American Economic Review 2023, 113(3): 738–765
https://doi.org/10.1257/aer.20210434

Not Too Late: Improving Academic Outcomes among Adolescents[†]

By Jonathan Guryan, Jens Ludwig, Monica P. Bhatt, Philip J. Cook, Jonathan M. V. Davis, Kenneth Dodge, George Farkas, Roland G. Fryer Jr., Susan Mayer, Harold Pollack, Laurence Steinberg, and Greg Stoddard*

Improving academic outcomes for economically disadvantaged students has proven challenging, particularly for children at older ages. We present two large-scale randomized controlled trials of a high-dosage tutoring program delivered to secondary school students in Chicago. One innovation is to use paraprofessional tutors to hold down cost, thereby increasing scalability. Participating in math tutoring increases math test scores by 0.18 to 0.40 standard deviations and increases math and nonmath course grades. These effects

* Guryan: School of Education and Social Policy, Northwestern University (email: j-guryan@northwestern.edu); Ludwig: Harris School of Public Policy, University of Chicago (email: jludwig@uchicago.edu); Bhatt: Education Lab, University of Chicago (email: mbhatt@uchicago.edu); Cook: Sanford School of Public Policy, Duke University (email: pcook@duke.edu); Davis: Department of Economics, University of Oregon (email: jdavis5@uoregon.edu); Dodge: Sanford School of Public Policy, Duke University (email: dodge@duke.edu); Farkas: School of Education, University of California, Irvine (email: gfarkas@uci.edu); Fryer: Department of Economics, Harvard University (email: rfryer@fas.harvard.edu); Mayer: Harris School of Public Policy, University of Chicago (email: s-mayer@uchicago.edu); Pollack: Crown Family School of Social Work, Policy, and Practice, University of Chicago (email: haroldp@uchicago.edu); Steinberg: Department of Psychology, Temple University (email: laurence.steinberg@temple.edu); Stoddard: Crime Lab, University of Chicago (email: gstoddard@uchicago.edu).
Chinhui Juhn was the coeditor for this article. 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, J-PAL North America, the Illinois Criminal Justice Information Authority, the city of Chicago, grant 2012-JU-FX-0019 from the Office of Juvenile Justice and Delinquency Prevention, Office of Justice Programs, US Department of Justice, and award 1P01HD076816 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health. We 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, we 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, Carolyn Silverman, Weiyu Wang, Robert Webber, Nathan Weil, and David Welgus for their invaluable contributions to the data analysis. For useful suggestions we thank seminar participants at Harvard, Microsoft Research, New York University, UCLA, and the Institute for Research on Poverty at the University of Wisconsin, as well as the editor and referees. This study was approved by the University of Chicago's committee on human subjects as IRB13-0691 on June 24, 2013. This RCT was registered in the American Economic Association Registry for randomized control trials under trial AEARCTR-0000041. All opinions (and any errors) 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.

[†] Go to https://doi.org/10.1257/aer.20210434 to visit the article page for additional materials and author disclosure statements.

persist into future years. The data are consistent with increased personalization of instruction as a mechanism. The benefit-cost ratio is comparable to many successful early childhood programs. (*JEL* H75, I21, I24, I26, I32, J13, J15)

Improving academic outcomes for economically disadvantaged students has proven a difficult challenge, particularly for older children. Studies of educational interventions for adolescents tend to yield more disappointing results than interventions for young children. This pattern gives rise to concerns that efforts to improve learning outcomes for teens may face intrinsic challenges, such as declining developmental plasticity (Carneiro and Heckman 2003; Knudsen et al. 2006; Heckman 2013).[1] More hopeful is Fryer's (2014) study showing that key components of "no excuses" charter schools—including a longer school day and year, replacing principals and teachers, and a culture of high expectations—improve outcomes for students of *all* ages. Yet the difficulty of implementing these changes raises questions of whether they can be scaled (Cullen et al. 2013).

There is one instructional technology that has been viewed for centuries as promising and replicable for students of any age: intensive or "high-dosage" tutoring. This method of instruction dates back at least to the fifteenth century at Oxford. High-dosage tutoring can be thought of in some sense as an extreme version of class-size reduction and has itself become a component of many no excuses charter schools.[2] Tutoring addresses what teachers report as the two most difficult challenges of classroom teaching:[3] classroom management and variability in the academic levels, and hence needs, across students. These challenges are more difficult with older students because as students age, disruptive behaviors become more prevalent,[4] and the variability in student academic levels (hence instructional needs) also becomes more pronounced (Cascio and Staiger 2012). This variation in instructional needs is difficult to address in classroom settings that focus on teaching grade-level curricula. Small-scale randomized controlled trials (RCTs) comparing tutoring to classroom instruction confirm tutoring to be "the best learning conditions we can devise" (B. Bloom 1984; Nickow, Oreopoulos, and Quan 2020). The challenge to widespread

[1] Part of the argument here is conceptual, based on economic models suggesting that skills developed early in life improve the productivity of later skill investments (see for example Cunha et al. 2006). Empirical support comes partly from studies of nonhuman animals finding "sensitive periods" in which skills are much easier to modify than in later periods of life, as well as some examples of the same phenomenon among humans as well, such as language acquisition (Cunha et al. 2006; Knudsen et al. 2006). And part of the argument also comes from studies of educational data for children of different ages, including descriptive data suggesting that long-term trends in the income gap in test scores are less encouraging for older than younger children (see for example Hanushek et al. 2020 and Hashim et al. 2020, although also see Reardon 2011). Cunha et al. (2006) note that some selected interventions for teens, such as financial incentives for school performance or Catholic school enrollment, can affect behavioral outcomes like crime or dropout, but impacts on test scores are more modest, and argue that the returns for older teens and young adults from programs like public job training are very low.

[2] See for example Fryer (2014); Dobbie and Fryer (2015); Dobbie and Fryer (2020); Tuttle et al. (2015); Angrist et al. (2016); and Abdulkadiroğlu et al. (2017).

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

[4] Disciplinary actions in school increase with age (https://nces.ed.gov/programs/raceindicators/indicator_RDA.asp), as do absences (https://www2.ed.gov/datastory/chronicabsenteeism.html) and arrests, including for serious crimes.

implementation has not been a pedagogical problem so much as an economic one—cost. Is it simply too costly to provide every student with tutoring without compromising effectiveness?

One potential solution to this cost (and hence scaling) challenge is to rethink the human resource model we use for tutoring versus classroom teaching. Given the enormous skill required for the difficult task of classroom teaching, and the variability across teachers in this skill,[5] many policies have focused on improving teacher skill and effort through some combination of teacher recruitment, training, or performance-based pay (e.g., Rivkin, Hanushek, and Kain 2005; Gordon, Kane, and Staiger 2006). An alternative idea is that simplifying the teaching task itself can reduce the amount of skill instructors require to teach effectively. The Pratham NGO in India tried this approach by hiring paraprofessionals to hold down costs (local community women hired at one-tenth the price of regular teachers), providing them with two weeks of training, and then directing them to work on remedial skill development with 15–20 children grouped based on their academic needs. Ability grouping is one way to make teaching easier because less skill is required of the instructor to personalize instruction across students. Banerjee et al. (2007) show test score gains from this intervention of 0.28 standard deviations (SD) in the short term, which were larger still for the students furthest behind, with gains that faded to 0.10 SD a year later.

In this paper we report on an intervention that seeks to deliver Oxford-style high-dosage tutoring at relatively low cost by employing a Pratham-style human resources model. The key premise is that dramatically reducing class size to two students per instructor (tutor) simplifies the teaching task enough that tutors without extensive prior pedagogical training or on-the-job experience can be effective. Saga Education, the developer of the tutoring model we study, hires paraprofessionals to serve as tutors for one year as a public service at a modest stipend.[6] Given evidence of substantial on-the-job learning by classroom teachers (Rockoff 2004; Clotfelter, Ladd, and Vigdor 2010; Henry, Bastian, and Fortner 2011) a staffing model with such high turnover would simply be infeasible for regular classroom instruction. But if tutors can be effective in their first year the benefit would be to help substantially hold down costs. Saga's cost during this study was $3,200 to $4,800 per year per student to deliver students nearly an hour of tutoring a day in school every day at a 2:1 tutor-student ratio (costs are lower now).[7] By comparison, Chicago Public Schools (CPS) spending per pupil is about $17,000 per year.

We carried out two separate RCTs of high-dosage tutoring with disadvantaged high school students, the first large-scale major tests of this strategy with students in this age range (the vast majority of previous tutoring studies focus on very young children; see Nickow, Oreopoulos, and Quan 2020). The pooled sample size of the two RCTs is 5,343; previous tutoring studies typically examine fewer

[5] See, for example, Chetty, Friedman, and Rockoff (2014a, b); Gilraine, Gu, and McMillan (2020); Rothstein (2010); Rockoff (2004); Kane and Staiger (2008); and Jackson (2018).

[6] Saga was initially part of the Match charter school organization in Boston, then spun off to become SAGA Innovations to focus on tutoring nationwide, then changed the name to Saga Education in 2019.

[7] Saga has since dropped its cost to $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.

than 200 students. The cost of the intervention we study relative to its intensity increases the chance of passing a benefit-cost test. Because high cost per student can be a barrier to scale up, evaluating programs that might support students' learning at reasonable cost also has the potential for large social impact.[8]

In the first RCT ("study 1"), we randomly assigned 2,633 rising ninth or tenth graders in 12 CPS high schools in the summer of 2013 to Saga tutoring versus control. This sample consists almost entirely of low-income male high school students of color. We focus on math because failing core math classes is a driver of dropout (Allensworth and Easton 2005), because math is important for later earnings (Duncan et al. 2007), and because math skills may be more responsive to school-based interventions than are reading skills (e.g., Fryer 2014, 2017). After one year of the program, the intention-to-treat effect (ITT) on standardized math test scores is 0.09 SD, and the treatment-on-the-treated (TOT) effect is 0.18 SD. These gains in test scores do not appear to be the result of tutors narrowly teaching to the test. We also estimate TOT effects on grades in regular classroom math courses equal to 0.57 points (on a 0 to 4 point scale), and a decline in the percent of math course failures of 48 percent.

Motivated partly to see if these results would replicate, in the summer of 2014 we randomized a separate sample of 2,710 ninth and tenth graders ("study 2"). The TOT effect on math scores after one year is more than twice as large as in study 1 (0.40 SD). We also find sizable positive effects on math grades in study 2 similar to the findings for study 1. This sample for study 2 includes females as well, who seem to benefit as much as male students do.

When we look at eleventh grade outcomes, a year or two after tutoring, we find persistent gains in math test scores of 0.23 SD (pooling study 1 and 2) and math grades of 0.25 GPA points. The estimate for on-time graduation is small and positive, 1.3 percentage points, but imprecise and not statistically significant.

These results may understate the learning gains students experience because of "floor effects," the idea that achievement tests do not measure gains in knowledge below the level the test questions target. For math course grades, we find positive treatment effects for all four quartiles of baseline achievement. In contrast, for math test scores, we find positive treatment effects for the top three baseline achievement quartiles, but no significant effects for the bottom quartile. These are patterns we might expect from floor effects because the gains in knowledge by students in the bottom quartile of the baseline test score distribution are most likely to be in topics not covered by end-of-grade standardized tests. The possibility of floor effects also complicates tests of candidate mechanisms that compare effects across students by baseline achievement levels.

So why does high-dosage tutoring generate such large gains in student learning? To guide exploration of the mechanisms by which high-dosage tutoring might generate learning gains we develop a simple model in the spirit of Lazear (2001) that suggests tutoring impacts should be larger in settings where there is more disruption to classroom learning time either because of disruptive behaviors (which are easier to manage in two-on-one tutoring) or relatively greater heterogeneity in student

[8] See also Dietrichson et al. (2017); Fryer (2017); Baye et al. (2019); and Pellegrini et al. (2021).

achievement levels within the classroom (since instruction is easier to personalize in tutoring). We find some suggestive support for the personalization hypothesis in the available data. An alternative hypothesis is that because tutors can really get to know students, this seemingly academic intervention is actually improving learning by serving as a nonacademic intervention through a "mentoring effect" that improves metacognitive or other nonacademic skills. But in survey data we see no signs that tutoring participants have more or stronger adult relationships than control youth, nor do we find detectable treatment effects on grit, conscientiousness, locus of control or other such skills.

Extrapolating the estimated test score impacts to earnings gains implies that the benefit-cost ratio 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.

I. The Intervention

Saga Education's high-dosage tutoring program provided students 50 minutes of small-group in-person tutoring (two students per tutor) every school day. About half of each tutoring session focused on earlier-grade topics the student had not yet mastered, for which Saga developed its own curriculum, and the other half was tied to grade-level material taught in the regular math classrooms. Saga used frequent formative assessments of student progress to individualize instruction. Tutors also discussed general study skills as part of both the formal tutoring program (like how to approach a difficult problem by breaking it down) and through informal discussions.

The key to making this high-dosage tutoring scalable and feasible from a cost perspective was to rely on paraprofessionals rather than teachers to serve as tutors. As with other public service programs, tutors were willing to serve at low wages ($17,000 plus benefits for the academic year during the study period, $20,000 plus benefits today). 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 had higher average SAT scores than is typical among big-city public school classroom teachers (Jacob et al. 2018). But Saga tutors neither had formal teacher training nor were licensed Illinois teachers. Roughly half of tutors hired during this study were Black or Hispanic, and around 50 percent were female. Each tutor participated in approximately 100 hours of training in the summer before school started.

Students in the treatment group were assigned—as part of their regular class schedule—to a tutoring session for one class period every day, which was a credit-bearing course and was in addition to their regular math class. For study 1 ninth graders (most of the study 1 sample), tutoring most commonly replaced a second hour of "double dose" Algebra. For study 2, tutoring typically replaced an elective course like art or physical education. The control group was eligible for all the status quo supports in the CPS high schools in our study including No Child Left Behind-funded supplemental educational services (SES) tutoring, which was of much lower dosage than Saga tutoring (and without the same structure, curriculum, or supervision).[9] While schools in the study had other supplemental programs, none focused on academic skills the way Saga tutoring did. Additional details about the Saga tutoring model are in the online Appendix.

II. Results

A. *Samples*

For study 1 we recruited 12 of the larger high schools in the CPS system, primarily located on the south and west sides of Chicago. Using administrative data the summer before the 2013–2014 academic year (AY) started, we identified a total of 2,633 students we expected to enroll in the study schools that fall. We then randomized them to the treatment group, which was offered up to two years of Saga tutoring, or else to control. Of the 2,633 students we randomized, 2,103 enrolled in study schools at the beginning of the school year. Our study sample represented 86.4 percent of all ninth and tenth grade male students in the study schools.

In study 1, we also independently randomized students to a behavioral science-informed metacognitive intervention from the nonprofit Youth Guidance called Becoming a Man (BAM), which we had tested in previous RCTs as reported in Heller et al. (2017). The inclusion of BAM in a 2 × 2 factorial design for this RCT was the motivation for the limitation of study 1's sample to males. At the time of randomization, our study team's goal was to present the main BAM effects from this RCT in one paper (see Bhatt et al. 2021), the main tutoring effects in a separate paper (this one), and the interaction in a third paper that would attempt to measure any complementarity between academic and nonacademic skill building. In practice, what appeared to have been variable implementation of BAM in this sample led to an imprecise estimate of the tutoring-BAM interaction, as discussed below and in Bhatt et al. (2021).

In response to a preliminary analysis of the study 1 results, the city of Chicago began to support Saga tutoring in public schools in the 2014–2015 school year with public-sector funding. To validate the results from study 1, we worked with the city to randomize the offer of these additional tutoring slots in what we refer to as study 2. We randomized 2,645 students in 15 schools (12 of which were also in study 1) to treatment (one year of tutoring in this study) versus control.[10] Though study 2 took place in many of the same schools as study 1, the students in study 2 were a different set of students from those in study 1. Most of the students in study 2 were incoming ninth graders in the 2014–2015 school year, but the study included some tenth graders as well. Study 2 did not include independent randomization to BAM as study 1 had, alleviating the restriction of the study to male students. Thus, study 2 included both female and male students. For study 2, 68.8 percent of all ninth and tenth grade male students in our study 2 schools were randomized and included in our study sample, while 33.5 percent of all female ninth and tenth grade students in our study schools were randomized and included in our sample for study 2.

[9] For example, for study 1 we estimate about 25 percent of control students in our schools received SES tutoring, which involves 21 hours of writing tutoring and 20 hours of math tutoring per *year*. Previous nonexperimental studies of SES tutoring in Chicago find little effect on math scores (http://sesiq2.wceruw.org/documents/chicago_ses.pdf).

[10] Sixty-five students were missed in the de-duplication process and appear in the sample twice so the study 2 sample size is 2,710. All study 2 estimates are clustered by student to account for these duplicates.

B. *Data and Measures*

Our main source of information about study participants was from longitudinal student-level records maintained by CPS. These data captured basic demographics, enrollment, attendance, course grades, disciplinary actions, and standardized test scores. The standardized tests administered at the end of the two program years, which are primary outcomes for the study, are from the ninth grade EXPLORE and tenth grade PLAN tests, developed by ACT, Inc. We have CPS data at baseline for all students in both study samples (not every baseline item is available for every student, but we have some baseline data for all students). There was more missingness in postrandomization data, which can be seen in online Appendix Table 1. In study 1, school attendance was missing at a 4.4 percent rate for the control group and a 6.2 percent rate for the treatment group. The missingness rate was somewhat higher for course grades (about 16.2 percent for both groups for math grades and 14.6 percent for both groups for nonmath grades) and test scores (just under 30 percent for both groups and subject areas), presumably because of students dropping out, transferring to suburban or private schools, and missing school on testing days. Attrition rates were similar for study 2. Online Appendix Table 1 shows the vast majority of treatment-control differences in missingness were not statistically significant. The only outcome with a significant difference in missingness at the 5 percent level is for school attendance in study 1. Below we show our results are not sensitive to different approaches to dealing with missing outcomes.

From Saga we obtained tutoring attendance records, tutor characteristics, and student scores on Saga's internal math assessments. Because we hypothesized Saga tutoring would increase school persistence, and because of the link between educational attainment and crime involvement (Lochner and Moretti 2004), we also measured impacts on criminal behavior using administrative arrest records from the Chicago Police Department.

Our final data source came from two waves of in-person surveys carried out on behalf of the research team by the Institute for Social Research (ISR) at the University of Michigan. To develop this survey we drew on existing survey questions that have been used in previous studies of youth, including the Moving to Opportunity surveys.[11] The survey also included a math achievement test that we use as an alternative outcome to help mitigate missingness in the CPS test data. The ISR-administered test was based on a test designed for the eighth grade wave of the National Educational Longitudinal Study of 1988 (National Center for Educational Statistics 1988), which we supplemented with items from the fifth grade wave of the Early Childhood Longitudinal Study math assessment to help mitigate floor effects. The ISR survey was administered to a randomly selected subsample of 881 students in study 1 at the end of the first program year. The effective response rates were 88.2 percent and 90.6 percent for treatment and control, respectively. ISR administered a second wave of the survey to 1,238 randomly selected students in study 1 in the fall after the second intervention year (2014–2015). Response rates for the second

[11] See http://www2.nber.org/mtopublic/.

wave of the survey were 90.1 percent for the treatment group and 89.1 percent for the control group.

C. *Sample Characteristics*

Table 1 provides context for the study sample. The table shows the distribution of test scores measured the year before randomization. Panel A presents baseline test scores for study 1 (spring 2012–2013 AY) for all ninth and tenth grade male students in CPS, for all ninth and tenth grade students and male students in study schools, and students in the study sample; panel B presents the same for study 2 (spring 2013–2014 AY). The average test score among all ninth and tenth grade CPS students is close to the national median (forty-fifth percentile the year before study 1 for male students only and forty-ninth percentile the year before study 2), but the national norming of the test was based on assessments commonly taken in the fall. CPS students who took the tests in the spring were therefore slightly older and had more completed schooling than the norming sample of students. The average baseline test scores of students in our study samples were 8 to 15 percentile points lower than the CPS average, but similar to the average test scores of other ninth and tenth grade students in the study schools. This contrasts with many studies of "no excuses" charters where applicants tend to have slightly higher baseline scores than other students in the same school system (e.g., Angrist, Pathak, and Walters 2013).

Table 2 shows that the study 1 sample was split about evenly between Black and Hispanic youth. Almost 90 percent were eligible for subsidized lunch. The average GPA the year before our study was 2.11 on a 4-point scale. The study 2 sample was similar on these characteristics but included more Black and fewer Hispanic students and also included female students. Randomization appears to have been successful. We carried out an F-test of the null hypothesis that baseline characteristics are jointly the same for treatment and control by regressing a treatment-group indicator against all of our baseline covariates, controlling for randomization blocks. The p-values are 0.798 for study 1 and 0.273 for study 2.

D. *Analysis Plan*

Given our randomized design, the analysis plan is straightforward. We estimate both the ITT effect and the effect of TOT. The ITT estimate comes from estimating equation (1):

$$Y_i = \pi_0 + \pi_1 Z_i + X_i \pi_2 + B_i + \varepsilon_i, \tag{1}$$

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, ε_i is a random error term, and X_i is a set of baseline controls to improve precision.[12] To ensure our standard errors are not misleadingly small as an artifact of finite sampling issues (Young 2019), we also report p-values from a

[12] These include sociodemographics, average prerandomization test scores, and previous year GPA, days absent, days out-of-school suspension, disciplinary incidents, an indicator for having any arrests, and number of

Table 1—Baseline Test Score Comparison of Study Samples versus All Chicago Public School (CPS) Students

Sample	*N*	Mean
Panel A. Study 1 sample: prerandomization math scores (school year 2012–2013)		
All CPS ninth/tenth grade boys	31,064	44.76
All study school ninth/tenth grade students	4,406	35.40
All study school ninth/tenth grade boys	2,434	35.22
All randomized students	2,633	37.12
All participating treatment students	526	32.20
Panel B. Study 2 sample: prerandomization math scores (school year 2013–2014)		
All CPS ninth/tenth grade students	61,824	48.73
All study school ninth/tenth grade students	5,068	34.89
All randomized students	2,645	33.56
All participating treatment students	571	30.05

Notes: This table shows how the baseline characteristics of our high-dosage tutoring study samples compare to those of other students in the CPS system as a whole, as well as other students in the study subjects' schools specifically. These are percentile scores in the national test score distributions for the EXPLORE and PLAN tests in the 2012–2013 year and the Northwest Evaluation Association, EXPLORE, and PLAN tests in the 2013–2014 school year.

Table 2—Baseline Characteristics for High-Dosage Tutoring Study 1 and Study 2 Cohorts

	Study 1, $N = 2{,}633$		Study 2, $N = 2{,}710$	
Variable	Control mean	Treatment/control contrast	Control mean	Treatment/control contrast
Panel A. Demographic				
Age	14.807	−0.036 (0.025)	14.430	0.007 (0.022)
Female	0.002	0.000 (0.001)	0.307	0.012 (0.007)
Black	0.461	−0.003 (0.011)	0.643	−0.006 (0.012)
Hispanic	0.487	0.011 (0.013)	0.326	−0.001 (0.013)
Other	0.052	−0.008 (0.008)	0.031	0.007 (0.007)
Panel B. Education				
Free/reduced lunch recipient	0.871	0.002 (0.013)	0.902	−0.019 (0.012)
Has individualized education plan (IEP)	0.167	0.012 (0.015)	0.159	−0.002 (0.014)
Grade at study start	9.441	0.008 (0.014)	9.072	0.010 (0.008)
Math test score (CPS-wide *z*-score)	−0.357	−0.021 (0.034)	−0.465	−0.033 (0.032)
Reading test score (CPS-wide *z*-score)	−0.455	0.006 (0.033)	−0.465	−0.032 (0.034)
GPA	2.109	0.009 (0.036)	2.386	−0.043 (0.031)
Days absent	20.551	0.329 (0.838)	15.901	−0.332 (0.724)
Panel C. Disciplinary				
Out-of-school suspensions	1.612	−0.005 (0.165)	1.197	0.050 (0.147)
Disciplinary incidents	1.242	0.000 (0.101)	0.786	0.002 (0.082)
Panel D. Arrest				
Number of arrests for violent crimes	0.131	0.007 (0.021)	0.117	−0.014 (0.019)
Number of arrests for property crimes	0.090	−0.016 (0.017)	0.067	−0.006 (0.015)
Number of arrests for drug crimes	0.070	−0.004 (0.014)	0.046	−0.009 (0.011)
Number of arrest for other crimes	0.226	−0.015 (0.032)	0.190	−0.036 (0.031)
Ever arrested for any crime	0.187	−0.009 (0.014)	0.148	0.008 (0.014)

Notes: F-test for treatment-control comparison for all baseline characteristics: study 1: $p = 0.798$; study 2: $p = 0.273$; pooled study 1 and 2: $p = 0.525$. All tests control for randomization block fixed effects, which were defined at the school-grade level in study 1, which only included boys, and at the school-grade-gender level in study 2. Randomization block fixed effects do not perfectly predict gender because of differences in the administrative records at the time of randomization and the time baseline covariates were pulled. The study 2 sample was 30.8 percent female. Average test scores are averaged across all assessments taken in the baseline year. For each assessment, test scores are standardized and converted to *z*-scores within grade among all CPS students that took that assessment. The 2013–2014 and 2014–2015 school years included 178 and 180 days of school, respectively. Some students ($N = 65$) were randomized into study 2 twice. Both observations and treatment assignments are retained ~~in the table~~ above. Heteroskedasticity robust standard errors, clustered by individual in study 2, are in parentheses.

nonparametric permutation test (Efron and Tibshirani 1993). We randomly reassign the treatment indicator 100,000 times, storing the t-test statistic (T) in each replication, then calculate the share of replications where this exceeds the t-test statistic from using actual treatment assignment, T^*, or $\frac{1}{p}\sum_{i=1}^{p}\mathbf{1}\{|T_i| > |T^*|\}$. We also show that our results are robust to clustering by math teacher.

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

$$D_i = \gamma_0 + \gamma_1 Z_i + X_i \gamma_2 + B_i + \mu_i, \tag{2}$$

where D_i is an indicator for having participated in Saga tutoring (defined as having participated in at least one Saga tutoring session), the γs are parameters to be estimated, μ is a random error term, and all other variables are defined as above. The relationship of interest is

$$Y_i = \beta_0 + \beta_1 \hat{D}_i + X_i \beta_2 + B_i + \vartheta_i. \tag{3}$$

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).

Another methodological issue has to do with statistical inference in the presence of multiple testing. We grouped our outcomes into four different "families" that we expect to be affected in a similar way by the intervention: (i) mathematics achievement; (ii) achievement in other academic subjects, which could be improved if tutoring either improves overall study skills or a student's commitment to school; (iii) school behavior, which could change if students become more committed to school or teachers see students differently as a result of higher math or nonmath achievement; and (iv) out-of-school behavior (arrests), which previous research links to educational attainment (Lochner and Moretti 2004). We calculate the false discovery rate (FDR) "q-value," defined as the share of statistically significant estimates within a family that are expected to be false positives (Benjamini and Hochberg 1995). The selection of outcome "families" is currently more art than science, leaving the door open to multiple testing concerns across not just within families. Replication of our results across two experiments provides some additional protection against "false positives" in our analysis.

The final analyses presented here deviate slightly from our preanalysis plan in two ways. The preanalysis plan specified separate analyses for math achievement test scores; arrests for violent, property, drug, and other crimes; and a single index of CPS schooling outcomes that consists of the student absentee rate, number of student misconducts, total courses failed, and school persistence (enrollment or graduation status by the end of the academic year).[13] One change we have made is to split

violent, property, drug, and other arrests. Missing values are imputed with zeros and an indicator for missingness is included when necessary.

[13] AEA RCT Registry ID AEARCTR-0000041 (https://www.socialscienceregistry.org/trials/41).

out separately the two types of outcomes within the CPS schooling outcome index: school behavior (absences, misconducts, persistence) versus course grades (failures in all subjects), motivated partly by the observation that the "signal" captured by student misconducts has changed over time as CPS relies less on formal disciplinary actions (Stevens et al. 2015; Adukia, Feigenberg, and Momeni 2022).[14] Of course, we recognize that looking at additional outcomes increases the risk of false positives, so the second change we make is to carry out the multiple testing corrections described above, which were not specified in the preanalysis plan. We also again view the replication of results across the two large-scale RCTs as some additional protection against risk of false positives.

E. *Main Results*

The rates of participation in at least one Saga tutoring session for the treatment group were 40.2 percent in study 1 and 36.9 percent in study 2 (for controls the rates were 1.1 percent and 7.8 percent, respectively). The most common reasons for nonparticipation were (i) the student did not attend the school expected at the time of randomization (20.0 percent of the study 1 sample and 31.1 percent of the study 2 sample), or (ii) the student had a scheduling conflict and could not add Saga tutoring as a class in their schedule. It was rare for students to decline Saga or ask to be rescheduled if Saga tutoring had already been added to their schedule by default.

Table 3 shows the estimated ITT effect in study 1 on math achievement test scores (EXPLORE and PLAN tests) was 0.09 SD, with a TOT effect of 0.18 SD. A potential concern is that perhaps the tutors were just "teaching to the tests" rather than helping students to build broad knowledge. So it is notable that we saw changes in math *grades* as well, with a TOT impact of 0.57 points on a 1–4 GPA scale, relative to the control complier mean (CCM) of 1.62; this represents a change from about a C− to a C+ in the student's regular math class, which as a reminder was taught and graded by a classroom teacher not the Saga tutor.[15] We also estimated a decline in percent of math courses failed of 48 percent of the CCM ($-0.086/0.178$). Other evidence that our results are not artifacts of tutors teaching narrowly to the primary CPS accountability tests is that we found TOT effects of 0.20 SD on the supplemental math tests administered to students on our behalf by ISR, which were low stakes for both students and teachers (and which they did not know we would administer, nor did the tutors). These estimated effects are also statistically significant with respect to the FDR q-value that accounts for the number of tests in this family of outcomes.

The second panel of Table 3 shows that tutoring may have had positive spillovers on some outcomes in subject areas other than math. Reading test scores did not show significant impacts, but the TOT estimates indicate tutoring increased grades in nonmath courses by 0.17 points (relative to a CCM of 1.61). Furthermore, the TOT estimate implies that tutoring reduced the percentage of courses failed in nonmath

[14] When we analyze the CPS index as originally defined the result is statistically insignificant because while the academic coursework component is significant (course failures), the school behavior measures are not—results that are consistent with what we report in the tables of this paper.

[15] 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." (https://bigfuture.collegeboard.org/plan-for-college/college-basics/how-to-convertgpa-4.0-scale).

Table 3—Estimated Effects of High-Dosage Tutoring on Academic and Behavioral Outcomes in Study 1, Year 1

Outcome	*N*	Control mean	ITT estimate	TOT estimate	Control complier mean	FDR *q*-value
Panel A. Mathematics outcomes						
CPS-administered math test (study sample *Z*)	1,852	0.000	0.091 (0.035)	0.179 (0.066)	−0.111	0.009
Math GPA	2,215	1.760	0.279 (0.040)	0.571 (0.079)	1.617	0.001
Math courses failed (percent)	2,215	0.191	−0.042 (0.013)	−0.086 (0.026)	0.178	0.002
Research team-administered math test (study sample *Z*)	617	0.000	0.116 (0.056)	0.199 (0.090)	−0.102	0.029
Panel B. Nonmath academic outcomes						
CPS reading test (study sample *Z*)	1,851	0.000	0.017 (0.039)	0.033 (0.074)	−0.135	0.657
Nonmath GPA	2,244	1.739	0.083 (0.033)	0.173 (0.068)	1.611	0.018
Nonmath core courses failed (percent)	2,244	0.210	−0.027 (0.011)	-0.057 (0.022)	0.221	0.018
Panel C. Disciplinary outcomes						
Disciplinary incidents	2,494	1.513	0.082 (0.104)	0.189 (0.235)	1.505	0.631
Days absent	2,633	23.182	0.180 (0.812)	0.441 (1.960)	23.971	0.823
Out-of-school suspensions	2,494	1.515	0.184 (0.153)	0.424 (0.345)	1.562	0.631
Panel D. Arrest outcomes						
Number of arrests for violent crimes	2,633	0.086	−0.016 (0.015)	−0.038 (0.035)	0.104	0.624
Number of arrests for property crimes	2,633	0.061	−0.010 (0.010)	−0.025 (0.025)	0.056	0.624
Number of arrests for drug crimes	2,633	0.057	0.019 (0.013)	0.047 (0.032)	-0.003	0.624
Number of arrests for other crimes	2,633	0.178	−0.002 (0.022)	−0.005 (0.053)	0.121	0.929
Ever arrested for any crime	2,633	0.176	−0.007 (0.013)	−0.018 (0.031)	0.148	0.825
Number of arrests for any crime	2,633	0.382	−0.008 (0.036)	−0.021 (0.087)	0.279	0.929

Notes: This table shows the impact of high-dosage tutoring on academic and behavioral outcomes in the first postrandomization school year for study 1. Nonmath GPA is calculated using grades in all nonmath courses in core subject areas (English, science, social science). All regressions control for randomization block fixed effects and baseline covariates, including sociodemographics; average prerandomization test scores; and previous year GPA; days absent; days out-of-school suspension; disciplinary incidents; an indicator for ever having been arrested; and number of violent, property, drug, and other arrests. Missing baseline covariate values are imputed zeros with indicators for missing covariates included. Only observations with observed outcomes are included. False discovery rate (FDR) *q*-values are the smallest level at which we can control the share of false positives in a family of outcomes and still reject the null for that outcome (Benjamini and Hochberg 1995). Families are defined by panels of the table. Heteroskedasticity robust standard errors are in parentheses.

classes by 26 percent $(-0.057/0.221)$. Both of these impacts are statistically significant based on analytic t-statistics and with respect to the FDR q-value. There did 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 do not take this as evidence that improved academic outcomes have no impact on crime involvement, so much as an indication that arrests are noisy outcomes and our statistical power to detect impacts on these outcomes is somewhat limited. (Survey-based measures of risky behavior show a similar pattern; see online Appendix Table 2).

Table 4 shows the test score effects of tutoring were at least as large in study 2 as in study 1. This suggests the learning gains in study 1 are not statistical flukes or the result of unusually good program implementation that cannot be replicated. The TOT effects for study 2 were a 0.40 SD increase on the EXPLORE and PLAN math tests, a 0.41 point increase in math grades (relative to a CCM of 1.80), and a 43 percent decline $(-0.080/0.185)$ in math courses failed. There were no statistically significant indications of spillovers on nonmath courses or behavioral outcomes in study 2 once we account for multiple testing.

Table 4—Estimated Effects of High-Dosage Tutoring on Academic and Behavioral Outcomes in Study 2, Year 1

Outcome	N	Control mean	ITT estimate	TOT estimate	Control complier mean	FDR q-value
Panel A. Mathematics outcomes						
CPS math test (study sample Z)	1,865	0.008	0.135 (0.036)	0.398 (0.105)	−0.172	0.001
Math GPA	2,061	1.859	0.144 (0.043)	0.412 (0.122)	1.795	0.002
Math courses failed (percent)	2,061	0.149	−0.028 (0.013)	−0.080 (0.037)	0.185	0.029
B. Nonmath academic outcomes						
CPS reading test (study sample Z)	1,865	0.007	0.002 (0.039)	0.005 (0.115)	−0.069	0.965
Nonmath GPA	2,110	1.936	0.063 (0.034)	0.181 (0.100)	1.823	0.208
Nonmath core courses failed (percent)	2,110	0.138	−0.010 (0.010)	−0.030 (0.028)	0.165	0.434
Panel C. Disciplinary outcomes						
Disciplinary incidents	2,474	1.556	−0.012 (0.139)	−0.037 (0.437)	1.762	0.933
Days absent	2,710	20.748	0.570 (0.789)	1.899 (2.625)	22.536	0.713
Out-of-school suspensions	2,474	0.733	0.065 (0.092)	0.206 (0.288)	0.569	0.713
Panel D. Arrest outcomes						
Number of arrests for violent crimes	2,710	0.104	−0.011 (0.016)	−0.038 (0.052)	0.137	0.559
Number of arrests for property crimes	2,710	0.072	−0.027 (0.017)	−0.090 (0.056)	0.137	0.219
Number of arrests for drug crimes	2,710	0.051	0.001 (0.012)	0.003 (0.040)	0.033	0.950
Number of arrests for other crimes	2,710	0.225	−0.048 (0.028)	−0.160 (0.092)	0.333	0.219
Ever arrested for any crime	2,710	0.164	−0.018 (0.013)	−0.059 (0.043)	0.194	0.255
Number of arrests for any crime	2,710	0.452	−0.086 (0.045)	−0.286 (0.151)	0.640	0.219

Notes: This table shows the impact of high-dosage tutoring on academic and behavioral outcomes in the first postrandomization school year for Study 2. Nonmath GPA is calculated using grades in all nonmath courses in core subject areas (English, science, social science). All regressions control for randomization block fixed effects and baseline covariates, including sociodemographics; average prerandomization test scores; previous year GPA; days absent; days out-of-school suspension; disciplinary incidents; an indicator for ever having been arrested; and number of violent, property, drug, and other arrests. Missing baseline covariate values are imputed zeros with indicators for missing covariates included. Only observations with observed outcomes are included. False discovery rate (FDR) q-values are the smallest level at which we can control the share of false positives in a family of outcomes and still reject the null for that outcome (Benjamini and Hochberg 1995). Families are defined by panels of the table. Some students ($N = 65$) were randomized into study 2 twice. Both observations and treatment assignments are retained in the table above. Heteroskedasticity robust standard errors, clustered by individual, are in parentheses.

Table 5 reports the results of pooling together the year 1 data from studies 1 and 2 to improve statistical power, which is particularly useful for examining impacts on outcomes that were not the main focus of the intervention (such as nonmath outcomes). In the pooled sample, the TOT estimate was a 0.28 SD increase in math test scores, a 0.52 point increase in math GPA relative to a CCM of 1.68, and a decline of 0.09 in the share of math courses failed, equal to 47 percent of the CCM. These effects remain statistically significant when inference is based on analytic standard errors (Table 5), standard errors clustered by math teacher (online Appendix Table 3), or a permutation test (online Appendix Table 4), and even using the FDR q-values from each of these variants to account for multiple testing. Even with the added power of the pooled sample we cannot detect effects on reading test scores or arrests and out-of-school suspensions (despite the large size of the TOT effects in proportional terms), although we do detect a 0.18 point increase in nonmath GPA, and a decline of 0.05 in share of nonmath courses failed.

The 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

Table 5—Estimated Effects of High-Dosage Tutoring on Academic and Behavioral Outcomes, Pooling Study 1 and 2

Outcome	N	Control mean	ITT estimate	TOT estimate	Control complier mean	FDR q-value
Panel A. Mathematics outcomes						
CPS math test (study sample Z)	3,717	0.004	0.119 (0.025)	0.282 (0.059)	-0.143	0.001
Math GPA	4,276	1.803	0.217 (0.029)	0.516 (0.069)	1.675	0.001
Math courses failed (percent)	4,276	0.173	−0.036 (0.009)	−0.086 (0.022)	0.184	0.001
Panel B. Nonmath academic outcomes						
CPS reading test (study sample Z)	3,716	0.003	0.008 (0.028)	0.019 (0.065)	−0.104	0.774
Nonmath GPA	4,354	1.825	0.076 (0.024)	0.184 (0.058)	1.699	0.005
Nonmath core courses failed (percent)	4,354	0.178	−0.020 (0.007)	−0.048 (0.018)	0.198	0.012
Panel C. Disciplinary outcomes						
Disciplinary incidents	4,968	1.533	0.042 (0.086)	0.111 (0.230)	1.612	0.631
Days absent	5,343	22.054	0.403 (0.564)	1.144 (1.598)	23.188	0.631
Out-of-school suspensions	4,968	1.162	0.129 (0.089)	0.343 (0.238)	1.078	0.449
Panel D. Arrest outcomes						
Number of arrests for violent crimes	5,343	0.094	−0.012 (0.011)	−0.035 (0.031)	0.117	0.287
Number of arrests for property crimes	5,343	0.066	−0.018 (0.010)	−0.050 (0.028)	0.091	0.287
Number of arrests for drug crimes	5,343	0.054	0.010 (0.009)	0.027 (0.025)	0.014	0.287
Number of arrests for other crimes	5,343	0.200	−0.024 (0.018)	−0.068 (0.051)	0.214	0.287
Ever arrested for any crime	5,343	0.171	−0.011 (0.009)	−0.030 (0.025)	0.164	0.287
Number of arrests for any crime	5,343	0.414	−0.044 (0.029)	−0.125 (0.083)	0.436	0.287

Notes: This table shows the impact of high-dosage tutoring on academic and behavioral outcomes in the first postrandomization school year pooling both studies. Nonmath GPA is calculated using grades in all nonmath courses in core subject areas (English, science, social science). All regressions control for randomization block fixed effects and baseline covariates, including sociodemographics; average prerandomization test scores; and previous year GPA; days absent; days out-of-school suspension; disciplinary incidents; an indicator for ever having been arrested; and number of violent, property, drug, and other arrests. Missing baseline covariate values are imputed zeros with indicators for missing covariates included. Only observations with observed outcomes are included. False discovery rate (FDR) q-values are the smallest level at which we can control the share of false positives in a family of outcomes and still reject the null for that outcome (Benjamini and Hochberg 1995). Families are defined by panels of the table. Some students ($N = 65$) were randomized into study 2 twice. Both observations and treatment assignments are retained in the table above. Heteroskedasticity robust standard errors, clustered by individual, are in parentheses.

cumulative.[16] Because we did not randomly assign the number of years students were offered tutoring we cannot cleanly distinguish the effects of receiving tutoring for one versus two years. Nevertheless, Angrist and Pischke (2009) show that if we use an average causal response model by adjusting equations (2) and (3) so D_i is defined by years of participation instead of any participation, we can interpret the TOT estimate for study 1 youth in their second postrandomization school year, $\tilde{\beta}_1$, as a weighted average of the persistent impact of the first year of participation, call this Δ_1, and the impact of a second year of participation, Δ_2. While we cannot separately identify Δ_1 and Δ_2 with a single randomization, the weights are equal to the first-stage impact on the probability of having at least one year of participation

[16] Even though students were offered a second year of tutoring as an option, take-up rates were fairly low for reasons that we believe are more relevant for questions about maximum possible dosage rather than maximum possible scale. Specifically, while Math Lab is a credit-bearing course in the CPS system for students, CPS also has a set of specific course requirements students need to take for graduation as well, and so while students could replace one elective in one year with Math Lab, that was difficult to do for two years in a row and still complete all the other courses that would be required for graduation.

or having exactly two years of participation, normalized by the first-stage impact on the overall years of participation.

The TOT estimate of the effect per year of tutoring participation on math test scores at the end of the second year is 0.23 SD (see Table 6, panel B). The first stage impacts shown in Table 6, panel A imply the weights on Δ_1 and Δ_2 are 0.75 (0.42/0.56) and 0.25 (0.14/0.56), respectively. Assuming both of these effects are nonnegative, we can estimate an upper bound for Δ_1 by assuming $\Delta_2 = 0$, i.e., that all of the observed impacts at the end of year 2 are persistent effects from the first year of participation, so $0 \leq \Delta_1 \leq \tilde{\beta}_1/w_1$. Table 6, panel B shows that the upper bound of this persistent effect is 0.31 SD. The final column shows the analogous upper bound for Δ_2, assuming that all of the observed impacts at the end of year 2 are the result of participating in two years of tutoring with no persistent effect from the students who participated in the first year but not the second, is 0.84 SD.

We also find that the effects of tutoring seem to persist over time. Table 7 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 eleventh grade year if they were not retained in a grade. Interpretation of these results could be complicated if there are treatment-control differences in being held back in school, but we can rule out effects on grade retention larger than plus or minus 3 percentage points in ITT terms. We find TOT effects on math test scores in eleventh grade of 0.23 SD, similar to the pooled impact measured at the end of tutoring, while the TOT effect on math grades is 0.25 GPA points, just under half of the short-term effect.

Table 7 also shows that the estimated TOT effect on graduating on time from high school, pooling the study 1 and 2 samples again, is 1.3 percentage points relative to a CCM of 78.3 percent, but is imprecisely estimated. The standard error of 3.2 percentage points means we cannot rule out a decline in graduation rates as large as -5.0 percentage points or an increase as large as $+7.6$ points.[17] The point estimate for the effects of tutoring on graduation is close to the effect we might expect from higher math test scores alone. This comes from multiplying the experimental impact on math test scores (0.28 SD, Table 5) by the coefficient of a nonexperimental regression of graduating on time on ninth grade math test scores controlling for student characteristics, which suggests higher test scores would boost graduation by 3.2 percentage points—well within the confidence interval of our estimated effect of tutoring on graduation.[18]

F. *Robustness Checks and Extensions*

The estimated effects on our main outcomes are robust to a range of estimation decisions. Online Appendix Table 6 shows what happens when we change the set of baseline covariates we control for in our regression, while online Appendix Table 7 shows 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

[17] Results for the study 1 and 2 cohorts separately are in online Appendix Table 5.

[18] That regression uses data on $N = 24{,}782$ students in ninth grade in AY2013–14 for whom we have valid test scores and later graduation outcomes (82 percent of the total cohort of AY2013–14 ninth graders). The coefficient on ninth grade math scores in that regression equals 0.116.

Table 6—Estimated Effects of High-Dosage Tutoring on Academic and Behavioral Outcomes in Study 1, Year 2

Outcome	*N*	Explanatory variable: year 1 assignment to high-dosage tutoring
Panel A. First stage impacts		
Years of high-dosage tutoring	2,633	0.563 (0.020)
1 or more years of high-dosage tutoring	2,633	0.423 (0.014)
2 years of high-dosage tutoring	2,633	0.139 (0.010)

Outcome	*N*	Control mean	Explanatory variable: years of tutoring	Explanatory variable: at least 1 tear of tutoring	Explanatory variable: 2 years of tutoring
Panel B. Mathematics impacts					
CPS math test (study sample *Z*)	1,640	0.000	0.228 (0.050) [0.001]	0.305 (0.067) [0.001]	0.835 (0.189) [0.001]
Math GPA	1,841	1.879	0.180 (0.065) [0.012]	0.243 (0.088) [0.012]	0.630 (0.230) [0.013]
Math courses failed	1,841	0.295	−0.080 (0.035) [0.021]	−0.109 (0.047) [0.021]	−0.281 (0.123) [0.023]
Research-team administered math test (study sample *Z*)	878	0.000	0.139 (0.059) [0.021]	0.188 (0.080) [0.021]	0.512 (0.223) [0.023]

Notes: This table shows the impact of high-dosage tutoring on academic and behavioral outcomes in the second postrandomization school year for study 1. Panel A shows the first-stage impacts on years of participation, having at least one year of participation, and having two years of participation. Panel B shows the impacts on our main mathematics outcomes. The first set of results uses years of schooling as the endogenous variable in equations (2) and (3). The second and third sets of results show the upper bound on the impact of having at least one year of participation and of having two years of participation, respectively. See text for details. All regressions control for randomization block fixed effects and baseline covariates, including sociodemographics; average prerandomization test scores; and previous year GPA; days absent; days out-of-school suspension; disciplinary incidents; an indicator for ever having been arrested; and number of violent, property, drug, and other arrests. Missing baseline covariate values are imputed zeros with indicators for missing covariates included. Only observations with observed outcomes are included. False discovery rate (FDR) *q*-values are shown in brackets. These are the smallest level at which we can control the share of false positives in a family of outcomes and still reject the null for that outcome (Benjamini and Hochberg 1995). Families are defined by panels of the table.

Table 7—Estimated Effects of High Dosage Tutoring on 11th Grade Outcomes and High School Graduation

Outcome	*N*	Control mean	ITT estimate	TOT estimate	Control complier mean	FDR *q*-value
Panel A. 11th grade outcomes						
11th grade CPS math test (study sample *Z*)	2,973	0.006	0.099 (0.028)	0.232 (0.065)	−0.147	0.001
11th grade math GPA	3,019	1.993	0.109 (0.037)	0.251 (0.086)	1.837	0.004
Panel B. High school graduation outcomes						
Graduated on time	3,594	0.761	0.005 (0.012)	0.013 (0.032)	0.783	0.677
Graduated ever	3,614	0.831	0.000 (0.011)	0.000 (0.029)	0.865	0.996

Notes: This table shows the impact of high-dosage tutoring on long-run academic outcomes pooling both studies. All regressions control for randomization block fixed effects and baseline covariates, including sociodemographics; average prerandomization test scores; and previous year GPA; days absent; days out-of-school suspension; disciplinary incidents; an indicator for ever having been arrested; and number of violent, property, drug, and other arrests. Missing baseline covariate values are imputed zeros with indicators for missing covariates included. Only observations with observed outcomes are included. False discovery rate (FDR) *q*-values are the smallest level at which we can control the share of false positives in a family of outcomes and still reject the null for that outcome (Benjamini and Hochberg 1995). Families are defined by panels of the table. Some students ($N = 65$) were randomized into study 2 twice. Both observations and treatment assignments are retained in the table above. Heteroskedasticity robust standard errors, clustered by individual, are in parentheses.

assignment but wound up not enrolling at those schools in the fall. Online Appendix Table 8 shows the results using multiple imputation for missing outcomes, which assumes outcomes are missing at random, and, for continuous outcomes, quantile regression where missing values are imputed with arbitrarily low (in this case, zero) values, given baseline data suggesting those with missing tests are disproportionately students with low baseline test scores and grades. Online Appendix Table 9 shows Lee bounds (Lee 2009) of the impacts on our main outcomes include only beneficial impacts, except the lower bound on math test scores in study 1 is −0.008 and the lower bound on nonmath GPA in study 2 is −0.009.

As noted above, study 1 was part of a 2 × 2 factorial design. Including indicators into our estimating equation for BAM assignment does not materially change our tutoring effect estimates (see online Appendix Table 10). The same online Appendix exhibit shows that estimating the interaction between the academic and nonacademic interventions yields coefficients of 0.08, −0.04, and 0.004 on math test scores, math GPA, and math course failures. These interactions are imprecisely estimated and are statistically insignificant at conventional levels. Replication of study 1's tutoring impacts in the study 2 sample (which did not have any BAM randomization) provides further confirmation that the study 1 results are not influenced by BAM interactions.

Online Appendix Table 11 shows results for ninth graders, for whom the counterfactual for tutoring was typically a second period of "double dose" algebra while online Appendix Table 12 shows results for tenth graders, for whom the counterfactual was an elective course a student would have chosen to take. The similarity in results for ninth and tenth graders is consistent with Nomi and Allensworth's (2009) findings that CPS double dose algebra has limited benefits.

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. Online Appendix Table 13 shows that the results for female students were not so different from those of the full study 2 results that pool males and females together. The racial/ethnic composition of the study 1 sample (46 percent Black students) was also somewhat different from the study 2 sample (64 percent Black students). We tested the interaction of treatment with student race/ethnicity but did not see detectably different effects for Black and Hispanic students (see online Appendix Table 14).

Our estimates understate true effects if the SUTVA assumption is violated—e.g., if control students benefit from higher-achieving treatment peers. Under the assumption that this attenuation is more pronounced when controls are more exposed to tutoring participants, in online Appendix Figure 1 we plot randomization-block-specific TOT effects against block-specific treatment assignment rates. We find the treatment effect seemed to *increase* (rather than decrease) with a larger share of individuals within a block randomized to treatment. This is the opposite of what we would expect if treatment spillovers were attenuating our estimates.

A final reason we may understate impacts is possible floor effects in achievement tests.[19] To test for floor effects we group students into baseline achievement

[19] For example if a ninth grade student started the intervention year with third grade-level math skills and tutoring increased those skills by three grade levels, we would not measure these learning gains if the tests include no items below (say) a seventh or eighth grade level.

bins in two ways: first, by using the average of every baseline math assessment we have for each student (80 percent of all students have results for at least 2 tests); and second, by building a machine learning model that predicts achievement at the end of the tutoring intervention year as a function of all baseline achievement measures (see online Appendix II for details). In the left-hand panel of Figure 1 we plot the point estimate and 95 percent confidence interval for the effects on math GPA by baseline achievement quartiles under both grouping approaches. The left-hand panel of the figure shows that we find positive treatment effects on math GPA for every baseline achievement. In contrast, the right-hand panel shows that we find positive treatment effects on math test scores for all but the bottom baseline achievement quartile.

This sort of comparison in the magnitude of test score gains across baseline achievement quartiles is complicated by the fact that grades and test scores are ordinal not cardinal measures (see for example Bond and Lang 2018). So in online Appendix III we show that we get similar results when we anchor test scores in adult earnings as in Cunha and Heckman (2008). These findings taken together are consistent with floor effects.

III. Mechanisms

To understand the potential mechanisms through which tutoring may affect student learning we develop a simple model in the spirit of Lazear (2001). The purpose of the model is to help us understand the relative benefits of putting students in a regular classroom with a relatively highly-paid credentialed teacher, versus greatly reducing class size in a budget-neutral way by using a lower-paid, less-trained instructor. This comparison corresponds roughly to the treatment and control conditions in the experiment, thus the model can make predictions about conditions that might generate larger versus smaller treatment effects from tutoring. Lazear notes the public good nature of a classroom means class size reduction should be relatively more beneficial in settings where the level of behavioral disruptions is higher. Our model extends this idea by making the frequency of disruption endogenous to the heterogeneity in student achievement in the classroom. If teachers must pick some achievement level to teach to, students further from that target achievement level are more likely to create disruptions because they are confused or bored, or disrupt regular instruction by asking questions that are not relevant for the rest of the class. Therefore the higher the variance in achievement in a classroom, the greater the frequency of classroom questions and disruptions. Small-group tutoring would therefore generate larger reductions in disruptions, and therefore larger learning gains, relative to counterfactual classrooms that are more heterogeneous.

A similar mechanism can arise if teachers try to "personalize" instruction by working with small groups of similar-ability students in the classroom. The more heterogeneous is achievement in the class, the less teacher time each student receives (because to personalize instruction to a more heterogeneous classroom, the teacher has to create more ability groups). Since each student would receive less direct attention from the teacher in more heterogeneous classrooms, the contrast in direct instruction time would also be larger and tutoring would generate larger learning gains relative to more heterogeneous classrooms.

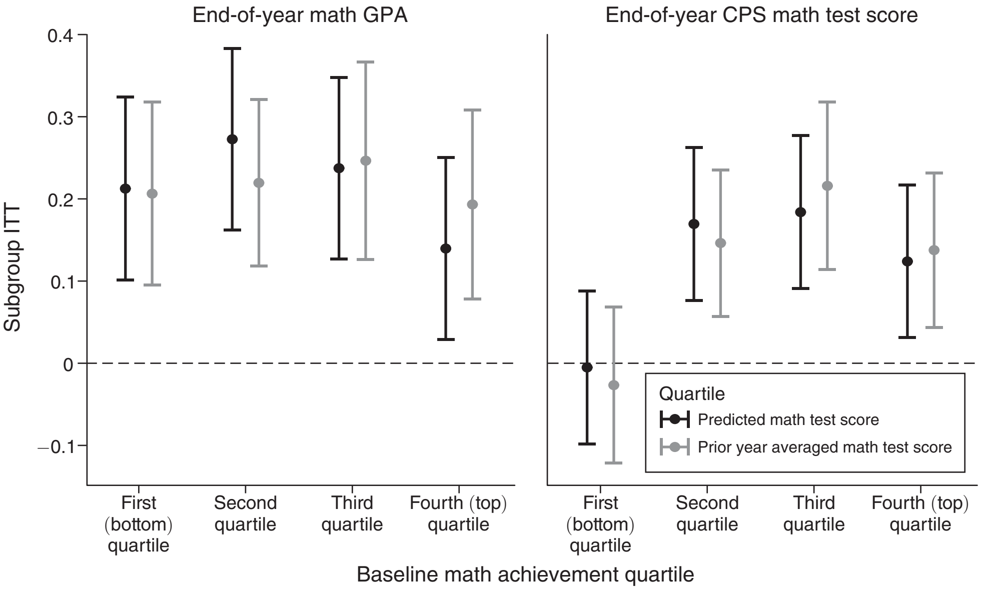

Figure 1. High-Dosage Tutoring Effects on Math Test Scores and Math GPA by Baseline Achievement Quartile

Notes: Figure shows the effects of high-dosage tutoring on math GPA (left panel) and CPS-administered math test score (right panel) separately for each baseline math achievement quartile, defined in two different ways. First, we use the average of all the baseline math test scores we have for each student. Second, we build a machine learning model to predict end-of-treatment year math test scores for the control group using all the baseline covariate information available for students (see online Appendix III). Estimates are from our ITT specification replacing treatment assignment with treatment assignment interacted with indicators for each group with appropriate main effects added, including block fixed effects and our usual set of baseline covariates. Because we include the full set of treatment interactions, estimates are interpretable as the ITT within each group. Error bars show 95 percent confidence intervals.

To model this situation, we assume a school has S students and a budget M to spend on m teachers. Teacher quality depends on the wage according to $V(w)$, where $V'(w) > 0$. The school can hire M/w teachers yielding an average class size of $n = wS/M$. Each student's skill level is drawn independently from a distribution $N(\mu, \sigma^2)$. Mirroring Lazear (2001), students only learn when there are no distractions, which occurs with probability p^n (or $p^{wS/M}$) where p depends on classroom heterogeneity: $p(\sigma^2) = e^{-\sigma^2}/(1 + e^{-\sigma^2})$. Alternatively, one could assume this is capturing that each student only learns when their teacher is focused on their achievement level. The school chooses wages according to the following optimization problem:

$$\max_w SV(w)p(\sigma^2)^{wS/M}.$$

The comparative static of the optimal wage with respect to classroom heterogeneity is

$$\partial w^*/\partial\sigma^2 = \frac{\{S/M[1 - p(\sigma^2)]\}V(w^*)^2}{V(w^*)V''(w^*) - V'(w^*)^2}.$$

We show in online Appendix IV that this is negative so long as teacher quality is not too convex in wages. This implies a school should trade teacher quality for smaller class sizes (that is, shift in the direction of high dosage tutoring) as heterogeneity in student achievement increases or, holding wages fixed, that we should see bigger benefits from tutoring in classrooms with greater dispersion in students' baseline skills.

Figure 2 shows that the ability of tutoring to help address behavioral disruptions does not seem to be a key mechanism behind these effects because impacts do not seem to be larger for students whose classroom settings have higher levels of disruption. One data challenge we face is that CPS accurately records what *teacher* a given student has, but seems to record less reliably which specific *section* of the teacher's class any given student would have been in. It is not obvious which approach to measuring classroom environments is better (assigning the average classroom characteristics of all sections taught by a given teacher, or assigning the features of a given course section given it might be the wrong section) so we show the estimates both ways. Figure 2 shows the results of re-estimating the TOT model with interactions between treatment assignment and prevalence of classroom disruptions measured in different ways (misconducts or suspensions). Whether we use math GPA (left-hand panel) or math test scores (right-hand panel) as the outcome, the coefficients on the interaction term are modest in magnitude with confidence intervals that include zero, whether we assign students to classrooms by either section or teacher (see online Appendix Table 15 for more details). The only exception is that treatment effects on math test scores are significantly smaller in classrooms with a higher percentage of students with any baseline misconduct. This is the opposite of what we would expect to find if the ability of tutoring to help address behavioral disruptions was a key mechanism driving our effects.

In contrast, Figure 3 suggests that the benefits of assignment to tutoring may be larger for students whose classroom environments have higher variance in student achievement levels. This comes from re-estimating the TOT model interacting treatment assignment with different measures of heterogeneity in student achievement in the classroom environment. The impact of classroom heterogeneity on tutoring treatment effects is positive in all but one case, although somewhat imprecisely estimated, whether we measure achievement using either math GPA (left-hand panel) or math test scores (right-hand panel). The interaction with math GPA heterogeneity is particularly pronounced if we assign students to classroom characteristics based on teacher (rather than teacher and section).[20] This finding is consistent with the importance of personalization of instruction implied by our model, and also with the results of Banerjee et al.'s (2007) study of the Pratham NGO in India and Duflo, Dupas, and Kremer's (2011) study of tracking in Kenya.

In further support of personalization being a likely mechanism for the benefits of tutoring, we show the results are likely not due to a generic "mentoring" effect that arises simply from connecting youth to a prosocial adult for so many hours over the school year. In the developing-country context of India, Banerjee et al. (2007) found qualitatively similar impacts from Pratham's tutoring-like intervention and a

[20] See online Appendix Table 15 for more details. Also, online Appendix Figure 4 shows that the classroom math heterogeneity measures are almost completely uncorrelated with impacts on reading outcomes.

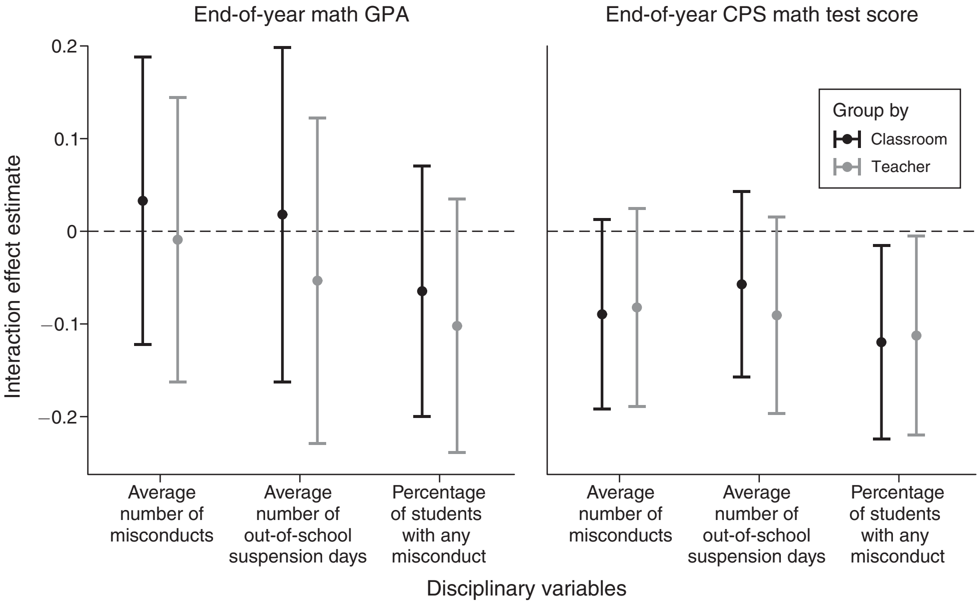

Figure 2. Heterogeneity by Classroom Discipline

Notes: Figure shows the coefficient on the interaction between treatment assignment and different measures of heterogeneity in baseline classroom behavior for each student in the study sample. The CPS data on classroom assignments for students are noisy for assigning students to a specific classroom or "section," but we believe are more reliable for assigning students at least to the correct teacher. So we replicate the results first defining classroom at what we believe to be the actual classroom section (recognizing that is noisy), and then replicate counting all students assigned to the same teacher as a "classroom" (recognizing that adds measurement error of a different sort). Estimates are from our TOT specification with the interaction between treatment and each measure added along with the appropriate main effects. All specifications also control for block fixed effects and our usual set of baseline covariates. Figure plots the coefficient on the interaction between treatment and the specified measure and 95 percent confidence intervals.

computer-assisted learning intervention, which is more consistent with the influence of the two mechanisms highlighted by our model than with a generic mentoring effect. In our survey data collected for this study, we find no detectable effects on the number of adults students say they have to talk to or who care about them, although the confidence intervals do not let us rule out the possibility this has increased by one adult (see online Appendix Table 16).[21] We also do not see detectable changes in the nonacademic or socioemotional skills we might expect any mentoring effects to operate through (for example Heller et al. 2017). Our 95 percent confidence intervals let us rule out ITT effects on z-score indices of 0.10 SD for grit, 0.16 SD for conscientiousness, and 0.10 SD for locus of control. There would seem to be limited scope for these mechanisms to drive sizable test score gains unless there is some important interaction effect between any mentoring effect and the academic elements of tutoring.[22]

[21] Online Appendix Table 16 shows the results from our first wave of ISR surveys done at the end of the first program year (2013–2014). These results are comparable to the findings from our second wave of ISR surveys administered at the end of the second program year (results for second wave of ISR surveys can be provided upon request).

[22] For example, the program we study here tries to have students work with the same tutor all year to strengthen the relationship; it is possible that rotating tutors across students could lead to smaller impacts. We cannot test that possibility with the data we have available here.

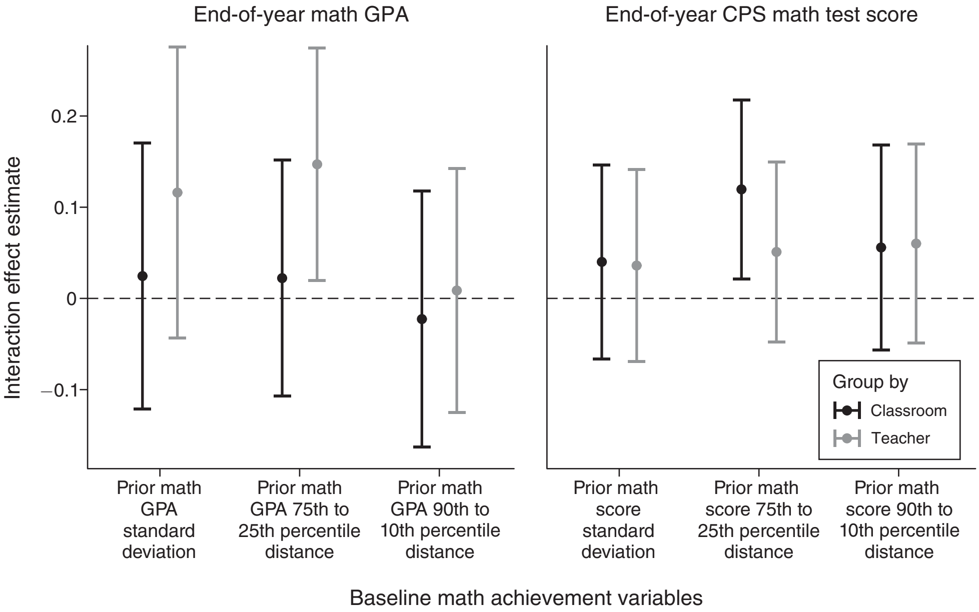

FIGURE 3. HETEROGENEITY BY CLASSROOM MATH SKILLS

Notes: Figure shows the coefficient on the interaction between treatment assignment and different measures of heterogeneity in classroom math achievement for each student in the study sample. Estimates are from our TOT specification replacing treatment assignment with treatment assignment interacted with indicators for each group with appropriate main effects added, including block fixed effects and our usual set of baseline covariates. Because we include the full set of treatment interactions, estimates are interpretable as the TOT within each group. Figure plots point estimates and 95 percent confidence intervals. The CPS data on classroom assignments for students are noisy for assigning students to a specific classroom or "section," but we believe are more reliable for assigning students at least to the correct teacher. So we replicate the results first defining classroom at what we believe to be the actual classroom section (recognizing that is noisy), and then replicate counting all students assigned to the same teacher as a "classroom" (recognizing that adds measurement error of a different sort).

IV. Benefit-Cost Analysis and Scaling

The benefit-cost ratio of Saga tutoring appears 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 class size reduction as implemented in the Tennessee STAR experiment.[23] At the time of our studies (2013–2015) the per-pupil cost was approximately $3,500 (defensible range of $3,200 to $4,800).[24] To estimate benefits, we adapt the approach of Kline and Walters (2016) and calculate that the present discounted value of the earnings gains induced by Saga's tutoring is about $11,500 for study 1 and $25,700 for study 2.[25] So the implied

[23] 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 0.06 in study 1 and 0.08 to 0.12 in study 2. Success for All, however, also improves reading scores (a 0.09 SD improvement per $1,000).

[24] See online Appendix V for a more detailed discussion of the cost estimates.

[25] This approach to extrapolating earnings benefits based on test score gains may be conservative. García, Heckman, and Ronda (forthcoming) show that this type of approach underestimates the observed long-run benefits of the Perry Preschool Program by two-thirds because it ignores the impacts on noncognitive skills.

benefit-cost ratios are 2.4–3.6 in study 1 and 5.4–8.0 in study 2.[26] (See the online Appendix for details). By way of comparison, estimated benefit-cost ratios for model early childhood programs are 1.9–2.2 for the Abecedarian Project (Masse and Barnett 2002), 3.9–6.8 for Perry Preschool (Heckman et al. 2010), and about 2 for a 7-student reduction in class sizes in grades K–3 (Krueger 2003).[27]

The biggest challenge to this scale strategy is likely to be cost, as we can see for example in the aftermath of the global COVID-19 pandemic. The US Secretary of Education, Miguel Cardona, encouraged districts to support high-dosage tutoring with at least part of the $122 billion the federal government provided to overcome pandemic-related learning loss (Locke 2022). Even with this one-time infusion of resources, however, districts can provide tutoring only to a modest share of all the students who would benefit.

A related challenge is labor supply. The "great resignation" that has come on the heels of the pandemic has made it difficult for school districts around the country to recruit core instructional staff like full-time teachers, much less tutors. Obviously, this comes back to cost: a high enough wage would lure more tutors but also (holding the budget fixed) increase per pupil costs and so reduce the number of children who could be served. Some districts have responded by, for example, hiring "virtual" tutors, which expands labor supply geographically and also allows schools to hire part-time tutors as well. The degree to which virtual versus in-person tutoring affects student learning is currently not known. And early feedback from districts trying to work with part-time tutors is that scheduling tutors and students together can be a challenge.

What the labor supply of potential tutors looks like, and how quickly tutor effectiveness might decline as scale increases, is also an open question. One suggestive data point is that in the literature review of high-dosage tutoring by Nickow, Oreopoulos, and Quan (2020), they compare the average effectiveness of tutoring carried out with (expensive) full-time, credentialed teachers with tutoring done by (less expensive) trained paraprofessionals. We might hypothesize that the difference in "quality" between teachers and paraprofessionals could be comparable to how quality might change within the pool of paraprofessionals as more and more are hired as tutoring expands. The Nickow, Oreopoulos, and Quan (2020) finding suggests paraprofessionals are nearly as effective, if not exactly as effective, as teachers, especially for the setting we study here (math instruction, with tutoring delivered at high dosages). That is consistent also with the results of Davis et al. (2017), who show that over the range of tutors hired by Saga Education in these studies to date, we see little difference in effectiveness (or "value added") between tutors at the top

[26] 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 Hispanic individuals from Chicago in the 2019 American Community Survey (and discounted to age 15 at a 5 percent rate) implies slightly smaller benefit-cost ratios of 1.4 to 2.2 for study 1 and 3.2 to 4.8 for study 2.

[27] Krueger's calculation uses a 4 percent discount rate and so would be below 2 with a 5 percent discount rate, which is the discount rate used in our calculations and those for Abecedarian and Perry.

versus bottom of Saga's ranked hiring list. That is, quality does not decline at the prevailing tutor wage at least at the current scale of hiring we have seen.[28]

One way to expand scale by simultaneously reducing the number of tutors that must be hired to serve a given number of students, and reduce cost per student, is to incorporate elements of computer assisted learning (CAL) into tutoring. In ongoing work, our team in partnership with CPS and Saga has implemented a "hybrid" model in which tutors work with students in person every other day (rather than every day), and students spend the off days at a separate desk next to the tutor on CAL software. This reduces tutoring costs by about a third and cuts the number of tutors that need to be hired in half. In our preliminary estimates, the treatment effects of the tutor-CAL hybrid model are similar to those reported here (Bhatt et al. 2022). Saga is also now trying a model that increases student-tutor caseloads by 50 percent again. To what degree would it be possible to increase student-tutor ratios and incorporate more CAL without reducing effectiveness is an open question.

A final thought about scaling is that our target scale will hopefully be somewhat lower once we are through the full after-effects of the pandemic. As has been noted the consequences of the pandemic for student learning have not only been severe, but also widespread. So almost all students would currently benefit from tutoring, since so many students are currently behind grade level and so will have trouble engaging with grade-level instruction. But eventually the hope would be that more students catch up, and so tutoring could be focused just on the students who are behind, with just enough tutoring delivered until they can fully engage again with regular classroom instruction. Additional discussion about scaling strategies are in Ander, Guryan, and Ludwig (2016) and Kraft and Falken (2021).

V. Conclusion

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 these changes (like lengthening the school day and year, or replacing all the principals and half the teachers) were only possible to implement within these particular Houston public schools because they were low-performing schools in danger of being taken over by the state. In Fryer's study some grades but not others got tutoring on top of the other "no excuses" changes. Among middle and high school students the gains for tutoring and nontutoring grades was 0.61 versus 0.21 SD, respectively, suggesting that for teens, tutoring might be the most important component strategy used in no excuses schools.

In the present paper, pooling the two study samples together we find the average effect of tutoring on high school students' math test scores is 0.28 SD, not

[28] The study by Davis et al. (2017) was carried out in the context of a tutoring program that provided tutors with four weeks of training and gave them access to a high-quality curriculum. There is some suggestion in past work that having a structured, quality curriculum can help accommodate tutors with lower levels of pedagogical training (see, e.g., Nickow, Oreopoulos, and Quan 2020; Rowan, Camburn, and Correnti 2004), although we have no data directly to examine how tutor effectiveness might change as the number of tutors hired increases in a setting with less training and a weaker curriculum. The field's investments in educational technology also raises the possibility of a different solution to the curriculum challenge, which is the possibility that tutors and the students they are tutoring essentially use a computer assisted learning program (of which there are now an increasing number of options) as the tutoring curriculum.

so dissimilar to Fryer's implied 0.4 SD effect (0.61–0.21) once we account for the standard errors around the estimates. By way of comparison, the gap in math test scores for Black and White eighth grade students in the National Assessment of Educational Progress is 0.8 SD.[29] So even a few years of high-dosage tutoring alone could substantially reduce educational disparities. As Nickow, Oreopoulos, and Quan (2020) described the results from the two RCTs we report on here, these effect sizes are "exceptional relative to the potential alternatives at the secondary level" (p. 36). The most important barrier to truly large-scale adoption is likely to be cost, despite the fact that a key innovation of the tutoring intervention we study here by its developer, Saga Education, is to lower cost by hiring paraprofessionals (rather than full-fledged teachers) as tutors.

The lesson is that it is possible to substantially improve academic skills by accounting for the challenges of individualizing instruction—among other things—and that these strategies can be effective even when implemented in traditional public high schools to broad, representative samples of students. These strategies seem to work even with secondary school students, yielding benefit-cost ratios comparable to promising early childhood programs. Evidently adolescence is not too late to realize large social benefits from human capital investment.

REFERENCES

▶**Abdulkadiroğlu, Atila, Joshua D. Angrist, Yusuke Narita, and Parag A. Pathak.** 2017. "Research Design Meets Market Design: Using Centralized Assignment for Impact Evaluation." *Econometrica* 85 (5): 1373–1432.

Adukia, Anjali, Benjamin Feigenberg, and Fatemeh Momeni. 2022. "From Retributive to Restorative: An Alternate Approach to Justice." Unpublished.

Allensworth, Elaine M., and John Q. Easton. 2005. *The On-Track Indicator as a Predictor of High School Graduation*. University of Chicago Consortium on Chicago School Research.

Ander, Roseanna, Jonathan Guryan, and Jens Ludwig. 2016. *Improving Academic Outcomes for Disadvantaged Students: Scaling Up Individualized Tutorials*. Washington, DC: Hamilton Project.

▶**Angrist, Joshua D., Sarah R. Cohodes, Susan M. Dynarski, Parag A. Pathak, and Christopher R. Walters.** 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.

▶**Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin.** 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91 (434): 444–55.

▶**Angrist, Joshua D., Parag A. Pathak, and Christopher R. Walters.** 2013. "Explaining Charter School Effectiveness." *American Economic Journal: Applied Economics* 5 (4): 1–27.

Angrist, Joshua D., and Jorn Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

▶**Banerjee, Abhijit V., Shawn Cole, Esther Duflo, and Leigh Linden.** 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *Quarterly Journal of Economics* 122 (3): 1235–64.

▶**Baye, Ariane, Amanda Inns, Cynthia Lake, and Robert E. Slavin.** 2019. "A Synthesis of Quantitative Research on Reading Programs for Secondary Students." *Reading Research Quarterly* 54 (2): 133–66.

▶**Benjamini, Yoav, and Yosef Hochberg.** 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.

[29] According to the National Assessment of Educational Progress Data Explorer, in 2019, White eighth grade students scored, on average, 32 points higher than Black eighth grade students—a gap equivalent to 0.8 standard deviations (the standard deviation was 40 for 2019).

Bhatt, Monica, Jonathan Guryan, Salman Khan, and Michael LaForest. 2022. "Increasing the Reach of Promising Individualized Academic Supports: Experimental Evidence of a Lower-Cost Technology-Infused Intervention." Unpublished.

Bhatt, Monica, Jonathan Guryan, Jens Ludwig, Anuj Shah, and the Chicago Youth Violence Project Team. 2021. "Scope Challenges to Social Impact." NBER Working Paper 28406.

▶**Bloom, Benjamin 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, Howard S.** 1984. "Accounting for No-Shows in Experimental Evaluation Designs." *Evaluation Review* 8 (2): 225–46.

▶**Bond, Timothy N., and Kevin Lang.** 2013. "The Evolution of the Black-White Test Score Gap in Grades K–3: The Fragility of Results." *Review of Economics and Statistics* 95 (5): 1468–79.

▶**Bond, Timothy N., and Kevin Lang.** 2018. "The Black–White Education Scaled Test-Score Gap in Grades K-7." *Journal of Human Resources* 53 (4): 891–917.

▶**Borman, Geoffrey D., and Gina M. Hewes.** 2002. "The Long-Term Effects and Cost-Effectiveness of Success for All." *Educational Evaluation and Policy Analysis* 24 (4): 243–66.

Carneiro, Pedro, and James J. Heckman. 2003. "Human Capital Policy." In *Inequality in America: What Role for Human Capital Policies?* edited by James J. Heckman and Alan B. Krueger, 77–240. Cambridge, MA: MIT Press.

Cascio, Elizabeth U., and Douglas O. Staiger. 2012. *Knowledge, Tests, and Fadeout in Educational Interventions*. NBER Working Paper 18038.

▶**Chetty, Raj, John N. Friedman, and Jonah E. Rockoff.** 2014a. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review* 104 (9): 2593–2632.

▶**Chetty, Raj, John N. Friedman, and Jonah E. Rockoff.** 2014b. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review* 104 (9): 2633–79.

Chicago Police Department. 2006–2015. "Chicago Police Department: Administrative Data System." Chicago, IL (accessed on a flow basis from 2015 to November 11, 2022).

Chicago Public Schools. 2011–2020. "Chicago Public Schools: Student Administrative Data System." Chicago, IL (accessed on a flow basis from 2015 to November 11, 2022).

▶**Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor.** 2010. "Teacher Credentials and Student Achievement in High School: A Cross-Subject Analysis with Student Fixed Effects." *Journal of Human Resources* 45 (3): 655–81.

Cook, Philip J., Kenneth Dodge, George Farkas, Roland G. Fryer Jr., Jonathan Guryan, Jens Ludwig, Susan Mayer, Harold Pollack, and Laurence Steinberg. 2014. "The (Surprising) Efficacy of Academic and Behavioral Intervention with Disadvantaged Youth: Results from a Randomized Experiment in Chicago." NBER Working Paper 19862.

Cook, Philip J. et al. 2014. "Improving Life Chances of Disadvantaged Youth: Testing Best-Practice Academic vs. Non-Academic Supports Through a Large-Scale Randomized Control Trial in Chicago." AEA RCT Registry. September 12. https://doi.org/10.1257/rct.41-1.0.

▶**Cullen, Julie Berry, Steven D. Levitt, Erin Robertson, and Sally Sadoff.** 2013. "What Can Be Done to Improve Struggling High Schools?" *Journal of Economic Perspectives* 27 (2): 133–52.

▶**Cunha, Flavio, and James J. Heckman.** 2008. "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Journal of Human Resources* 43 (4): 738–82.

Cunha, Flavio, James J. Heckman, Lance Lochner, and Dimitriy V. Masterov. 2006. "Interpreting the Evidence on Life Cycle Skill Formation." In *Handbook of the Economics of Education*, Vol. 1, edited by Eric Hanushek and Finis Welch, 697–812. Amsterdam: Elsevier.

▶**Cunha, Flavio, James J. Heckman, and Susanne M. Schennach.** 2010. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica* 78 (3): 883–931.

Davis, Jonathan M. V., Jonathan Guryan, Kelly Hallberg, and Jens Ludwig. 2017. "The Economics of Scale-Up." NBER Working Paper 23925.

▶**Dietrichson, Jens, Martin Bøg, Trine Filges, and Anne-Marie Klint Jørgensen.** 2017. "Academic Interventions for Elementary and Middle School Students with Low Socioeconomic Status: A Systematic Review and Meta-Analysis." *Review of Educational Research* 87 (2): 243–82.

▶**Dobbie, Will, and Roland G. Fryer Jr.** 2015. "The Medium-Term Impacts of High-Achieving Charter Schools." *Journal of Political Economy* 123 (5): 985–1037.

▶**Dobbie, Will, and Roland G. Fryer.** 2020. "Charter Schools and Labor Market Outcomes." *Journal of Labor Economics* 38 (4): 915–57.

▶**Duflo, Esther, Pascaline Dupas, and Michael Kremer.** 2011. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." *American Economic Review* 101 (5): 1739–74.

Duncan, Greg J., Chantelle J. Dowsett, Amy Claessens, Katherine Magnuson, Aletha C. Huston, Pamela Klebanov, Linda Pagani et al. 2007. "School Readiness and Later Achievement." *Developmental Psychology* 43 (6): 1428–46.

Efron, Bradley, and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Friedman, Jerome H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics* 29 (5): 1189–1232.

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.

Fryer, Roland. G., Jr. 2017. "The Production of Human Capital in Developed Countries: Evidence From 196 Randomized Field Experiments." In *Handbook of Economic Field Experiments*, Vol. 2, edited by Abhijit Vinayak Banerjee and Esther Duflo, 95–322. North-Holland: Elsevier.

García, Jorge Luis, James J. Heckman, and Victor Ronda. Forthcoming. "The Lasting Effects of Early Childhood Education on Promoting the Skills and Social Mobility of Disadvantaged African Americans." *Journal of Political Economy.*

Gilraine, Michael, Jiaying Gu, and Robert McMillan. 2020. "A New Method for Estimating Teacher Value-Added." NBER Working Paper 27094.

Gordon, Robert, Thomas J. Kane, and Douglas O. Staiger. 2006. *Identifying Effective Teachers Using Performance on the Job*. Washington, DC: Brookings Institution Press.

Guryan, Jonathan, Jens Ludwig, Monica P. Bhatt, Philip J. Cook, Jonathan M. V. Davis, Kenneth Dodge, George Farkas et al. 2013–2015. "University of Michigan Institute for Social Research Administered Surveys and Math Tests Data for 2013-14 and 2014-15." Chicago, Illinois (accessed on a flow basis from 2015 to November 11, 2022).

Guryan, Jonathan, Jens Ludwig, Monica P. Bhatt, Philip J. Cook, Jonathan M. V. Davis, Kenneth Dodge, George Farkas et al. 2023. "Replication Data for: Not Too Late: Improving Academic Outcomes among Adolescents." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E182903V1.

Hanushek, Eric A., Jacob D. Light, Paul E. Peterson, Laura M. Talpey, and Ludger Woessmann. 2020. "Long-Run Trends in the U.S. SES-Achievement Gap." NBER Working Paper 26764.

Hanushek, Eric A., and Ludger Woessmann. 2008. "The Role of Cognitive Skills in Economic Development." *Journal of Economic Literature* 46 (3): 607–68.

Hashim, Shirin A., Thomas J. Kane, Thomas Kelley-Kemple, Mary E. Laski, and Douglas O. Staiger. 2020. "Have Income-Based Achievement Gaps Widened or Narrowed?" NBER Working Paper 27714.

Heckman, James J. 2013. *Giving Kids a Fair Chance*. Cambridge, MA: MIT Press.

Heckman, James J., Seong Hyeok Moon, Rodrigo Pinto, Peter A. Savelyev, and Adam Yavitz. 2010. "The Rate of Return to the HighScope Perry Preschool Program." *Journal of Public Economics* 94 (1-2): 114–28.

Heller, Sara B., Anuj K. Shah, Jonathan Guryan, Jens Ludwig, Sendhil Mullainathan, and Harold A. Pollack. 2017. "Thinking, Fast and Slow? Some Field Experiments to Reduce Crime and Dropout in Chicago." *Quarterly Journal of Economics* 132 (1): 1–54.

Henry, Gary T., Kevin C. Bastian, and C. Kevin Fortner. 2011. "Stayers and Leavers: Early-Career Teacher Effectiveness and Attrition." *Educational Researcher* 40 (6): 271–80.

Illinois Report Card. n.d. "Illinois Report Card." Illinois Report Card. https://www.illinoisreportcard.com/District.aspx?source=trends&source2=sat&Districtid=15016299025 (accessed February 14, 2022).

Jackson, C. Kirabo. 2018. "What Do Test Scores Miss? The Importance of Teacher Effects on Non–Test Score Outcomes." *Journal of Political Economy* 126 (5): 2072–2107.

Jacob, Brian A., Jonah E. Rockoff, Eric S. Taylor, Benjamin Lindy, and Rachel Rosen. 2018. "Teacher Applicant Hiring and Teacher Performance: Evidence from DC Public Schools." *Journal of Public Economics* 166: 81–97.

Kane, Thomas J., and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." NBER Working Paper 14607.

Kline, Patrick, and Christopher R. Walters. 2016. "Evaluating Public Programs with Close Substitutes: The Case of Head Start*." *Quarterly Journal of Economics* 131 (4): 1795–1848.

Knudsen, Eric I., James J. Heckman, Judy L. Cameron, and Jack P. Shonkoff. 2006. "Economic, Neurobiological, and Behavioral Perspectives on Building America's Future Workforce." *Proceedings of the National Academy of Sciences* 103 (27): 10155–62.

Kraft, Matthew A., and Grace T. Falken. 2021. "A Blueprint for Scaling Tutoring and Mentoring Across Public Schools." *AERA Open* 7 (1): 1–21.

Krueger, Alan B. 2003. “Inequality, Too Much of a Good Thing.” In *Inequality in America: What Role for Human Capital Policies?* Edited by James J. Heckman and Alan B. Krueger, 1–76. Cambridge, MA: MIT Press.

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

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

Lochner, Lance, and Enrico Moretti. 2004. “The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports.” *American Economic Review* 94 (1): 155–89.

Locke, Charley. 2022. “American Schools Got a $190 Billion Covid Windfall. Where Is It Going?” *New York Times*, September 8. https://www.nytimes.com/2022/09/08/magazine/covid-aid-schools.html.

Masse, Leonard N., and W. Steven Barnett. 2002. “A Benefit Cost Analysis of the Abecedarian Early Childhood Intervention.” National Institute for Early Education Research (NIEER), Rutgers, The State University of New Jersey,

National Center for Educational Statistics. 1988. “National Education Longitudinal Study of 1988 (NELS:88).” United States Department of Education. https://nces.ed.gov/surveys/nels88/.

Nickow, Andre, Philip Oreopoulos, and Vincent Quan. 2020. “The Impressive Effects of Tutoring on PreK-12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence.” NBER Working Paper 27476.

Nomi, Takako, and Elaine Allensworth. 2009. “Double-Dose Algebra as an Alternative Strategy to Remediation: Effects on Students’ Outcomes.” *Journal of Research on Educational Effectiveness* 2 (2): 111–48.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel et al. 2011. “Scikit-Learn: Machine Learning in Python.” *Journal of Machine Learning Research* 12 (10): 2825–30.

Pellegrini, Marta, Cynthia Lake, Amanda Neitzel, and Robert E. Slavin. 2021. “Effective Programs in Elementary Mathematics: A Meta-Analysis.” *AERA Open* 7 (January): 2332858420986211.

Reardon, Sean F. 2011. “The Widening Academic Achievement Gap between the Rich and the Poor: New Evidence and Possible Explanations.” In *Whither Opportunity? Rising Inequality, Schools, and Children’s Life Chances*, edited by Greg J. Duncan and Richard J. Murnane, 91–116. New York: Russell Sage Foundation Press.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. “Teachers, Schools, and Academic Achievement.” *Econometrica* 73 (2): 417–58.

Rockoff, Jonah E. 2004. “The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data.” *American Economic Review* 94 (2): 247–52.

Rothstein, Jesse. 2010. “Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement*.” *Quarterly Journal of Economics* 125 (1): 175–214.

Rowan, Brian, Eric Camburn, and Richard Correnti. 2004. “Using Teacher Logs to Measure the Enacted Curriculum: A Study of Literacy Teaching in Third-Grade Classrooms.” *Elementary School Journal* 105 (1): 75–101.

Saga Education. 2013–2015. “Saga Education: Student Attendance and Programmatic Data for CPS.” Chicago, Illinois (accessed on a flow basis from 2015 to November 11, 2022).

Stevens, W. David, Lauren Sartain, Elaine M. Allensworth, and Rachel Levenstein. 2015. *Discipline Practices in Chicago Schools*. Chicago: University of Chicago Consortium on Chicago School Research.

Tuttle, Christiana C., Philip Gleason, Virginia Knechtel, Ira Nichols-Barrer, Kevin Booker, Gregory Chojnacki, Thomas Coen, and Lisbeth Goble. 2015. *Understanding the Effect of KIPP as It Scales: Volume I, Impacts on Achievement and Other Outcomes*. Princeton, NJ: Mathematica Policy Research.

Young, Alwyn. 2019. “Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results.” *Quarterly Journal of Economics* 134 (2): 557–98.

Zou, Hui, and Trevor Hastie. 2005. “Regularization and Variable Selection via the Elastic Net.” *Journal of the Royal Statistical Society Series B* 67 (2): 301–20.