For Online Publication

Online Appendix for Not Too Late: Improving Academic Outcomes Among Adolescents

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

Technical Appendix

I. Saga Program Model and Tutor Selection Process

A. *Overview of Saga Tutoring Model*

Saga Education's high-dosage, two-on-one math tutoring model was developed at the Match Charter Public High School in Boston by Alan Safran (who subsequently co-founded Saga Education) and Michael Goldstein in 2004, about a decade before our RCT of the program. The program was introduced in Chicago Public Schools in academic years 2013-14 (study 1) and 2014-15 (study 2) for our RCT, across a total of 15 CPS high schools.[1] During the school day, students as part of their regular class schedule were assigned to participate in a tutoring session for one class period every day of the 180-day school year (45-50 minutes a day), for a total potential dosage of about 135 contact hours per year. Tutors taught six periods a day and worked with two students at a time during each tutoring session.

Each tutoring session was divided into two segments:

- The focus of the beginning of each tutoring session was on remediating students' skill deficits – focusing on Saga's self-developed Algebra 1 curriculum but teaching foundational mathematics skills where needed to access these algebraic concepts.
- The second part of each session was tied to what youth learn in their Algebra 1 classrooms. For example, a student would first do four to five minutes of warm-up problems before receiving 40 minutes of tutoring on material tailored to that student.

Saga also used frequent internal formative and summative assessments of student progress to continuously individualize instruction and benchmark achievement.

- Saga conducted daily one to three-question mini-assessments at the end of each day's lesson that allowed tutors to assess student understanding of the material covered during the class period and revise the next day's lesson as needed.
- Saga also divided the year into 7 to 10 course units, each with a pre-test and post-test to help tutors determine how much review time was needed before the next unit.
- Quarterly proficiency assessments consisting of 50 questions of basic math skills, were also administered at the beginning of the school year and up to four other times during the year. These tests assisted tutors in targeting specific areas that students had not yet mastered that would be taught in the next quarter.

These numerous assessments allowed tutors to constantly and consistently measure student progress and tailor curricula to meet their students' needs.

[1] For study 1, Saga was implemented in 12 CPS high schools. This number increased to 15 CPS high schools in Study 2. Guidance on how to incorporate the intervention into the CPS system came from a small-scale pilot study our team carried out the previous academic year (2012-13), which involved delivering our own version of the tutoring model in one high school. Details are reported in Cook et al. (2014).

In addition, each study school and team of tutors was overseen by a Saga site director who worked with mathematics teachers on a weekly basis to understand what standards were being taught in mainstream math classes so the Saga tutorial covered complementary concepts. In addition to overseeing communication with math teachers and other school staff, Saga site directors also handled behavioral issues in the tutoring room and offered daily feedback and professional development to the team of tutors at each school. Site directors observed each tutor briefly each day, and at greater length for a portion of a period about once a week – meeting with the tutor after the period to provide feedback and advice. Saga's curriculum team (which consisted of certified math teachers) also provided substantive training on math teaching skills, math content, and lesson preparation to tutors throughout the school year.

B. *Tutor Selection Process*

Saga hired 139 total tutors across both study years out of an estimated pool of approximately 1,200 applicants. As noted in the main paper, tutors were mostly recent college graduates hired based on their exhibit of strong math skills and strong interpersonal skills during Saga's hiring process.

The first stage of Saga's hiring process involved applicants submitting an online application with their resume. Applicants deemed promising were then screened in a phone interview by Saga staff. Those candidates who made it through the phone interview completed a screening assessment of high school math proficiency. Those applicants who passed the math assessment were invited to on-site interviews in Chicago. The on-site interview process included multiple interviews with Saga leadership, former tutors, and site directors. During these interviews, tutors were screened for strong math and interpersonal skills and required to demonstrate their ability to build relationships with students in a mock tutoring session with local youth. One of the key inputs that Saga hiring managers considered during this process was whether the students would want a particular applicant as a tutor, i.e., was an applicant able to make a connection with the students in a high-stakes situation.

C. *Saga Tutor Training*

For both study years, each tutor participated in roughly 100 hours of training prior to the start of the school year (full-time for most of four weeks in the summer). This training included: workshops on math pedagogy, specific tutoring techniques, sample tutorials, preparation for working in the classroom, and lectures and discussions with outside speakers about the landscape of the Chicago Public Schools and the issues confronting Chicago youth in underserved neighborhoods.

Significant time was spent on the teaching and practice of tutorial techniques – largely adapted from Doug Lemov's *Teach like a Champion* – to increase student engagement, set high behavioral and academic expectations for students, and give students the resources they needed to meet those expectations. Tutors also spent time learning how to deal with student behavioral problems and how to effectively de-escalate challenging situations. Several sessions were reserved for special trainings, including how to work with students with Individualized Education Plans (IEPs), how to work with students who are non-native English speakers, and how to creatively break down math concepts for different types of learners. Additional training

was also dedicated to programmatic and logistical information, including professionalism, logging in grades, and tracking students' performance. Finally, the remaining training was dedicated to cultural competence, parent engagement, and developing a deeper understanding of the unique environment tutors would face in the Chicago Public Schools.

II. Using Machine Learning to Build a Baseline Skill Proxy

Our goal is to generate a proxy for baseline academic skills that can both serve as a baseline covariate and help us explore mechanisms. A typical approach in the literature is to proxy for baseline ability with a single pre-randomization covariate like a baseline math test score or prior year grade. There are several reasons to think we can improve on this benchmark. A single test score is a noisy measure of baseline skills, may be missing more than other measures, and ignores all information other measures have about baseline skills.

We try to improve on this benchmark by developing a proxy for prior skills $\hat{Y}_0$ based on *predicted test scores*. Specifically, we seek to estimate a function that predicts a student's expected end-of-year test scores based on their covariates X if they did not receive the intervention. Formally, we can write this as:

$$\hat{Y}_i \;=\; \hat{Y}_0(X_i) \;=\; E[Y \mid X, T = 0]$$

We explore how the accuracy of $\hat{Y}_0(\cdot)$ is affected by both the choice of models used to estimate the function as well as the choice of what covariates are included in X, which include a single test score from the pre-randomization school year, the average of all baseline tests from year (t-1),[2] all other (t-1) covariates we have for students, and adding averages of (t-2) test scores.

In order to use predicted test scores in downstream statistical inference, we want the predicted test scores for each student to be an out-of-sample prediction - meaning that the predicted test score for student i is from a model that didn't use student i's data during the model training process. We accomplish this using a *cross-fitting* procedure where the dataset is first split into the treatment group and the control group. We train one model on the entire control group and use that model to generate predictions for students in the treatment group. Then, among the control group, we split students into K different partitions. Each partition is iteratively held-out, a model is trained on the remaining K-1 partitions, and predictions are generated for the students in the held-out partition.

$\hat{Y}_i \;=\; M_C(X)$ if i is in the treatment group
$\hat{Y}_i \;=\; M_{C-K_i}(X)$ if i is in control group

We use gradient boosting to generate these machine learning estimates, which is an ensemble method that combines many decision trees into a single, more accurate predictor (Friedman, 2001). The intuition behind gradient boosting is that the first decision tree fits a tree T(X) to model the relationship between the covariates and the target variable. The second decision tree

[2] While most 10th graders in our study samples have either no (t-1) test score available (16.4%) or only 1 (t-1) test score available (83.6%), most (83.7%) of the 9th graders in our study sample have 2 or more time (t-1) baseline tests in our dataset.

then fits a model T(X) to the residual between the target variable and the prediction from the first tree, allowing for the second tree to partially correct for errors made by the first. In general, the Kth tree is fit to the residual between the target variable and the discounted sum of predictions from the first K-1 trees. The output of a gradient boosting model is a discounted sum of the predictions from each tree. Formally, the optimization procedure of gradient boosting and the form of its predictions can be written as:

$$T_0 = argmin_{t \in T()} \sum_i (y_i - t(x_i))^2$$

$$T_K = argmin_{t \in T()} \sum_i ((y_i - \sum_{k=0}^{K-1} \alpha^k T_k(x_i)) - t(x_i))^2$$

$$T_K(X) = \sum_{k=0}^{K} \alpha^k T_k(X)$$

Unfortunately, finding the optimal decision tree is a computationally intractable problem, so most implementations of gradient boosting use a heuristic algorithm to approximate the optimization problem. Our work specifically uses scikit-learn's implementation of gradient boosting (Pedregosa et al. 2011), which in turn uses the CART algorithm to fit each decision tree. Furthermore, this method requires specifying a number of hyperparameters like the maximum depth of any given decision tree (at a decision tree of depth j can model interactions of order j-1), the number of trees in the gradient boosted forest, and the discount rate $\boldsymbol{\alpha}$ that scales the predictions of each tree. We use cross-validation to choose optimal values for these hyperparameters.

Finally, we tested one variant of gradient boosting that was modified in two ways to address two shortcomings of gradient-boosting. The first modification is that we replaced the first estimator in the gradient boosting ensemble with an ElasticNet so the initial estimator captures the linear relationship between the covariates and the outcome variable. The subsequent boosting rounds then use decision trees as usual - allowing the boosted trees to model the residual error after accounting for the linear structure. The second modification is that we repeat the training procedure 20 times with different random seeds and then average together the results to form the final predictions. This technique is known as "bagging" and is used to improve out-of-sample accuracy and stability for high-variance models such as gradient boosting[3].

The final refinement we implement capitalizes on the fact that while our study sample consists of only around 5,000 students total, we have data from the larger population of CPS students. Let O denote this set of 9th and 10th graders in CPS during the study years who were not in the treatment or control group. We trained a gradient boosting model on this observational sample to construct an "observational model" $M_O(X)$ that estimates $E_{(X,Y)\sim O}[Y|X]$. We then used that

[3] We also tested using OLS and Elastic Net regression (Zou and Hastie 2005), a regularized version of OLS. We found that OLS performed consistently worse out-of-sample than Elastic Net and gradient boosting. We found that Elastic Net performed similarly to vanilla gradient boosting and slightly worse than the modified gradient boosting algorithm when we used the standard set of features. However, when we expanded the feature set to include the predictions from the observational model, Elastic Net had the same level of accuracy as the modified gradient boosting algorithm.

observational model to make predictions for all students in the treatment and control group and include that prediction as a feature, in the gradient boosting algorithm described above access to in order to predict test scores for our actual study sample.

The upshot is that these methods do indeed let us construct a measure of baseline achievement that has much more signal than a single test score from baseline period (t-1). A simple OLS regression of a single test score against a student's test score from the post-treatment year yields an R-squared of 0.349. In contrast, our preferred machine learning algorithm as described above yields an R-squared of 0.543.

III. Anchoring Test Scores to Earnings

One methodological challenge of looking at heterogeneity in test score outcomes is that test scores are an ordinal measure of skills. In other words, the practical value of raising test scores by 5 points, for example, may vary depending on where the student is in the test score distribution. That might be a very large effect for a lower performing student but a small one for a high-performing student, or vice versa. To test how sensitive our floor effect results are to the ordinality in test scores, we flexibly anchor test scores to earnings (Cunha and Heckman 2008; Cunha, Heckman, and Schennach 2010; Bond and Lang 2013) to examine how gains in test scores translate into gains in earnings. This earnings analysis mirrors our findings that students in the upper quartiles of baseline math achievement benefit more from the intervention than students in the bottom quartile, indicating that floor effects are real and are not purely artifacts of using test scores as the main outcome.

The main empirical challenge is that we do not directly observe earnings for youth in our study (and would have to wait many years for youth to be closer to their prime earning ages). We overcome this data limitation by leveraging the fact that a subset of youth in our study were administered the math assessment from the National Educational Longitudinal Study of 1988 (NELS:88) by Educational Testing Services (ETS). Using the NELS:88, we flexibly estimate the relationship between a student's performance, as measured by their "ability score",[4] and future earnings, and then use that mapping to compute estimated future incomes for the students in our sample who take the same assessment.[5] However, not all youth in our sample took the NELS:88 assessment, so we impute ability scores scores using end-of-year standardized test scores when they are missing.[6] Further details on the estimation are provided below.

---

[4] A student's ability score is calculated using ETS's PARSCALE IRT program using students' NELS assessment responses. Scores were intentionally calibrated to be comparable between our sample and the NELS:88.

[5] The NELS:88 sample includes 12,144 individuals. We use estimated ability scores (e.g. "theta scores") based on students' performance on a standardized 8th grade math test. We measure adult income using the employment income of the respondent in 1999. After dropping individuals who are missing income or ability score data, our sample includes 10,098 observations. All estimates are weighted by the panel weight for the fourth follow-up sample. We estimate the relationship between ability scores and earnings using gradient boosting with a monotonicity constraint to enforce that earnings are increasing in scores. We tune the learning rate using 10 repetitions of 10-fold cross-validation and use early stopping to select the optimal number of trees. Our final estimates use the full dataset with gradient boosting with a learning rate of 0.3.

[6] We predict ability scores using polynomial regression with the single math test score predictor. We use 10-fold cross-validation to select the degree of polynomial that generates the highest out-of-sample R2. A 5th order polynomial yielded the highest out-of-sample R2.

Appendix Figure 2 shows the mapping between ability scores and earnings. The left panel shows estimates for students where we directly observe their ability scores. The right panel shows that the estimates for students with predicted ability scores look very similar to the estimates for students with observed ability scores.

Appendix Figure 3 replicates our analysis of how test score impacts vary with baseline math achievement quartiles using the predicted earnings instead of test scores. Mirroring the main estimates in the paper that directly use test scores, the estimates indicate no impact for the bottom quartile and increasingly positive impacts for the higher quartiles. The middle quartiles impacts, however, are sometimes imprecisely estimated and so are not significantly different from zero when we split the sample by whether or not the ability score is observed because of the reduced sample size.

IV. A Model of Class Size and Teacher Quality with Endogenous Classroom Disruption

Consider the model presented in Section III of the main text. The FOC to the school's optimization problem is given by:

$$(w)\colon V'(w^*)/V(w^*) = -S/M \; ln\, p.$$

Taking the derivative with respect to classroom heterogeneity, σ^2, accounting for the fact that the wage, w, is an implicit function of classroom heterogeneity, we find that the comparative static with respect to classroom heterogeneity is:

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

S/M is the marginal impact of increasing the wage on the number of teachers. This is positive for all feasible wage offers. $1 - p(\sigma^2)$ is the proportional change in p from increasing σ^2 (because $1 - p = p'/p$) and is positive for finite σ^2. The final term is the inverse rate of change of the elasticity of teacher quality with respect to the wage ($= [\partial^2/\partial w^{*2} lnV(w^*)]^{-1}$). If $V(w)$ is concave (so $V'' < 0$) then the whole derivative is always negative. If the function is convex the whole derivative will be negative if:

$$V(w^*)V''(w^*) > V'(w^*)^2.$$

This implies the expression will be negative so long as the gradient of teacher quality with respect to wages is not too steep.

V. Estimating Program Cost

We measure the program's nominal cost directly using Saga's proposed budget for the two years of the experiment and a planning year. The total cost in year one and two was $2,582,140 and $3,648,153, respectively.

This budget information is shown in Appendix Table 17. To support thinking about how average costs might change with program scale, line items are placed into approximate variable and fixed cost groups in Panels A and B, respectively. We say these are approximate categories because there is some discretion in the labels. Tutor stipends and benefits are clearly variable costs as more tutors would be required if the program is to serve more students. Program management, on the other hand, increases with the scale of the program, but possibly at a slower rate than tutors. Curriculum development costs should be relatively fixed.

The largest expenses are tutor stipends and benefits which account for just under half of expenses across the two program years. Program management and instructional support are the next largest expenses accounting for 21 and 8 percent of overall costs, respectively. Variable costs account for about 82% of the overall program cost.

Appendix Table 18 shows how we combine this total cost information with program size details to calculate per pupil costs. The program budget assumed the program would serve 670 students in year one, the 2013-14. school year, and 1,130 students in year two, the 2014-15 school year. At full capacity, the average total cost per slot is $3,854 in year one and $3,228 in year two. The average variable costs are $3,135 and $2,651 in these years. The total column includes the costs in the planning year so yields average costs closer to the higher year 1 values.

In the main text we report the per-pupil cost of Saga is approximately $3,500 with a defensible range of $3,200 to $4,800. This is roughly the average total cost per treatment slot across the two program years, ignoring the sunk costs from the planning year. Outside the context of an RCT, Saga is likely to operate at closer to full capacity because it has more flexibility in filling slots. The range of estimates is defined by the average total cost per treatment slot in year two ($3,228) and the average total cost per participant in year one ($4,835).

If the program is scaled as it was implemented in the study years, the average variable cost may better represent the marginal cost of students. Using an analogous approach as above would yield a cost estimate of about $2,800 with a defensible range of $2,600 to $3,900. As we mention in the paper, however, 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.

Of course, there are other complications that arise when trying to estimate the economic cost of the program. In their analysis of the benefits and costs of the Perry Preschool program, Heckman et al. (2010) highlight that these types of programs are often financed with public funds and there may be a deadweight loss from the taxation required to raise these funds. To account for this deadweight loss, they present cost estimates that are inflated by 0, 50, and 100 percent. Our baseline estimates do not make this adjustment but doing so would deflate the benefit-cost ratio by a third or half.

Appendix Tables

Appendix Table 1: Missing Outcome Data by Study

Variable	Control Mean	Treatment/Control Contrast
Study 1, N = 2633		
Missing Math Test - Program Year 1	0.298	-0.016 (0.018)
Missing Reading Test - Program Year 1	0.297	-0.014 (0.018)
Missing Math GPA - Program Year 1	0.162	-0.008 (0.014)
Missing Non-Math Core GPA - Program Year 1	0.146	0.002 (0.013)
Missing Attendance - Program Year 1	0.044	0.018 (0.009)
Missing Math Test - Program Year 2	0.374	-0.009 (0.018)
Missing Reading Test - Program Year 2	0.375	-0.011 (0.018)
Missing Math GPA - Program Year 2	0.305	-0.020 (0.018)
Missing Non-Math Core GPA - Program Year 2	0.284	-0.015 (0.017)
Missing Attendance - Program Year 2	0.147	-0.010 (0.014)
Study 2, N = 2710		
Missing Math Test - Program Year 1	0.312	-0.001 (0.018)
Missing Reading Test - Program Year 1	0.310	0.003 (0.018)
Missing Math GPA - Program Year 1	0.257	-0.029 (0.016)
Missing Non-Math Core GPA - Program Year 1	0.237	-0.024 (0.016)
Missing Attendance - Program Year 1	0.089	-0.002 (0.011)

Notes: All tests control for block fixed effects. 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, in parentheses.

Appendix Table 2: Impacts on Self-Reported Risky Behavior and Crime Victimization by Study 1 Subjects: End of Second Program Year

Outcome	N	Control Mean	Intent-to-Treat Estimate	FDR q-value
A. Risky Behavior				
During your life, how many days have you had at least one drink of alcohol? (Z)	888	0.000	-0.197 (0.063)	0.033
During the past 30 days, on how many days did you have at least one drink of alcohol? (Z)	890	0.000	-0.181 (0.064)	0.044
During your life, how many times have you used marijuana? (Z)	884	0.000	-0.102 (0.067)	0.417
During the past 30 days, how many times did you use marijuana? (Z)	886	0.000	-0.048 (0.067)	0.703
During your life, how many times have you tried any other sort of illegal drug/inhalant/prescription drug? (Z)	889	0.000	-0.119 (0.066)	0.348
Do any of your brothers, sisters, cousins, or friends belong to a gang? (Dummy)	887	0.318	-0.015 (0.032)	0.773
Do you belong to a gang? (Dummy)	889	0.079	-0.014 (0.017)	0.703
Have you ever sold marijuana or any other drug to your friends? (Dummy)	888	0.133	-0.033 (0.022)	0.417
Have you ever sold marijuana or any other drug to people you didn't know? (Dummy)	888	0.105	-0.025 (0.019)	0.428
During the past 3 months with how many people did you have sexual intercourse? (Z)	557	0.000	-0.184 (0.150)	0.464
How many times have you gotten someone pregnant? (Z)	558	0.000	0.051 (0.087)	0.703
In the past year, how many times did you get in a physical fight in which you were so badly injured that you were treated by a doctor or a nurse? (Z)	895	0.000	-0.016 (0.074)	0.888
In the past year, how often did you hurt someone badly enough in a physical fight that he or she needed to be treated by a doctor or nurse? (Z)	895	0.000	-0.143 (0.066)	0.188
During the past 30 days, on how many days did you carry a weapon – such as a gun, knife, or club – to school? (Z)	892	0.000	-0.039 (0.065)	0.703
In the past year, how often did you paint graffiti or signs on someone else's property or in a public place? (Z)	895	0.000	-0.044 (0.064)	0.703
In the past year, how often did you deliberately damage property that didn't belong to you? (Z)	896	0.000	-0.047 (0.066)	0.703
In the past year, how often did you take something from a store without paying for it? (Z)	894	0.000	-0.014 (0.069)	0.888
In the past year, how often did you drive a car without owner's permission? (Z)	895	0.000	-0.011 (0.079)	0.888
In the past year, how often did you break into someone's home in order to steal? (Z)	893	0.000	-0.075 (0.053)	0.424
B. Crime Victimization				
In the past year, how often did someone pull a gun/knife on you? (Z)	894	0.000	0.023 (0.070)	0.742
In the past year, how often did you get into a physical fight? (Z)	894	0.000	-0.061 (0.069)	0.742
In the past year, how often did you get jumped? (Z)	896	0.000	0.038 (0.072)	0.742
In the past year, how often did you get beaten up and something was stolen from you? (Z)	894	0.000	0.046 (0.092)	0.742

Notes: All items are coded so the desired effect direction is positive. Baseline covariates and randomization block fixed effects included in all models (see text). Heteroskedasticity-robust standard errors in parentheses. 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.

Appendix Table 3: Estimated Pooled 1 Year Treatment Effects on Academic and Behavioral Outcomes - Clustering by Math Teacher

Outcome	N	Control Mean	Intent-to-Treat Estimate	Treatment-on-the-Treated Estimate	Control Complier Mean	FDR q-value
A. Mathematics Outcomes						
CPS Math Test (Study Sample Z)	3364	-0.005	0.116 (0.028)	0.260 (0.060)	-0.121	0.001
Math GPA	4013	1.834	0.211 (0.032)	0.497 (0.073)	1.695	0.001
Math Courses Failed (%)	4013	0.163	-0.037 (0.010)	-0.088 (0.022)	0.186	0.001
B. Non-math Academic Outcomes						
CPS Reading Test (Study Sample Z)	3363	-0.015	0.009 (0.027)	0.021 (0.061)	-0.106	0.734
Non-Math GPA	4053	1.864	0.063 (0.023)	0.150 (0.054)	1.734	0.012
Non-Math Core Courses Failed (%)	4053	0.165	-0.019 (0.007)	-0.045 (0.017)	0.195	0.012
C. Disciplinary Outcomes						
Disciplinary Incidents	4079	1.671	-0.032 (0.097)	-0.075 (0.232)	1.798	0.813
Days Absent	4079	23.951	0.209 (0.549)	0.498 (1.307)	23.834	0.813
Out-of-School Suspensions	4079	1.218	0.022 (0.093)	0.052 (0.221)	1.369	0.813
D. Arrest Outcomes						
# Arrests for Violent Crimes	4079	0.092	-0.020 (0.011)	-0.049 (0.027)	0.131	0.348
# Arrests for Property Crimes	4079	0.060	-0.011 (0.011)	-0.027 (0.025)	0.068	0.348
# Arrests for Drug Crimes	4079	0.049	0.009 (0.011)	0.021 (0.027)	0.020	0.433
# Arrests for Other Crimes	4079	0.184	-0.021 (0.019)	-0.049 (0.045)	0.196	0.348
Ever Arrested for Any Crime	4079	0.167	-0.013 (0.011)	-0.031 (0.027)	0.165	0.348
# Arrests for Any Crime	4079	0.386	-0.043 (0.034)	-0.103 (0.082)	0.414	0.348

Notes: This table shows our main effect estimates pooling both studies when we cluster standard errors by students' math teacher (526 clusters). Non-math GPA is calculated using grades in all non-math courses in core subject areas (English, Science, Social Science). All regressions control for randomization block fixed effects and baseline covariates, including socio-demographics, average pre-randomization 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 math teacher, in parentheses.

Appendix Table 4: Estimated 1 Year Effects on Academic and Behavioral Outcomes, Pooling Study 1 and 2: Permutation Test

Outcome	N	Control Mean	Intent-to-Treat Estimate	Permutation P-value	FDR q-value
A. Mathematics Outcomes					
CPS Math Test (Study Sample Z)	3717	0.004	0.119 (0.025)	0.007	0.008
Math GPA	4276	1.803	0.217 (0.029)	0.000	0.001
Math Courses Failed (%)	4276	0.173	-0.036 (0.009)	0.001	0.002
B. Non-math Academic Outcomes					
CPS Reading Test (Study Sample Z)	3716	0.003	0.008 (0.028)	0.759	0.760
Non-Math GPA	4354	1.825	0.076 (0.024)	0.092	0.138
Non-Math Core Courses Failed (%)	4354	0.178	-0.020 (0.007)	0.066	0.138
C. Disciplinary Outcomes					
Disciplinary Incidents	4968	1.533	0.042 (0.086)	0.285	0.427
Days Absent	5343	22.054	0.403 (0.564)	0.556	0.557
Out-of-School Suspensions	4968	1.162	0.129 (0.089)	0.199	0.427
D. Arrest Outcomes					
# Arrests for Violent Crimes	5343	0.094	-0.012 (0.011)	0.359	0.431
# Arrests for Property Crimes	5343	0.066	-0.018 (0.010)	0.035	0.210
# Arrests for Drug Crimes	5343	0.054	0.010 (0.009)	0.598	0.599
# Arrests for Other Crimes	5343	0.200	-0.024 (0.018)	0.328	0.431
Ever Arrested for Any Crime	5343	0.171	-0.011 (0.009)	0.244	0.431
# Arrests for Any Crime	5343	0.414	-0.044 (0.029)	0.172	0.431

Notes: Permutation tests were performed by randomly shuffling treatment assignment at the randomization block level and performing a (2-sided) t-test at each repetition. We then calculate the share of replications where this exceeds the t-test statistic using actual treatment assignment. This process is repeated for 100000 repetitions for each outcome. Non-math GPA is calculated using grades in all non-math courses in core subject areas (English, Science, Social Science).

Appendix Table 5: High-Dosage Tutoring Effects on 11th Grade Outcomes and High School Graduation - by Study

Outcome	N	Control Mean	Intent-to-Treat Estimate	Treatment-on-the-Treated Estimate	Control Complier Mean	FDR q-value
A. Study 1						
i. Eleventh Grade Outcomes						
11th Grade CPS Math Test (Study Sample Z)	1528	0.010	0.159 (0.039)	0.304 (0.074)	-0.219	0.001
11th Grade Math GPA	1554	2.015	0.132 (0.053)	0.250 (0.100)	1.865	0.013
ii. High School Graduation Outcomes						
Graduated On-Time	1819	0.752	0.001 (0.017)	0.001 (0.037)	0.779	0.972
Graduated Ever	1825	0.832	0.004 (0.015)	0.008 (0.033)	0.871	0.815
B. Study 2						
i. Eleventh Grade Outcomes						
11th Grade CPS Math Test (Study Sample Z)	1445	0.000	0.027 (0.040)	0.082 (0.122)	0.008	0.502
11th Grade Math GPA	1465	1.964	0.082 (0.053)	0.242 (0.156)	1.808	0.121
ii. High School Graduation Outcomes						
Graduated On-Time	1775	0.772	0.011 (0.018)	0.033 (0.057)	0.786	0.557
Graduated Ever	1789	0.830	-0.004 (0.016)	-0.011 (0.052)	0.861	0.831

Notes: This table shows the impact of high-dosage tutoring on long-run academic outcomes separated by study. Non-math GPA is calculated using grades in all non-math courses in core subject areas (English, Science, Social Science). All regressions control for randomization block fixed effects and baseline covariates, including socio-demographics, average pre-randomization 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, in parentheses. Students may have multiple 11th grade years.

Appendix Table 6: Sensitivity of Intent-to-Treat Estimates to Choice of Baseline Covariates (Pooled Data from Study 1 and 2)

Outcome	N	All Baselines	Sociodemographic Baselines	Academic Baselines	Arrest Baselines	No Baselines
A. Mathematics Outcomes						
CPS Math Test (Study Sample Z)	3717	0.119 (0.025)	0.093 (0.031)	0.122 (0.025)	0.099 (0.034)	0.095 (0.034)
Math GPA	4276	0.217 (0.029)	0.197 (0.034)	0.228 (0.030)	0.202 (0.034)	0.199 (0.035)
Math Courses Failed (%)	4276	-0.036 (0.009)	-0.034 (0.010)	-0.039 (0.009)	-0.035 (0.010)	-0.034 (0.010)
B. Non-math Academic Outcomes						
CPS Reading Test (Study Sample Z)	3716	0.008 (0.028)	-0.013 (0.033)	0.004 (0.028)	-0.009 (0.034)	-0.014 (0.034)
Non-Math GPA	4354	0.076 (0.024)	0.056 (0.029)	0.080 (0.025)	0.061 (0.029)	0.059 (0.030)
Non-Math Core Courses Failed (%)	4354	-0.020 (0.007)	-0.016 (0.008)	-0.021 (0.008)	-0.018 (0.008)	-0.017 (0.009)
C. Disciplinary Outcomes						
Disciplinary Incidents	4968	0.042 (0.086)	0.063 (0.092)	0.027 (0.091)	0.047 (0.091)	0.043 (0.094)
Days Absent	5343	0.403 (0.564)	0.591 (0.645)	0.343 (0.569)	0.516 (0.633)	0.483 (0.657)
Out-of-School Suspensions	4968	0.129 (0.089)	0.150 (0.097)	0.120 (0.096)	0.147 (0.094)	0.138 (0.098)
D. Arrest Outcomes						
# Arrests for Violent Crimes	5343	-0.012 (0.011)	-0.011 (0.011)	-0.014 (0.011)	-0.011 (0.011)	-0.012 (0.011)
# Arrests for Property Crimes	5343	-0.018 (0.010)	-0.017 (0.010)	-0.019 (0.010)	-0.018 (0.010)	-0.018 (0.010)
# Arrests for Drug Crimes	5343	0.010 (0.009)	0.008 (0.010)	0.006 (0.010)	0.010 (0.009)	0.007 (0.010)
# Arrests for Other Crimes	5343	-0.024 (0.018)	-0.025 (0.020)	-0.032 (0.019)	-0.024 (0.018)	-0.030 (0.020)
Ever Arrested for Any Crime	5343	-0.011 (0.009)	-0.009 (0.010)	-0.014 (0.010)	-0.010 (0.009)	-0.012 (0.010)
# Arrests for Any Crime	5343	-0.044 (0.029)	-0.046 (0.033)	-0.058 (0.032)	-0.043 (0.030)	-0.053 (0.034)

Notes: This table explores the sensitivity of the impact of high-dosage tutoring on academic and behavioral outcomes in the first post-randomization school year pooling both studies to the set of baseline covariates that are included. We group our standard set of baseline covariates into the following groups. Socio-demographics: indicators for age, having a learning disability, being eligible for free or reduced-price lunch, being Black or Hispanic. Academic baselines: Average pre-randomization math and reading test scores, number of disciplinary incidents, and number of out-of-school suspensions, number of As, Bs, Cs, Ds, and Fs. Arrest baselines: An indicator ever having been arrested and number of arrests for violent, property, drug, and other crimes. Each column shows the ITT estimate controlling for the set of baselines described in the column title, missing indicators for the set of covariates that are included, and block fixed effects. 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 parentheses.

Appendix Table 7: Estimated Pooled 1 Year Treatment Effects on Academic and Behavioral Outcomes - Omitting No-Shows

Outcome	N	Control Mean	Intent-to-Treat Estimate	Treatment-on-the-Treated Estimate	Control Complier Mean	FDR q-value
A. Mathematics Outcomes						
CPS Math Test (Study Sample Z)	3013	-0.003	0.124 (0.028)	0.249 (0.056)	-0.100	0.001
Math GPA	3694	1.808	0.231 (0.031)	0.495 (0.066)	1.718	0.001
Math Courses Failed (%)	3694	0.169	-0.036 (0.010)	-0.077 (0.020)	0.171	0.001
B. Non-math Academic Outcomes						
CPS Reading Test (Study Sample Z)	3011	-0.022	0.009 (0.031)	0.019 (0.062)	-0.109	0.760
Non-Math GPA	3741	1.826	0.068 (0.025)	0.146 (0.054)	1.751	0.022
Non-Math Core Courses Failed (%)	3741	0.179	-0.019 (0.008)	-0.041 (0.017)	0.186	0.023
C. Disciplinary Outcomes						
Disciplinary Incidents	3959	1.591	0.019 (0.090)	0.044 (0.203)	1.628	0.829
Days Absent	3977	24.550	0.310 (0.650)	0.704 (1.474)	23.580	0.829
Out-of-School Suspensions	3959	1.322	0.068 (0.103)	0.154 (0.233)	1.279	0.829
D. Arrest Outcomes						
# Arrests for Violent Crimes	3977	0.102	-0.014 (0.013)	-0.033 (0.030)	0.118	0.394
# Arrests for Property Crimes	3977	0.074	-0.024 (0.011)	-0.054 (0.026)	0.089	0.229
# Arrests for Drug Crimes	3977	0.057	0.008 (0.011)	0.018 (0.025)	0.021	0.481
# Arrests for Other Crimes	3977	0.207	-0.020 (0.021)	-0.046 (0.047)	0.180	0.394
Ever Arrested for Any Crime	3977	0.186	-0.016 (0.010)	-0.037 (0.024)	0.165	0.268
# Arrests for Any Crime	3977	0.440	-0.051 (0.034)	-0.115 (0.076)	0.408	0.268

Notes: This table shows the impact of high-dosage tutoring on academic and behavioral outcomes in the first post-randomization school year pooling students from both studies when we restrict the sample to students who attended a study school in the fall after randomization (as expected). Non-math GPA is calculated using grades in all non-math courses in core subject areas (English, Science, Social Science). All regressions control for randomization block fixed effects and baseline covariates, including socio-demographics, average pre-randomization 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, in parentheses.

Appendix Table 8: Variations on Missing Outcome Data Imputation, Pooling Studies 1 and 2

Outcome	Intent-to-Treat Estimate	Quantile Regression	Multivariate Imputation via Chained Equations
A. Study 1			
i. Math Outcomes			
CPS Math Test (Study Sample Z)	0.091 (0.035)	0.055 (0.024)	0.081 (0.042)
Math GPA	0.279 (0.040)	0.313 (0.053)	0.274 (0.04)
Math Courses Failed (%)	-0.042 (0.013)		-0.043 (0.013)
ii. Other Outcomes			
CPS Reading Test (Study Sample Z)	0.017 (0.039)	-0.002 (0.026)	-0.001 (0.039)
Non-Math GPA	0.083 (0.033)	0.047 (0.048)	0.084 (0.033)
Non-Math Core Courses Failed (%)	-0.027 (0.011)		-0.027 (0.011)
Days Absent	0.180 (0.812)	-0.196 (0.466)	0.16 (0.803)
B. Study 2			
i. Math Outcomes			
CPS Math Test (Study Sample Z)	0.135 (0.036)	0.077 (0.021)	0.087 (0.04)
Math GPA	0.144 (0.043)	0.216 (0.061)	0.136 (0.042)
Math Courses Failed (%)	-0.028 (0.013)		-0.026 (0.013)
ii. Other Outcomes			
CPS Reading Test (Study Sample Z)	0.002 (0.039)	-0.008 (0.022)	-0.021 (0.039)
Non-Math GPA	0.063 (0.034)	0.125 (0.044)	0.066 (0.035)
Non-Math Core Courses Failed (%)	-0.010 (0.010)		-0.01 (0.01)
Days Absent	0.570 (0.789)	0.205 (0.434)	0.572 (0.789)

Notes: We present our standard results alongside different approaches to imputing missing data. We run median quantile regression after imputing 0's for the outcome variables. We calculate bootstrap standard errors. We also perform multiple imputation via chained equations (denoted 'MI'). We impute M=50 datasets and pool the estimated effects and robust standard errors.

Appendix Table 9: Lee Bounds - Study 1 and Study 2

Outcome	N	Control Mean	Treatment Mean	Intent-to-Treat Estimate	Lower Bound	Upper Bound	Confidence Interval
A. Study 1							
i. Mathematics							
CPS Math Test (Study Sample Z)	1854	-0.007	0.056	0.062 (0.050)	-0.008	0.127	[-0.148, 0.255]
Math GPA	2212	1.742	1.999	0.258 (0.052)	0.237	0.278	[0.131, 0.385]
Math Courses Failed (%)	2212	0.194	0.151	-0.043 (0.015)	-0.052	-0.041	[-0.088, -0.015]
ii. Non-math Academics							
CPS Reading Test (Study Sample Z)	1853	-0.008	-0.012	-0.004 (0.050)	-0.072	0.040	[-0.219, 0.146]
Non-Math GPA	2243	1.723	1.784	0.061 (0.046)	0.056	0.068	[-0.043, 0.177]
Non-Math Core Courses Failed (%)	2243	0.214	0.191	-0.023 (0.013)	-0.024	-0.021	[-0.049, 0.014]
iii. Disciplinary & Attendance							
Disciplinary Incidents	2494	1.521	1.597	0.076 (0.129)	0.046	0.349	[-0.167, 0.594]
Days Absent	2494	24.438	25.266	0.827 (1.054)	0.354	2.625	[-1.409, 4.645]
Out-of-School Suspensions	2494	1.554	1.742	0.188 (0.177)	0.158	0.594	[-0.136, 0.926]
iv. Graduation							
Graduated On-Time	1823	0.745	0.739	-0.006 (0.021)	-0.011	0.008	[-0.049, 0.057]
Graduated Ever	1829	0.825	0.829	0.004 (0.018)	0.001	0.020	[-0.031, 0.069]
B. Study 2							
i. Mathematics							
CPS Math Test (Study Sample Z)	1868	0.022	0.130	0.108 (0.049)	0.101	0.117	[-0.079, 0.310]
Math GPA	2058	1.853	1.980	0.127 (0.050)	0.054	0.199	[-0.057, 0.309]
Math Courses Failed (%)	2058	0.149	0.128	-0.021 (0.014)	-0.053	-0.017	[-0.093, 0.007]
ii. Non-math Academics							
CPS Reading Test (Study Sample Z)	1868	0.014	-0.007	-0.021 (0.048)	-0.035	0.004	[-0.161, 0.183]
Non-Math GPA	2109	1.932	1.978	0.046 (0.044)	-0.009	0.103	[-0.103, 0.202]
Non-Math Core Courses Failed (%)	2109	0.139	0.131	-0.008 (0.012)	-0.033	-0.004	[-0.069, 0.016]
iii. Disciplinary & Attendance							
Disciplinary Incidents	2474	1.718	1.695	-0.023 (0.190)	-0.044	-0.022	[-0.824, 0.303]
Days Absent	2474	22.654	23.291	0.637 (0.997)	0.542	0.653	[-3.169, 2.599]
Out-of-School Suspensions	2474	0.786	0.852	0.066 (0.112)	0.047	0.066	[-0.621, 0.264]
iv. Graduation							
Graduated On-Time	1775	0.773	0.768	-0.004 (0.021)	-0.008	0.006	[-0.045, 0.059]
Graduated Ever	1788	0.834	0.820	-0.014 (0.018)	-0.018	0.002	[-0.050, 0.053]

Notes: This table shows Lee Bounds on the impact of high-dosage tutoring on academic and behavioral outcomes in the first post-randomization school year for study 1 (panel A) and study 2 (panel B). (Lee, 2009). In contrast to our other tables, we control for blocking in this table using inverse propensity score weights.

Appendix Table 10: Main effects with BAM 2x2

Outcome	N	Controlling for BAM	Full Treatment Interactions		
		Assigned to Tutoring	Assigned to Tutoring	Assigned to BAM	BAM x Tutoring Assignment Interaction
A. Mathematics Outcomes					
CPS Math Test (Study Sample Z)	1852	0.093 (0.035)	0.059 (0.046)	-0.066 (0.049)	0.076 (0.069)
Math GPA	2215	0.281 (0.04)	0.298 (0.053)	-0.019 (0.056)	-0.038 (0.08)
Math Courses Failed (%)	2215	-0.042 (0.013)	-0.044 (0.017)	0.007 (0.019)	0.004 (0.026)
B. Non-math Academic Outcomes					
CPS Reading Test (Study Sample Z)	1851	0.02 (0.039)	0.034 (0.052)	-0.042 (0.054)	-0.032 (0.077)
Non-Math GPA	2244	0.084 (0.034)	0.078 (0.044)	-0.011 (0.047)	0.012 (0.067)
Non-Math Core Courses Failed (%)	2244	-0.028 (0.011)	-0.024 (0.014)	0.011 (0.016)	-0.008 (0.022)
C. Disciplinary Outcomes					
Disciplinary Incidents	2494	0.079 (0.105)	0.138 (0.134)	0.132 (0.156)	-0.13 (0.21)
Days Absent	2633	0.166 (0.81)	-0.918 (1.075)	-0.975 (1.096)	2.396 (1.627)
Out-of-School Suspensions	2494	0.174 (0.153)	0.179 (0.184)	0.207 (0.198)	-0.012 (0.305)
D. Arrest Outcomes					
# Arrests for Violent Crimes	2633	-0.016 (0.015)	-0.018 (0.019)	0.006 (0.019)	0.004 (0.028)
# Arrests for Property Crimes	2633	-0.011 (0.01)	-0.02 (0.013)	0.006 (0.017)	0.02 (0.022)
# Arrests for Drug Crimes	2633	0.018 (0.014)	0.024 (0.017)	0.026 (0.018)	-0.014 (0.028)
# Arrests for Other Crimes	2633	-0.005 (0.022)	0 (0.027)	0.053 (0.03)	-0.011 (0.043)
Ever Arrested for Any Crime	2633	-0.01 (0.013)	-0.021 (0.017)	0.025 (0.018)	0.025 (0.025)
# Arrests for Any Crime	2633	-0.014 (0.037)	-0.013 (0.045)	0.091 (0.053)	0 (0.072)

Notes: This table shows the impact of high-dosage tutoring on academic and behavioral outcomes in the first post-randomization school year for study 1 accounting for the 2x2 factorial design. study 2 did not have a second treatment. The first results column shows the impact of high-dosage tutoring when we include a control for being assigned to the Becoming a Man (BAM) treatment group. The next three columns show the full set of interacted treatment effects in the 2x2 design. The first column in this set shows the impact of being assigned to high-dosage tutoring only. The second column shows the impact of being assigned to BAM only. The final column shows the difference in impacts if assigned to both high-dosage tutoring and BAM. All regressions also control for block fixed effects and baseline covariates, including socio-demographics, average pre-randomization 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. Heteroskedasticity robust standard errors in parentheses

Appendix Table 11: Estimated 1 Year Treatment Effects: Pooling Study 1 and 2 - 9th Grade Student Subsample Only

Outcome	N	Control Mean	Intent-to-Treat Estimate	Treatment-on-the-Treated Estimate	Control Complier Mean	FDR q-value
A. Mathematics Outcomes						
CPS Math Test (Study Sample Z)	2735	0.006	0.114 (0.029)	0.290 (0.073)	-0.089	0.001
Math GPA	3019	1.833	0.156 (0.035)	0.384 (0.086)	1.807	0.001
Math Courses Failed (%)	3019	0.164	-0.025 (0.011)	-0.061 (0.026)	0.160	0.021
B. Non-math Academic Outcomes						
CPS Reading Test (Study Sample Z)	2734	0.001	0.015 (0.032)	0.039 (0.080)	-0.111	0.629
Non-Math GPA	3083	1.875	0.057 (0.029)	0.141 (0.072)	1.823	0.150
Non-Math Core Courses Failed (%)	3083	0.165	-0.014 (0.009)	-0.034 (0.021)	0.171	0.162
C. Disciplinary Outcomes						
Disciplinary Incidents	3579	1.441	0.084 (0.099)	0.233 (0.275)	1.417	0.396
Days Absent	3905	21.059	0.587 (0.645)	1.749 (1.921)	21.432	0.396
Out-of-School Suspensions	3579	1.028	0.097 (0.097)	0.270 (0.269)	0.979	0.396
D. Arrest Outcomes						
# Arrests for Violent Crimes	3905	0.092	-0.008 (0.013)	-0.024 (0.038)	0.108	0.612
# Arrests for Property Crimes	3905	0.072	-0.020 (0.012)	-0.061 (0.035)	0.094	0.133
# Arrests for Drug Crimes	3905	0.053	0.005 (0.011)	0.016 (0.031)	0.019	0.612
# Arrests for Other Crimes	3905	0.212	-0.055 (0.020)	-0.165 (0.061)	0.314	0.041
Ever Arrested for Any Crime	3905	0.165	-0.017 (0.010)	-0.052 (0.030)	0.181	0.133
# Arrests for Any Crime	3905	0.429	-0.079 (0.034)	-0.235 (0.101)	0.535	0.063

Notes: This table shows the impact of high-dosage tutoring on academic and behavioral outcomes in the first post-randomization school year pooling all 9th grade students from both studies. Non-math GPA is calculated using grades in all non-math courses in core subject areas (English, Science, Social Science). All regressions control for randomization block fixed effects and baseline covariates, including socio-demographics, average pre-randomization 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, in parentheses.

Appendix Table 12: Estimated 1 Year Treatment Effects: Pooling Study 1 and 2 - 10th Grade Student Subsample Only

Outcome	N	Control Mean	Intent-to-Treat Estimate	Treatment-on-the-Treated Estimate	Control Complier Mean	FDR q-value
A. Mathematics Outcomes						
CPS Math Test (Study Sample Z)	976	-0.002	0.130 (0.050)	0.257 (0.098)	-0.262	0.010
Math GPA	1235	1.749	0.331 (0.052)	0.718 (0.113)	1.477	0.001
Math Courses Failed (%)	1235	0.191	-0.050 (0.017)	-0.109 (0.038)	0.204	0.007
B. Non-math Academic Outcomes						
CPS Reading Test (Study Sample Z)	976	0.009	-0.011 (0.057)	-0.021 (0.112)	-0.098	0.851
Non-Math GPA	1248	1.714	0.086 (0.042)	0.189 (0.093)	1.498	0.115
Non-Math Core Courses Failed (%)	1248	0.209	-0.025 (0.014)	-0.055 (0.031)	0.240	0.115
C. Disciplinary Outcomes						
Disciplinary Incidents	1340	1.780	-0.175 (0.180)	-0.403 (0.418)	2.314	0.997
Days Absent	1371	24.721	-0.005 (1.158)	-0.013 (2.751)	27.363	0.997
Out-of-School Suspensions	1340	1.506	0.077 (0.207)	0.178 (0.478)	1.683	0.997
D. Arrest Outcomes						
# Arrests for Violent Crimes	1371	0.100	-0.027 (0.021)	-0.064 (0.049)	0.143	0.563
# Arrests for Property Crimes	1371	0.053	-0.008 (0.017)	-0.018 (0.041)	0.079	0.796
# Arrests for Drug Crimes	1371	0.055	0.016 (0.017)	0.038 (0.041)	0.018	0.712
# Arrests for Other Crimes	1371	0.161	0.055 (0.035)	0.130 (0.085)	0.012	0.563
Ever Arrested for Any Crime	1371	0.187	0.001 (0.019)	0.002 (0.046)	0.146	0.967
# Arrests for Any Crime	1371	0.370	0.036 (0.054)	0.086 (0.128)	0.252	0.755

Notes: This table shows the impact of high-dosage tutoring on academic and behavioral outcomes in the first post-randomization school year pooling all 10th grade students from both studies. Non-math GPA is calculated using grades in all non-math courses in core subject areas (English, Science, Social Science). All regressions control for randomization block fixed effects and baseline covariates, including socio-demographics, average pre-randomization 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, in parentheses.

Appendix Table 13: Estimated 1 Year Treatment Effects: Pooling Study 1 and 2 - Female Student Subsample Only

Outcome	N	Control Mean	Intent-to-Treat Estimate	Treatment-on-the-Treated Estimate	Control Complier Mean	FDR q-value
A. Mathematics Outcomes						
CPS Math Test (Study Sample Z)	595	-0.054	0.134 (0.059)	0.436 (0.193)	-0.183	0.037
Math GPA	657	2.007	0.172 (0.073)	0.554 (0.236)	1.776	0.037
Math Courses Failed (%)	657	0.117	-0.018 (0.020)	-0.059 (0.063)	0.170	0.349
B. Non-math Academic Outcomes						
CPS Reading Test (Study Sample Z)	596	0.083	0.057 (0.064)	0.185 (0.211)	-0.140	0.570
Non-Math GPA	678	2.163	0.062 (0.059)	0.202 (0.194)	1.921	0.570
Non-Math Core Courses Failed (%)	678	0.109	-0.006 (0.015)	-0.018 (0.051)	0.136	0.717
C. Disciplinary Outcomes						
Disciplinary Incidents	767	1.709	-0.356 (0.232)	-1.227 (0.826)	3.030	0.414
Days Absent	817	23.667	-0.768 (1.372)	-2.745 (4.929)	30.705	0.725
Out-of-School Suspensions	767	0.642	0.053 (0.152)	0.184 (0.521)	0.411	0.725
D. Arrest Outcomes						
# Arrests for Violent Crimes	817	0.065	-0.003 (0.021)	-0.011 (0.075)	0.064	0.884
# Arrests for Property Crimes	817	0.031	-0.014 (0.017)	-0.051 (0.060)	0.083	0.480
# Arrests for Drug Crimes	817	0.006	0.015 (0.010)	0.055 (0.037)	-0.031	0.205
# Arrests for Other Crimes	817	0.108	-0.088 (0.033)	-0.316 (0.122)	0.338	0.057
Ever Arrested for Any Crime	817	0.105	-0.031 (0.020)	-0.111 (0.071)	0.186	0.205
# Arrests for Any Crime	817	0.210	-0.090 (0.046)	-0.323 (0.168)	0.454	0.165

Notes: This table shows the impact of high-dosage tutoring on academic and behavioral outcomes in the first post-randomization school year pooling all female students from both studies. Non-math GPA is calculated using grades in all non-math courses in core subject areas (English, Science, Social Science). All regressions control for randomization block fixed effects and baseline covariates, including socio-demographics, average pre-randomization 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, in parentheses.

Appendix Table 14: Intent-to-Treat (ITT) Estimates: Black and Latinx Subsample, Pooling Both Studies

Outcome	N	Control Mean	Intent-to-Treat Estimate	ITT q-val	Intent to Treat effect x Latinx	ITT x Latinx q-val	ITT Joint Test P-val	ITT Joint Test q-val
A. Mathematics Outcomes								
CPS Math Test (Study Sample Z)	3554	-0.006	0.092 (0.034)	0.010	0.03 (0.050)	0.943	0.001	0.003
Math GPA	4085	1.799	0.198 (0.040)	0.001	-0.004 (0.060)	0.943	0.005	0.008
Math Courses Failed (%)	4085	0.173	-0.030 (0.013)	0.020	-0.003 (0.019)	0.943	0.237	0.237
B. Non-math Academic Outcomes								
CPS Reading Test (Study Sample Z)	3553	-0.007	0.005 (0.038)	0.905	-0.02 (0.055)	0.721	0.824	0.897
Non-Math GPA	4161	1.812	0.099 (0.033)	0.008	-0.063 (0.049)	0.493	0.896	0.897
Non-Math Core Courses Failed (%)	4161	0.179	-0.024 (0.010)	0.029	0.015 (0.015)	0.493	0.345	0.897
C. Disciplinary Outcomes								
Disciplinary Incidents	4757	1.572	-0.008 (0.138)	0.953	0.128 (0.169)	0.882	0.000	0.001
Days Absent	5105	22.298	0.512 (0.811)	0.791	-0.169 (1.130)	0.882	0.200	0.200
Out-of-School Suspensions	4757	1.187	0.157 (0.145)	0.791	-0.049 (0.173)	0.882	0.000	0.001
D. Arrest Outcomes								
# Arrests for Violent Crimes	5105	0.097	-0.009 (0.018)	0.726	-0.01 (0.019)	0.613	0.000	0.002
# Arrests for Property Crimes	5105	0.064	-0.019 (0.015)	0.380	0.011 (0.019)	0.613	0.134	0.201
# Arrests for Drug Crimes	5105	0.056	0.005 (0.015)	0.726	0.011 (0.017)	0.613	0.815	0.979
# Arrests for Other Crimes	5105	0.206	-0.046 (0.028)	0.380	0.051 (0.035)	0.613	0.989	0.989
Ever Arrested for Any Crime	5105	0.174	-0.016 (0.014)	0.380	0.014 (0.017)	0.613	0.001	0.003
# Arrests for Any Crime	5105	0.423	-0.069 (0.046)	0.380	0.063 (0.057)	0.613	0.059	0.118

Notes: This table tests for differences in the impact of high-dosage tutoring between the Black and Hispanic students in our pooled study sample. We interact treatment with an indicator variable for being Hispanic. The ITT coefficient gives the estimated impact on Black students in our sample. The coefficient on the interaction shows the estimated difference in impacts between Hispanic and Black students. We also report the p-value on the null hypothesis that the Black and Hispanic ITT effects are jointly zero. The compliance rate for Black students is 0.35 and the compliance rate for Hispanic students is 0.43.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. We report q-values for the null hypothesis that the Black ITT effect is zero, that the Hispanic-Black difference is zero, and on the joint ITT test. 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 parentheses.

Appendix Table 15: Heterogeneity by Math GPA and CPS Math Test, grouped by Classroom and Teacher

Outcome	N	I. Group by Classroom			II. Group by Teacher		
		Participation	Baseline Heterogeneity	Participation x Baseline Heterogeneity	Participation x Baseline Heterogeneity	Participation	Baseline Heterogeneity
A. End of Year Math GPA							
Average # of Misconducts	4013	0.498 (0.068)	0.015 (0.028)	0.033 (0.080)	0.499 (0.068)	0.056 (0.030)	-0.009 (0.079)
Average # of Out-of-School Suspension Days	4013	0.497 (0.069)	0.020 (0.026)	0.018 (0.093)	0.493 (0.068)	0.060 (0.028)	-0.053 (0.090)
Percentage of Students with Any Misconduct	4013	0.498 (0.068)	0.023 (0.028)	-0.065 (0.070)	0.500 (0.068)	0.071 (0.028)	-0.102 (0.071)
Prior Math GPA Standard Deviation	4010	0.493 (0.069)	-0.010 (0.024)	0.025 (0.075)	0.486 (0.069)	-0.026 (0.025)	0.116 (0.082)
Prior Math GPA 75-25th Percentile Distance	4013	0.496 (0.068)	-0.044 (0.023)	0.022 (0.067)	0.495 (0.068)	-0.052 (0.025)	0.147 (0.066)
Prior Math GPA 90-10th Percentile Distance	4013	0.497 (0.069)	-0.059 (0.025)	-0.023 (0.072)	0.496 (0.068)	-0.063 (0.024)	0.009 (0.069)
B. End of Year CPS Math Test (Z)							
Average # of Misconducts	3364	0.261 (0.058)	-0.033 (0.021)	-0.090 (0.053)	0.259 (0.058)	-0.058 (0.023)	-0.082 (0.055)
Average # of Out-of-School Suspension Days	3364	0.260 (0.057)	-0.017 (0.018)	-0.057 (0.052)	0.262 (0.058)	-0.039 (0.022)	-0.091 (0.055)
Percentage of Students with Any Misconduct	3364	0.267 (0.058)	-0.040 (0.024)	-0.120 (0.054)	0.264 (0.058)	-0.079 (0.024)	-0.113 (0.055)
Prior Math Score Standard Deviation	3362	0.252 (0.058)	0.014 (0.021)	0.040 (0.055)	0.252 (0.059)	0.002 (0.021)	0.036 (0.055)
Prior Math Score 75-25th Percentile Distance	3363	0.256 (0.057)	-0.006 (0.019)	0.120 (0.051)	0.255 (0.058)	0.014 (0.020)	0.051 (0.051)
Prior Math Score 90-10th Percentile Distance	3363	0.255 (0.058)	0.011 (0.022)	0.056 (0.058)	0.254 (0.058)	-0.025 (0.021)	0.060 (0.056)

Notes: This table shows how the impact of high-dosage tutoring on Math GPA (Panel A) and Math Test Scores (Panel B) in the first post-randomization school year varies with different dimensions of classroom heterogeneity. Each row shows heterogeneity using the baseline characteristic reported in the first column. The first three result columns group students by their math classroom and the final three result columns group students by their math teacher. All regressions are based on the TOT specification with baseline heterogeneity measure and the interactions between participation and treatment with the baseline heterogeneity measure added to the regression. All regressions also control for block fixed effects and baseline covariates, including socio-demographics, average pre-randomization 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. Heteroskedasticity robust standard errors, clustered by individual, in parentheses.

Appendix Table 16: Study 1 Sample - Estimated Effects of High-Dosage Tutoring on Outcomes from ISR Survey - End of First Program Year

Outcome	N	Control Mean	Intent-to-Treat Estimate	FDR q-value
A. Indices				
Adult Supports	623	-0.018	0.019 (0.067)	0.994
Grit	624	-0.041	0.011 (0.047)	0.994
Conscientiousness	624	-0.036	0.045 (0.059)	0.994
Locus of Control	624	-0.042	0.000 (0.050)	0.994
Social Networks	623	-0.013	-0.020 (0.045)	0.994
B. Adult Supports				
Number of adults to talk to (No Change)	622	4.297	0.016 (0.340)	0.962
Number of adults who care (No Change)	623	7.384	0.270 (0.589)	0.962
Would talk to adults at school (Dummy)	623	0.375	-0.010 (0.043)	0.962
C. Grit				
Agree: Setbacks don't discourage me (Z)	623	0.000	0.011 (0.087)	0.959
Agree: I am a hard worker (Z)	624	0.000	0.091 (0.085)	0.857
Disagree: I have difficulty maintaining focus (Z)	623	0.000	-0.088 (0.082)	0.857
Agree: I am diligent (Z)	624	0.000	0.021 (0.089)	0.959
Agree: I finish what I begin (Z)	624	0.000	-0.033 (0.087)	0.959
Agree: I can continue until everything is perfect (Z)	624	0.000	0.005 (0.087)	0.959
D. Conscientiousness				
Agree: I am always prepared (Z)	624	0.000	0.125 (0.090)	0.488
Agree: I continue until everything is perfect (Z)	624	0.000	0.005 (0.087)	0.990
Agree: I leave a mess in my room (Z)	624	0.000	0.001 (0.087)	0.990
E. Locus of Control				
Agree: I have control over direction of life (Z)	621	0.000	0.028 (0.087)	0.744
Disagree: Every time I try to get ahead, something or somebody stops me (Z)	624	0.000	0.053 (0.088)	0.744
Disagree: Luck is more important than hard work (Z)	624	0.000	0.135 (0.086)	0.285
Disagree: My plans never work out, planning makes me unhappy (Z)	622	0.000	0.041 (0.092)	0.744
Agree: I can make plans work (Z)	623	0.000	-0.215 (0.083)	0.051
F. Social Networks				
Reports No Close Friends (Dummy)	623	0.025	-0.011 (0.014)	0.752
Friends think it is important to attend classes regularly (Z)	607	0.000	-0.078 (0.092)	0.752
Friends think it is important to get good grades (Z)	607	0.000	-0.040 (0.082)	0.883
Friends think it is important to study (Z)	607	0.000	-0.250 (0.091)	0.046
Friends think it is important to continue education to college (Z)	607	0.000	0.020 (0.085)	0.948
Have stopped hanging around with someone (Recoded Dummy)	623	0.505	0.065 (0.045)	0.508
Have started hanging around with someone (Recoded Dummy)	622	0.616	-0.001 (0.044)	0.989

Notes: All items are coded so the desired effect direction is positive. Baseline covariates and randomization block fixed effects included in all models (see text). Heteroskedasticity-robust standard errors in parentheses. 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.

Appendix Table 17: Saga Program Costs

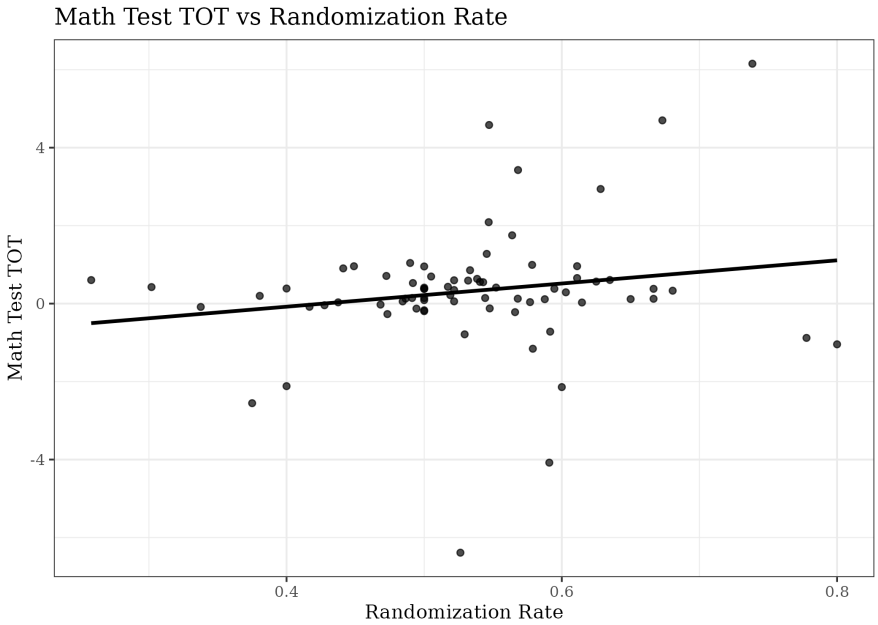
Input	Planning Year	Year One	Year Two	Total
A. Variable Costs				
Tutor stipends and transportation benefits	$0	$901,000	$1,445,000	$2,346,000
Tutor fringe benefits	$0	$265,795	$427,720	$693,515
Recruitment	$212,000	$269,500	$0	$481,500
Tutor Training	$0	$27,895	$49,399	$77,294
Supplies	$0	$53,000	$91,176	$144,176
Program Management	$20,600	$463,500	$842,358	$1,326,458
Administration and Back Office	$35,000	$120,000	$140,000	$295,000
B. Fixed Costs				
Curriculum Development	$75,000	$150,000	$175,000	$400,000
Data and instructional support	$25,750	$221,450	$315,000	$562,200
Communications/PR	$15,000	$35,000	$35,000	$85,000
Travel	$20,000	$35,000	$67,500	$122,500
Miscellaneous	$25,000	$40,000	$60,000	$125,000
C. Total Costs				
Total Cost	$428,350	$2,582,140	$3,648,153	$6,658,643
Total Variable Cost	$267,600	$2,100,690	$2,995,653	$5,363,943
Total Fixed Cost	$160,750	$481,450	$652,500	$1,294,700

Notes: This table shows details of Saga's planned budget over the study.

Appendix Table 18: Average Total and Variable Costs of Program

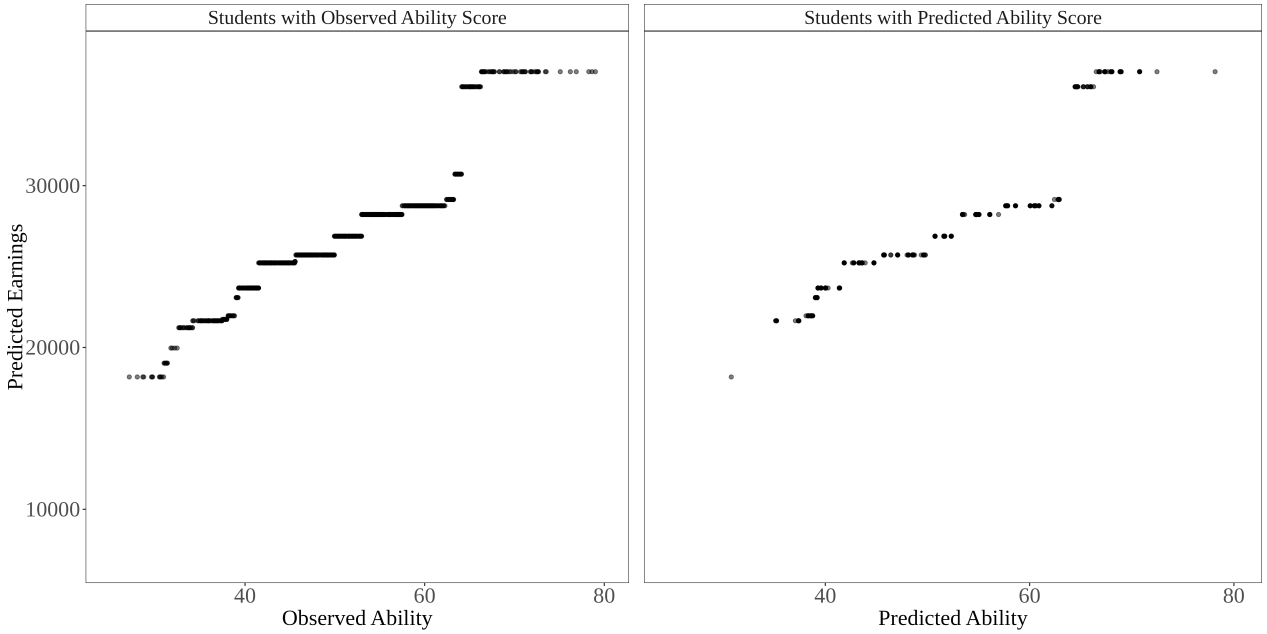
	Year One	Year Two	Total
A. Program Size			
Student Capacity	670	1130	1800
Participants	534	862	1396
Tutors	53	85	138
Students/Tutor	12.6	13.3	13
Schools	12	15	15
B. Costs			
Total Cost	$2,582,140	$3,648,153	$6,658,643
Total Variable Cost	$2,100,690	$2,995,653	$5,363,943
C. Average Total Cost			
Per Treatment Slot	$3,853.94	$3,228.45	$3,699.25
Per Participant	$4,835.47	$4,232.20	$4,769.80
D. Average Variable Cost			
Per Treatment Slot	$3,135.36	$2,651.02	$2,979.97
Per Participant	$3,933.88	$3,475.24	$3,842.37

Notes: This table shows how we calculate average program costs. Panel A summarizes program size in each year. Panel B reports the total costs implied by Appendix Table 20. Panels C and D use the information in the first two panels to calculate average total and variable costs per treatment slot and per participant.



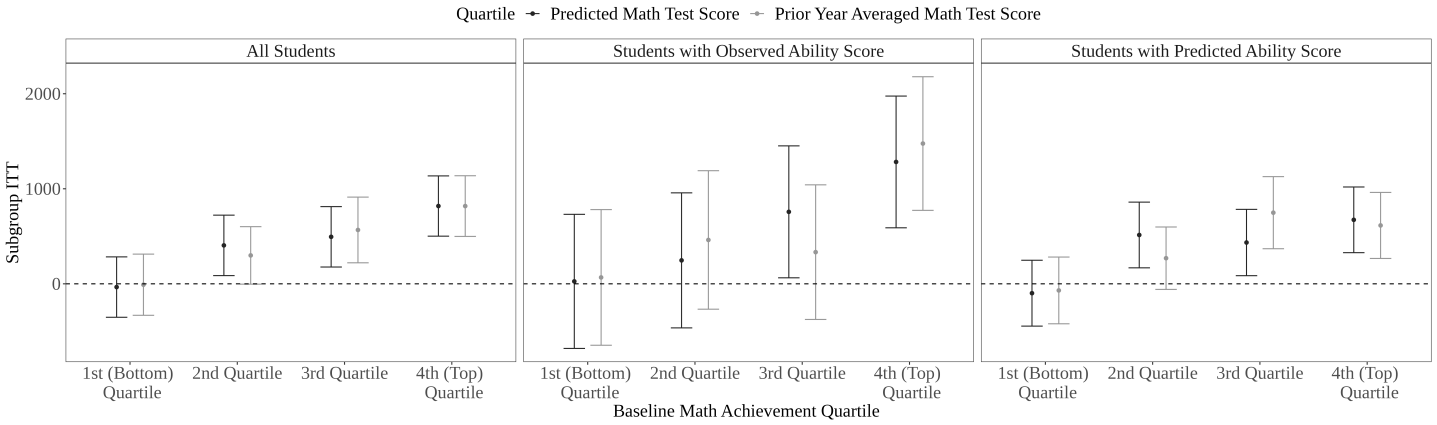
Appendix Figure 1: SUTVA Analysis: Block-Level Randomization Rate Plotted Against TOT Effect on Math Test Score

Notes: Figure plots randomization-block-specific TOT effects against block-specific treatment assignment rates. The results indicate that effects increase with a larger share of individuals within a block randomized to treatment. The coefficient on the randomization rate is 2.973 with a standard error of 1.943. This is inconsistent with what we would expect to see if treatment spillovers are attenuating our estimates.



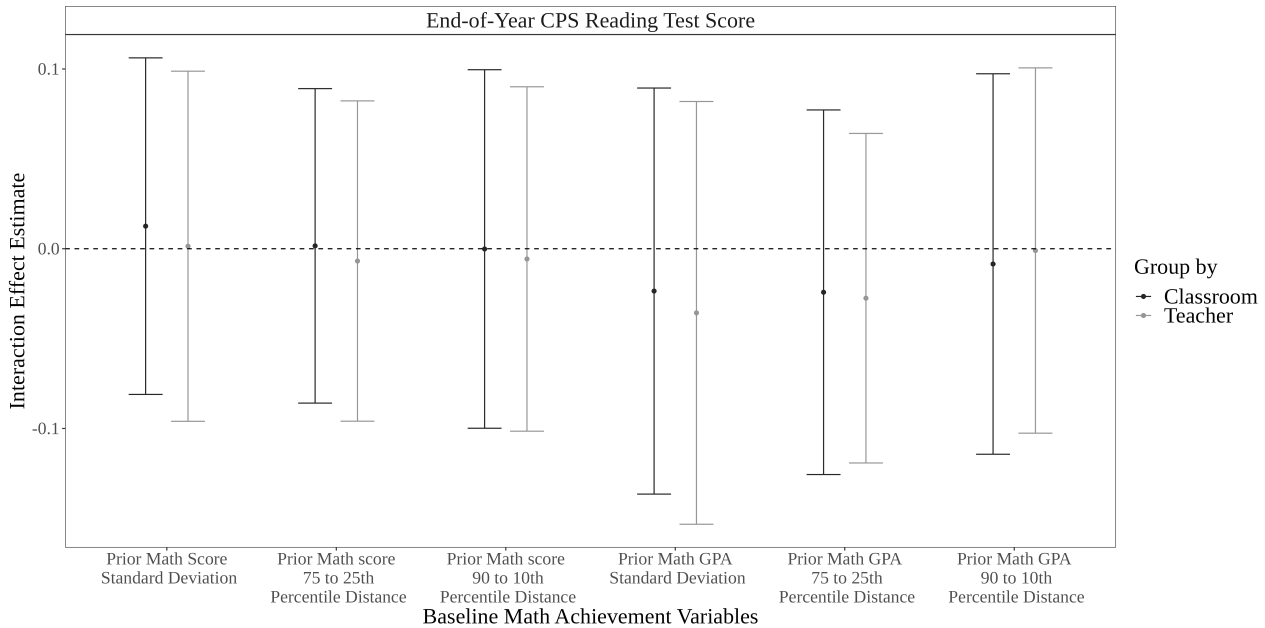
Appendix Figure 2: Mapping Ability Scores to Earnings

Notes: Figure shows the mapping between ability scores on the research team-administered math test and earnings. We use the NELS:88 dataset to flexibly estimate the relationship between a student's performance, as measured by their "ability score", and future earnings, and then use that mapping to compute estimated future incomes for the students in our sample who take the same assessment. The left panel shows estimates for students where we directly observe their ability scores. The right panel shows that the estimates for students with predicted ability scores look very similar to the estimates for students with observed ability scores.



Appendix Figure 3: Floor Effects with Earnings

Notes: Figure shows the effects of high-dosage tutoring on math GPA (left panel) and CPS-administered math test score (right panel) on predicted adult earnings 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 Appendix III). To predict earnings, we use the NELS:88 dataset to flexibly estimate the relationship between a student's performance, as measured by their "ability score", and future earnings, and then use that mapping to compute estimated future incomes for the students in our sample who take the same assessment. For students who did not take the research-team administered test score, we predict ability scores using performance on the CPS-administered math test. 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% confidence intervals.



Appendix Figure 4: Heterogeneity in Reading Impacts by Baseline Classroom Math Achievement

Notes: Figure shows the coefficient on the interaction between treatment assignment and different measures of heterogeneity in classroom reading 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% confidence intervals. The CPS data on classroom assignments for students are noisy for assigning students to a specific classroom or "section", but we believe is 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).