: Evidence From 196 Randomized Field Experimentsa. In Abhijit Vinayak Banerjee & E. Duflo (Eds.), *Handbook of Economic Field Experiments* (Vol. 2, pp. 95–322). North-Holland. https://doi.org/10.1016/bs.hefe.2016.08.006

Fryer, Roland G., & Howard-Noveck, M. (2020). High-Dosage Tutoring and Reading Achievement: Evidence from New York City. *Journal of Labor Economics*, *38*(2), 421–452. https://doi.org/10.1086/705882

Fryer, Roland G., Jr. (2014). Injecting Charter School Best Practices into Traditional Public Schools: Evidence from Field Experiments. *Quarterly Journal of Economics*, *129*(3), 1355–1407. https://doi.org/10.1093/qje/qju011

Gibbs, C., Ludwig, J., & Miller, D. L. (2013). Head Start origins and impacts. In M. J. Bailey & S. Danziger (Eds.), *Legacies of the War on Poverty* (pp. 39–65). Russell Sage Foundation Press.

Goldhaber, D. (2007). Everyone’s doing it, but what does teacher testing tell us about teacher effectiveness? *Journal of Human Resources*, *42*, 765–794.

Goldin, C., & Katz, L. F. (2008). *The Race between Education and Technology*. Harvard University Press.

Gormley, W. T., Phillips, D., & Gayer, T. (2008). Preschool programs can boost school readiness. *Science*, *320*(5884), 1723–1724.

Grogger, J., Neal, D., Hanushek, E., & Schwab, R. (2000). Further Evidence on the Effects of Catholic Secondary Schooling [with Comments. In *Brookings-Wharton Papers on Urban Affairs* (pp. 151–201).

Hanushek, E.A., Peterson, P. E., Talpey, L. M., & Woessmann, L. (2020). *Long-Run Trends in the U.S. SES-Achievement Gap*. National Bureau of Economic Research.

Hanushek, Eric A., & Woessmann, L. (2008). The Role of Cognitive Skills in Economic Development. *Journal of Economic Literature*, *46*(3), 607–668. https://doi.org/10.1257/jel.46.3.607

Heckman, J.J. (2012). *Giving Kids a Fair Chance*. MIT Press.

Heckman, J.J., Moon, S. H., Pinto, R., Savelyev, P. A., & Yaviz, A. (2010). The rate of return to the HighScope Perry Preschool Program. *Journal of Public Economics*, *94*, 114–128.

Heller, S. B., Shah, A. K., Guryan, J., Ludwig, J., Mullainathan, S., & Pollack, H. A. (2017). Thinking, Fast and Slow? Some Field Experiments to Reduce Crime and Dropout in Chicago. *Quarterly Journal of Economics*, *132*(1), 1–54.

Henry, G. T., Bastian, K. C., & Fortner, C. K. (2011). Stayers and leavers: Early-career teacher effectiveness and attrition. *Educational Researcher*, *40*(6), 271–280.

Hunt, D. E. (1975). Person-environment interaction: A challenge found wanting before it was tried. *Review of Educational Research*, *45*(2), 209–230.

Jackson, C. K. (2018). *Does School Spending Matter? The New Literature on an Old Question* (No. w25368). National Bureau of Economic Research. https://doi.org/10.3386/w25368

Jacob, B. A., Rockoff, J. E., Taylor, E. S., Lindy, B., & Rosen, R. (2018). Teacher Applicant Hiring and Teacher Performance: Evidence from DC Public Schools. *Journal of Public Economics, 166*, 81–97. https://doi.org/10.3386/w22054.

Kane, T., Rockoff, J., & Staiger, D. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review*, *27*, 615–631.

Kraft, M. A., & Falken, G. (2021). A Blueprint for Scaling Tutoring Across Public Schools. In *EdWorkingPapers.com* (No. 20–335). Annenberg Institute at Brown University. https://www.edworkingpapers.com/ai20-335

Krueger, A. B. (2003). Inequality, too much of a good thing. In James J. Heckman & A. B. Krueger (Eds.), *Inequality in America: What Role for Human Capital Policies?* (pp. 1–76). MIT Press.

Lake, R., Bowen, M., Demeritt, A., McCullough, M., Haimson, J., & Gill, B. (2012). Learning from Charter School Management Organizations: Strategies for Student Behavior and Teacher Coaching. *Mathematica Policy Research*.

Long, D., Mallar, C., & Thornton, C. V. D. (1981). Evaluating the Benefits and Costs of the Job Corps. *Journal of Policy Analysis and Management*, *1*(1), 55–76. https://doi.org/10.2307/3324110.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

Ludwig, J., & Phillips, D. A. (2007). *The benefits and costs of Head Start*. National Bureau of Economic Research.

Masse, L., & Barnett, W. (2002). *A Benefit Cost Analysis of the Abecedarian Early Childhood Intervention*.

Nomi, T., & Allensworth, E. (2009). Double-dose Algebra as an alternative strategy to remediation: Effects on students' outcomes. *Journal of Research on Educational Effectiveness*, *2*(2), 111–148.

Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. In E. G. J. Duncan & R. J. Murnane (Eds.), *Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances* (pp. 91–116). Russell Sage Foundation Press.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *2*, 417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, *94*(2), 247–252.

Rouse, C., & Barrow, L. (2006). U.S. Elementary and Secondary Schools: Equalizing Opportunity or Replicating the Status Quo? *The Future of Children*, *16*(2), 99–123.

Schweinhart, L. J., Montie, J., Xiang, Z., Barnett, W. S., Belfield, C. R., & Nores, M. (2005). *Lifetime Effects: The High/Scope Perry Preschool Study Through Age 40*. High/Scope Press.

Shonkoff, J. P., & Phillips, D. A. (2000). From Neurons to Neighborhoods: The Science of Early Childhood Development. *Zero to Three*, *5*. https://doi.org/10.17226/9824.

Steinberg, L. (2014). *Age of Opportunity: Lessons from the New Science of Adolescence*. Houghton Mifflin Harcourt.

Tuttle, C. C., Gleason, P., Knechtel, V., Nichols-Barrer, I., Booker, K., Chojnacki, G., Coen, T., & Goble, L. (2015). Understanding the Effect of KIPP as It Scales. In *Impacts on Achievement and Other Outcomes. Final Report of KIPP's "Investing in Innovation Grant Evaluation." In Mathematica Policy Research, Inc. Mathematica Policy Research, Inc: Vol. I*. https://eric.ed.gov/?id=ED560079

Wager, S., & S, A. (2018). Estimation and Inference of Heterogenous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242.

Young, A. (2019). Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results. *Quarterly Journal of Economics*, *134*(2), 557–598. https://doi.org/10.1093/qje/qjy029.

## Technical Appendix

### A. Building Grade Variables

Students participating in the intervention were assigned to participate in a tutoring session for one class period every day in addition to their regular math class. It is important to note that while the control condition for study 2 participants was overwhelmingly an elective course, the control condition for study 1 9th -graders—i.e., the majority of the study 1 sample—was a second hour or “double-dose” Algebra.

In building grade variables (overall GPA, non-math core GPA, math GPA, and course failures) used as covariates and outcomes, we want to ensure that we have comparable schedules between treatment and control students, taking the above into account. We use the following criteria to ensure we have comparable course schedules between treatment and control groups. We present our preferred specifications in the main results, and also include other variations in the appendices as robustness checks. Our results hold across different methods of variable calculation.

CPS defines ‘core’ courses as courses in core subjects: math, social science/history, science, and reading/English. Our baseline covariates include core GPA (and number of A’s, B’s, C’s, D’s, and F’s in core courses). We look at GPA and course failures (percent of courses failed) separately for math and non-math core subjects as our outcomes. CPS grades are administered each semester, and we use courses from both semesters of the intervention year to calculate our yearly grade variables.

#### ***Math Courses***

We first exclude ‘Math Lab’ (the Saga tutoring class) as well as double-dose Algebra (the class which Saga takes the place of) when calculating math course variables. A vast majority (over 97%) of study students for whom we have grade data only take one math class each semester after excluding Math Lab and double-dose Algebra.

For students who take multiple math classes in a semester, our preferred method is to select the appropriate grade-level math grade for each student as their main ’math course’—e.g. for a 9th grader, if they are in only one math course that semester, we take that course as their semester math class; if they are in multiple math courses that semester and take the grade-level appropriate math course (e.g. Algebra I for 9th graders, and Geometry for 10th graders), we use that course for variable calculation. If they are in multiple math classes in that semester and take a math course above their grade level (such as a 9th grader taking Geometry or a 10th grader taking Algebra II with Trigonometry), then we use that course as their ‘main’ math course. Otherwise, we randomly select a course and use that as their ‘math course’ for that semester.

The method where we ‘select’ the appropriate math course is presented in our main results and is used throughout our analysis. We also present a version of the results where we exclude students who we have to make a ‘selection’ for (presented in appendix table 23 and denoted with a ‘no selections’ marker).

### ***Non-Math Core Courses***

For non-math core courses (that is, courses in science, social studies, and reading/English), selecting the appropriate ‘grade-level’ course in each is less straightforward, and there is also possibility that including Saga in students’ schedules affects which other ‘core’ courses they can take. Thus, we present five methods to calculate non-math core grade variables. Our results calculated with each of these methods can be found in appendix table 23.

- Method 1: ‘All Classes’ – Include all courses taken by a given student in non-math core classes in the calculation of GPA and course failures. This is our preferred method.
- Method 2: ‘High grade by subject’—Selects the highest grade for each non-math core topic in each semester and uses those courses to calculate the outcomes. For example, if a student takes two science classes and two English classes in a semester, we take their highest science course grade and highest English course grade (as well as their one social science grade) and use those to calculate GPA and course failures.
- Method 3: ‘Low grade by subject’—Same as the above, but instead of selecting the highest grade for each core subject area in a semester, we use the lowest grade in each core subject area. This method in conjunction with the above method can help us bound the effects of the intervention on non-math core grades.
- Method 4: ‘Top 3 classes each semester’ uses a student’s three highest non-math grades in core subject areas in a semester regardless of subject.
- Method 5: ‘Top 6 classes in that year’ uses a student’s six highest non-math grades in that school year, regardless of subject.