NBER WORKING PAPER SERIES

CAN TECHNOLOGY FACILITATE SCALE? EVIDENCE FROM A RANDOMIZED EVALUATION OF HIGH DOSAGE TUTORING

Monica P. Bhatt
Jonathan Guryan
Salman A. Khan
Michael LaForest-Tucker
Bhavya Mishra

Working Paper 32510
http://www.nber.org/papers/w32510

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2024

This paper was made possible by the generous support of the AbbVie Foundation, Arnold Ventures, Griffin Catalyst, Overdeck Family Foundation, and the UChicago Crime Lab and Education Lab Investors' Council. For vital assistance in making this work possible, we thank Roseanna Ander, Brenda Benitez, Trayvon Braxton, Cathryn Cook, Ellen Dunn, Chris Dupuis, Jaureese Gaines, Antonio Gutierrez, Zach Honoroff, Julia Imperatore, Daniel Lopez, Sibella Matthews, Jacob Miller, Julia Quinn, Natalee Rivera, Alan Safran, Maitreyi Sistla, John Wolf, as well as the staffs of the Chicago Public Schools system, New York City Department of Education, and Saga Education. Thanks to Jeffrey Broom, Sarah Dickson, Kylie Klein, Jared Sell, and The Research & Policy Support Group at New York City Public Schools for their help in accessing the data we analyze here, and to Emily Gell, Cristobal Pinto, Catherine Schwarz, Anna Solow-Collins, and Erin Wright for their invaluable contributions to the data analysis. For useful suggestions we thank conference and seminar participants at SREE, APPAM, the Hoover Institution, and the University of Chicago Committee on Education, as well as Jonathan Davis, Max Kapustin, Jens Ludwig, Matteo Magnaricotte, and Greg Stoddard. This study was approved by the University of Chicago's committee on human subjects as IRB18-0574 on May 7, 2018. This RCT was registered on Open Science Framework registry for randomized control trials under trial DOI 10.17605/OSF.IO/UW8EH. All opinions and any errors are those of the authors and do not necessarily represent the views of the any partner or funder. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Can Technology Facilitate Scale? Evidence from a Randomized Evaluation of High Dosage Tutoring

Monica P. Bhatt, Jonathan Guryan, Salman A. Khan, Michael LaForest-Tucker, and Bhavya Mishra

NBER Working Paper No. 32510
May 2024
JEL No. I21,J24

ABSTRACT

High-dosage tutoring is an effective way to improve student learning (Nickow et al., 2024; Guryan et al., 2023). Finding ways to deliver high-dosage tutoring at large scale remains a challenge. Two primary challenges to scaling are cost and staffing. One possible solution is to reduce costs by substituting some tutor time with computer-assisted learning (CAL) technology. The question is: Does doing so compromise effectiveness? This paper provides evidence from a randomized controlled trial (RCT) of approximately 4,000 students in two large school districts in 2018- 19 and 2019-20. The RCT tested the effectiveness of an in-school math tutoring program where students worked in groups of four, with two students working with an in-person tutor while the other two worked on CAL, alternating every other day. The tutoring model had per-pupil costs approximately 30 percent lower than the 2- to-1 tutoring model studied in Guryan et al. (2023). We find gains in students' math standardized test scores of 0.23 standard deviations for participating students, which are almost as large as the effect sizes of the 2-to-1 tutoring model reported in Guryan et al. (2023). These findings suggest strategic use of technology may be a way to increase the scalability of HDT.

Monica P. Bhatt
University of Chicago
Crime and Education Labs
33 North LaSalle Street
Suite 1600
Chicago, IL 60602
mbhatt@uchicago.edu

Jonathan Guryan
Northwestern University
Institute for Policy Research
2040 Sheridan Road
Evanston, IL 60208
and NBER
j-guryan@northwestern.edu

Salman A. Khan
University of Chicago
Crime and Education Labs
33 North LaSalle Street
Suite 1600
Chicago, IL 60602
salmankhan@uchicago.edu

Michael LaForest-Tucker
U.S. Air Force Academy
michael.laforest@afacademy.af.edu

Bhavya Mishra
University of Chicago
Crime and Education Labs
33 North LaSalle Street
Suite 1600
Chicago, IL 60602
United States of America
bmishra@uchicago.edu

A data appendix is available at
http://www.nber.org/data-appendix/w32510

A randomized controlled trials registry entry is available at
https://osf.io/uw8eh/

1 Introduction

Improving student learning in America is not so much a problem of pedagogy as a problem of scale. For example, we have evidence that it is possible to realize large gains on student test scores in some charter schools (Angrist et al., 2012; Fryer, 2012; Angrist et al., 2013). But it has not borne out that practices used by these schools - including high dosage tutoring- can be scaled and retain their efficacy. As Cullen et al. (2013) argue, "we are skeptical that these achievements can be generalized on a large scale. More likely, attempts to do so would be extremely costly and largely ineffective." Much of the social science literature focuses on similar examples where success is found in pockets but is not broadly replicable.

In this paper we consider whether the development of educational technology and high-quality computer assisted learning (CAL) platforms might change this scale-up calculus. We focus on the strategic incorporation of technology into high dosage tutoring (HDT), a specific pedagogical strategy that the U.S. Secretary of Education has encouraged school districts to implement as part of efforts to overcome pandemic-induced learning loss. Tutoring delivered in small groups during the school day at substantial dosage with a structured curriculum can substantially improve student learning 1) by maximizing time on task relative to regular classroom instruction and 2) by addressing the problem of 'academic mismatch'- the idea that what the classroom teacher teaches may be above the level of what students who are behind grade level need (Nickow et al., 2024; Davis et al., 2017). As Bloom (1984) noted, small-group tutoring represents the "best learning conditions we can devise." While shown to be effective and cost-effective, this approach is largely cost-prohibitive from the perspective of the public sector. A key challenge to

scaling, therefore, is cost.

One important advance in solving this scale-up challenge comes from an insight by the non-profit Saga Education: that small-group tutoring and classroom teaching are fundamentally different tasks. Teachers consistently say the two hardest parts of teaching are classroom management and individualizing instruction.[1] For that reason, school districts commonly require substantial prior pedagogical training for classroom teachers. Even still, the steep gains in teacher effectiveness in their first few years of teaching suggest substantial additional on-the-job learning (Rockoff, 2004; Clotfelter et al., 2010; Henry et al., 2011). The hypothesis underlying tutoring is that if class size shrinks enough, the nature of teaching qualitatively changes. As the task of delivering instruction is simplified, people without much prior pedagogical training can be effective tutors much more quickly, which allows schools to deploy new staffing models, such as hiring recent college graduates or mid-career switchers to tutor for a year for a modest stipend as a public service. Two large-scale randomized controlled trials (RCTs) from Chicago confirm this hypothesis: For a cost of about $\$3{,}500$ per student per year, 2-to-1 tutoring for an hour a day, every day, showed double or triple the amount of math high school students learn per year(Guryan et al., 2023). Yet even with this innovative school staffing model, cost and the labor shortages we see on the heels of the pandemic both remain key barriers to scaling high dosage tutoring to every student who would benefit from it.[2]

[1]For example, in the National Teacher and Principal Survey (NTPS) for 2020-21, 58% of new elementary school teachers and 34% of new secondary school teachers say they felt not at all prepared to deal with classroom management; 55% of new elementary school teachers and 35% of new secondary school say they felt not at all prepared to differentiate instruction (from tabulations of NTPS data).

[2]For example, Congress allocated $189.5 billion to school districts as a part of the Elementary and Secondary School Emergency Relief Fund (ESSER). Even if every dollar were spent on evidence-based high-dosage tutoring, districts would only be able to serve a fraction of students who might benefit. In 2021 the Chicago Public Schools announced $50 million to hire over 500 tutors to implement an in-school tutoring

The present paper tests a second scale-up insight by Saga Education: That strategic use of technology might reduce cost and labor requirements without compromising effectiveness. High quality computer assisted learning (CAL) programs, like tutors, have the potential to both maximize time-on-task and individualize instruction to overcome academic mismatch (Escueta et al., 2017; Heffernan and Heffernan, 2014). What tutoring can do that CAL cannot is provide human connection. This human connection between the tutor and the student might help to sustain student engagement and motivation. It follows that the absence of a human connection may be one reason why there seems to be diminishing marginal returns to student time spent on CAL (Bettinger et al., 2023). During the COVID-19 pandemic, for example, it was made apparent that relying exclusively on technology as a mode of instruction can lead to student disengagement (Limone and Toto, 2021; Dorn et al., 2020; Akpınar et al., 2021; Drane et al., 2021; Huffman, 2020). Saga Education's insight is that diminishing marginal returns to time on CAL implies that not only is there a flat part of the social-returns curve, but a steep part as well.[3]

In this paper we present the results of a large-scale RCT that tests the effects of a 4-to-1 tutoring model in which four 9th grade students sit at a table with one in-person tutor and the students alternate days working either with the tutor in student pairs or working on CAL for the entirety of a class period (50 minutes). By way of comparison the tutoring model studied by Guryan et al. (2023) had students spend one 50-minute class period per day with a tutor at a 2-to-1 ratio every day. Saga delivered the 4-to-1 tutoring model (what we call the 'Saga Technology' model in what follows) as part of two large-scale

program, as part of an initiative it called the Chicago Tutor Corps. Despite this large investment, at a cost of thirty five hundred dollars per student the district would be able to serve less than ten percent of its total enrollment.

[3]See also Guryan et al. (2023); Nickow et al. (2020); Ritter et al. (2007); Escueta et al. (2017).

randomized controlled trials (RCTs) during the 2018-19 and 2019-20 school years in three public high schools in Chicago and another four high schools in New York City. Given the unexpected disruptions of the global COVID-19 pandemic that started in early 2020, we focus here primarily on the results from the 2018-19 academic year (AY2018-19).

Our pre-specified primary outcome is end-of-year standardized math achievement test scores. We also report estimated effects on other academic outcomes (e.g. GPA, course failures, and attendance). And, we assess the persistence of learning impacts by measuring treatment effects during the subsequent school year, after the tutoring intervention had ended.

For the first cohort (AY2018-19), we estimate the Saga Technology tutoring model had intent-to-treat (ITT) effects on standardized math test scores of 0.19 standard deviations (SD) and treatment-on-the-treated (TOT) effects of 0.23 SD. This gain in students' math standardized test scores is equivalent to between three-quarters and one full year of additional learning over the course of the program year (Reardon, 2011; Bloom et al., 2008). Among studies of educational interventions for secondary-school students, impacts of this magnitude are not common.

Perhaps even more remarkable is that the TOT effects from alternating days on tutoring versus CAL, 0.23 SD, are almost as large as the TOT effect of every-day tutoring—also with 9th graders in Chicago, also delivered by Saga Education within the same Chicago Public Schools system—reported by Guryan et al. (2023). In other words, relative to every-day tutoring, incorporating technology reduces costs by 30%, reduces the number of tutors required to serve a given number of students by 50%, and yet has effects on student learning almost as large. The two estimates are statistically indistinguishable from

one another when using a simple pairwise t-test.

Some indication that tutoring does not boost test scores merely by 'teaching to the test' comes from signs of beneficial impacts on other learning outcomes. For example, in the first cohort (AY2018-19) we find that participating in tutoring caused a 0.24 point increase in math course GPA (a 13% increase relative to the control complier mean) and a 22% decrease in math course failures. We do not find any detectable effect on non-math subjects or behavioral outcomes such as attendance and school suspensions.[4]

Some indication that these results are not a statistical fluke comes from signs that the first cohort (AY2018-19) results seem to largely replicate with the independent sample of students served in AY2019-20, at least for those outcomes we can consistently measure across cohorts despite the disruption of the COVID-19 pandemic in early 2020. Because the COVID-19 pandemic led both Chicago and New York to close schools, neither district administered end-of-year standardized tests. However, for both cohorts we are able to estimate effects on mid-year grades, which for the second cohort were issued before the school closures in March 2020. The mid-year effects are qualitatively similar for both cohorts in terms of both math GPA (0.20 versus 0.21 grade points for the TOT) and math course failures (-0.07 versus -0.05 for the TOT).

We also look at whether effects for the first cohort persisted into the following school year, recognizing this follow-up year was disrupted by the pandemic. We find that participating in tutoring during AY2018-19 generates positive and significant effects on mid-year math GPA during the following school year of 0.15 grade points (TOT) and a

[4]Similar to what we find here for the Saga Technology model, Guryan et al. (2023) did not find any significant effects of the traditional Saga tutoring model on behavioral outcomes such as attendance and school suspensions. However, they did find significant spillovers on some outcomes in non-math subjects other than reading which we do not find in this study.

20% reduction in first semester math course failures.[5]

The disruptions in schooling and unprecedented declines in student learning caused by the COVID-19 pandemic make the results of this study especially noteworthy. According to the results of the 2022 National Assessment of Educational Progress (NAEP), average math scores of fourth and eighth-grade students registered the largest decline seen since initial assessments in 1990 (LaFave et al., 2022). Test score declines have been broadly based but are most pronounced for the most disadvantaged children. The U.S. Secretary of Education encouraged districts all around the country to prioritize federal pandemic relief funds for tutoring to help overcome pandemic learning loss. Even with this substantial increase in federal funding, school districts around the U.S. have not been able to provide high-dosage tutoring to help every student adversely affected by the pandemic (which is to say, almost every student). The results presented here are consistent with the idea that at current levels CAL can be substituted on the margin for some tutor time to substantially lower the marginal costs of expanding the benefits of tutoring.

2 Conceptual Framework

The hypothesis behind the technology-scaling proposal here is illustrated by Figure 1a. The figure shows the relationship between student learning on the y-axis and student time spent on a computer assisted learning platform (CAL) on the x-axis. In the figure, Q_0 represents the baseline amount of time students spend on CAL in school, and T represents

[5]We do not see significant post-intervention year effects for cohort 2, which received treatment in AY2019-20. This is perhaps not surprising since cohort 2 received roughly half a year of tutoring before schools were closed due to the COVID-19 pandemic.

the increase in time spent on CAL as a result of the incorporation of technology into tutoring. Consistent with the findings in Bettinger et al. (2023) the relationship is assumed to be concave, i.e. to exhibit diminishing returns. The figure illustrates the possibility that at some point there are not just diminishing marginal positive returns, but even negative returns. Negative returns to time using CAL might occur if additional time on CAL crowds out time spent on more productive learning activities (Angrist and Lavy, 2002; Malamud and Pop-Eleches, 2011) or if time spent on CAL actively discourages students from engaging in school. There is evidence of this type of disengagement from remote schooling during the COVID pandemic – student chronic absenteeism rose dramatically, and many parents opted for in-person private instruction instead of remote public instruction (García and Weiss, 2020; Dee, 2023, 2024).

Figure 1a shows the best-case scenario for incorporating CAL into tutoring: The baseline level of CAL use in schools Q_0 is in the range where additional time on CAL yields positive learning effects, and the incremental CAL time in the tutoring program keeps students in the range where returns remain positive. In this case, substitution of technology for tutoring time offers an opportunity to improve student learning at large scale, without compromising implementation fidelity (since software can operate identically over and over), at low (perhaps even zero) marginal cost.

But it is also possible that adding technology into tutoring might not help– and might even harm– learning. Figure 1b illustrates one reason why: Even if the baseline level of CAL use in a school (Q_0) is in the range of positive returns, it is possible that infusion of CAL into tutoring adds too much time on the computer, so that marginal returns become negative and the level of learning at $Q_0 + T$ is no higher, and possibly lower, than at Q_0.

Another possibility is that schools are already using so much technology as part of the student's regular school day that Q_0 already lies on the negative returns part of the schedule, so that even judicious use of CAL as part of tutoring winds up inadvertently reducing learning, as shown in Figure 1c. Finally, it could be that schools are using just a modest amount of technology in school at baseline and the tutoring program incorporates a modest amount of CAL time, but students have very limited tolerance for time on CAL—that is, the relationship between time on CAL and student learning very quickly exhibits negative marginal returns, as shown in Figure 1d.

Our research design, discussed further below, essentially compares student learning at status quo school operations, including whatever baseline level of CAL schools use, Q_0, to a version of tutoring that incorporates time on CAL, $Q_0 + T$. In principle one could trace out the relationship between time on CAL and student learning shown in Figures 1a-1d with multiple treatment arms that randomized students to different levels of time on CAL. But given the single-treatment-arm design we have in the current project, the best we will be able to do is distinguish between whether we are in the scenario shown in Figure 1a or in one of the scenarios shown in Figures 1b-1d, without being able to distinguish among them.

Finally, we recognize that different students might have different tolerances for CAL; that is, the shape of the relationship between CAL time and learning might be different for different students. If there is heterogeneity in this regard, our current study essentially estimates movement along the average of these schedules across students.[6]

[6]Our research team has a separate project underway that is working to collect far larger samples to allow us to use new methods from machine learning to estimate heterogeneous treatment effects that essentially empirically recover separate CAL time / student learning schedules for different student 'types.' `https://educationlab.uchicago.edu/projects/personalized-learning-initiative/`

3 The Intervention: Integrating Computer Assisted Learning with High-Dosage Tutoring

In this paper, we measure the effect of a tutoring model we refer to as Saga Technology. This tutoring model builds on the high-dosage tutoring model developed by Saga Education that was evaluated in Guryan et al. (2023). We briefly describe the traditional Saga high-dosage tutoring model, and then explain how the Saga Technology model selectively incorporates technology to increase scalability.

In the 2-to-1 model, 9th and 10th grade students attended a tutoring class, called "Math Lab," that was included in their regular schedule each day. During Math Lab, each tutor met with a pair of students (for a 2-to-1 student-to-tutor ratio) in a classroom that had approximately 10 tutors and 20 students. Students generally worked with the same tutor throughout the school year, and tutoring sessions took place every day, during the school day, throughout the school year. All students attended a regular math class alongside the tutoring sessions.

Tutoring sessions covered a mix of topics spanning from earlier grade-level material the student had not yet mastered up to grade-level material that was being covered in the student's regular math class. Each Saga tutoring session followed a similar structure: about five minutes of warm-up problems, 40 minutes of individualized tutoring, and a few minutes at the end of the period to review the day's topics. Frequent assessments were given to students and were used to tailor the math topics and levels to each student's current needs and knowledge. Tutors used a structured curriculum that was designed by Saga. Tutors were typically recent college graduates, and Saga screened for strong math

skills and strong interpersonal skills during the hiring process.[7]

The Saga Technology tutoring model studied in this paper built on the structure of the traditional Saga tutoring model. In an effort to reduce the per-student cost, students met each day in groups of four with a tutor. On any given day, two of the students worked with the tutor in a manner that was similar to the daily sessions in the traditional Saga model, while the other two students worked on a CAL platform called ALEKS. The students alternated which days they spent working with the tutor versus working on the CAL platform. On alternate weeks each student would receive two days of tutoring and three days of CAL, or three days of tutoring and two days of CAL.

ALEKS is an educational technology platform that provides individualized instruction and adaptive questioning differentiated for each student's needs. The ALEKS CAL platform allows each student to work on math topics that are tailored to the students' assessed skill level and knowledge base. ALEKS uses artificial intelligence based on each student's performance to assign problems and to give hints as students attempt to solve problems. ALEKS was initially developed through a grant from the National Science Foundation (Canfield, 2001). A recent meta-analysis suggests that compared to regular classroom instruction, time on ALEKS boosts standardized test scores by 0.08 SD on average, comparable to what is seen from other online intelligent tutoring systems (Fang et al., 2019). Consistent with the idea that there might be diminishing marginal returns (or even negative returns) to time on CAL, the meta-analysis finds larger gains in studies where students spent relatively less time on ALEKS.

For the current study, Saga Education hired and trained a total of 72 tutors across

[7]For more details, please see Guryan et al. (2023).

both study years. Tutors were supervised by a site director at each of the seven study schools across Chicago and NYC. During the study period, Saga Education also hired two special program directors to oversee the implementation of the educational technology component. Saga Education tutors again mirrored those recruited in the 2-to-1 model- recent college graduates, strong math and interpersonal skills, and were willing to devote one year to public service.[8]

Tutors received teaching stipends, as well as benefits, during the nine-month academic year. Prior to the start of the school year, tutors received around 100 hours of training. During the school year, site directors observed tutors on a daily basis and provided regular feedback. Each tutor taught for six periods and had a caseload of approximately 24 students.

In both Chicago and New York City, the evaluations were structured as two-arm randomized studies, with some minor differences we describe below. In both cities randomization was at the student level, and the two arms of the study were: (1) a treatment group that was offered the chance to participate in the Saga Technology tutoring program; and (2) a control group that received status quo classes and services. Treatment group students who participated were enrolled in a course that was part of their daily schedules. The tutoring course typically took the place of an elective class such as foreign languages or study hall. In both cities, the research team received rosters from the individual study schools in the summer prior to the start of the academic year and then randomly assigned eligible students to one of the two conditions. Staff members at each school then used these random assignments to schedule students for the coming year. Note that, in New York

[8]Currently, Saga Education receives subsidies for some of the tutoring positions from AmeriCorps. During the study years, Saga Education had AmeriCorps members starting in the fall of 2019 (cohort 2).

City, NYC DOE allowed principals discretion in whether they followed the provided lists for program assignments, though school leadership largely chose to follow the suggested assignments.[9]

More details on the randomization process in Chicago and New York City are provided below.

4 Data, Randomization, and Baseline Balance

4.1 Data

Most of the data for this study are drawn from student-level administrative records from the Chicago Public Schools (CPS) and the New York City Department of Education (NYC DOE). These sources include test scores and academic outcomes measured prior to randomization and post-treatment, as well as student demographics. We supplement these data sources with records of tutoring participation that were input by Saga tutors and site directors, and with a series of observations of a sample of tutoring sessions carried out by the research team. We also analyze data collected by ALEKS that allow us to document student usage of the CAL platform.

The primary pre-specified post-treatment outcomes are standardized math test scores from tests taken at the end of the school year. For Chicago, we use the PSAT math test, which was required for students at the end of 9th and 10th grades. For New York City, we use the New York State Regents Examination math tests, which are statewide examinations

[9] As described in the pre-analysis plan, NYC DOE allowed principals discretion in whether they followed the provided lists for program assignments. In this study, school leadership largely chose to follow the suggested assignments.

given in core high school subjects, and are used in part to determine graduation eligibility. We also measure effects on standardized reading scores, mid-year and end-of-year course grades in math and non-math courses, the fraction of math and non-math courses failed, days absent from school, number of days suspended, and number of in-school disciplinary incidents.

We include baseline controls that are also drawn from the CPS and NYC DOE administrative records. These baseline controls include student demographic characteristics like age, race/ethnicity, English language learner status, diverse learner status, and free or reduced-price lunch eligibility, a commonly used measure of low family income. For baseline test scores in Chicago, we use math and reading scores from the NWEA MAP, a nationally normed test that is used in many school districts around the country and which is administered to CPS 3rd-8th graders. In New York City, we use math and English language arts (ELA) scores from New York State tests given to students in grades 3-8 each spring. [10]

4.2 Sample Selection and Randomization

Prior to the 2018-19 school year, the study team invited two schools in Chicago and four schools in NYC to participate in the study. Before the 2019-20 school year one additional school was added to the study in Chicago, bringing the total number of study schools to seven. Schools were selected with the goal of having a study sample with similar demographics and family income to the samples for previous studies of Saga tutoring, which included more than 90% students of color and more than 85% students who qualify

[10] We use the tests given at the end of the school year in AY2017-2018 and AY2018-2019.

for free and reduced-price lunch.

Prior to randomization, we asked schools to share names of entering, first-time 9th grade students who had a high probability of fitting Saga into their schedules. Though the factors influencing whether students could fit tutoring into their schedule varied by school, students in specialized programs, such as International Baccalaureate (IB), were typically excluded. Some schools chose to exclude high-achieving students, or students who were already participating in another specialized program.[11]

In an effort to focus the study on students who were most likely to participate and benefit, we collaborated with Saga Education, CPS, and NYC DOE to set eligibility criteria for the tutoring program. The resulting criteria excluded students who had failed 75% or more of their classes in the previous school year, missed more than 60% of their enrolled school days in the previous school year, and students who with significant disabilities that the tutoring program was not designed to support.[12]

Once the study sample was identified using the eligibility criteria above, we stratified eligible students by school and gender, and randomized 9th-grade students to treatment or control within those strata.[13] The number of students offered the chance to participate was approximately 15 percent more than the number of tutoring spots in the school. In cases where take-up was low enough that some tutoring spots remained unfilled, or in cases where students dropped out of the program and spots opened up, we conducted additional

[11]For further information on the criteria specific to each school, please refer to the pre-analysis plan: https://osf.io/uw8eh/.

[12]Students with the following disabilities were also omitted from the study and receipt of Saga Tech tutoring services at the request of the school and program provider: autism, "educable mentally handicapped," traumatic brain injury, speech/language disorders, and emotional/behavioral disorders.

[13]We further stratify by academy (business, law, arts, etc.) for the school with multiple academies and by class period for the school with multiple class periods.

rounds of randomization to fill the remaining spots. These additional randomization rounds did not include students who were in the original randomization round. Instead, we received enrollment rosters from schools that were updated after the beginning of the school year. Students who were on the updated rosters but not on the rosters used for the original round of randomization were included in these subsequent randomization rounds. The number of randomization rounds in a school ranged from one to five. In the ITT and TOT analyses, we include randomization blocks, which are school-by-gender-by-randomization round indicator variables.

In cohort 1 (AY2018-19) we randomized 2,005 students across six study schools in CPS and NYC DOE to either a treatment group that was offered Saga Technology tutoring, or to a control group that received status quo school classes and services. In cohort 2 (AY2019-20), we randomized an additional 1,841 students to treatment and control, for a total study sample of 3,906 students across the seven CPS and NYC DOE study schools.

Six of the seven study schools used a list-wise randomization procedure, in which individual students were assigned a randomly generated number and then ordered in ascending order in accordance with that random number. To account for incomplete take-up and ensure no program slots were left empty, the research team set a number of treatment slots to offer in each school that was 15% greater than the number of program slots available in the school. The research team then moved down the randomly ordered list in each school and provided that many names to schools. This random assignment occurred before students were programmed into classes by the school.[14]

[14]One study school used a randomization procedure which entailed pairing each Saga period with an elective class that was scheduled during the same period (e.g. art, music, Spanish, or double-dose algebra). School schedulers over-enrolled each elective class period and provided the research team with a list of students for each paired elective period. The research team then used this list to randomly assign students

NYC DOE allowed principals discretion in whether to follow the provided lists for program assignment. In practice, principals in general followed the suggested program assignments. The student assignment process was therefore designed to mimic a randomized experiment while still adhering to NYC DOE's preference to maintain some degree of principal discretion.

4.3 Baseline Characteristics

Table 1 presents sample sizes, randomization rates, and take-up rates by cohort. We define take-up, or participation, as a student having attended at least one tutoring session. In cohorts 1 and 2, respectively, the take-up rate was 79% and 75% among students randomized into treatment. There was minimal control cross-over – i.e., students assigned to the control group who participated in at least 1 tutoring session (0.1% and 6.7% for cohorts 1 and 2, respectively).

Tables 2 and 3 present baseline characteristics and baseline balance tests for cohorts 1 and 2. On average, the baseline characteristics of the students in cohort 1 and 2 were fairly similar. Approximately 91 percent of students in cohort 1 and 88 percent of students in cohort 2 were eligible for free or reduced-price lunch, a measure of low family income. In cohort 1, approximately 24 percent of students were Black and 57 percent were Latinx; in cohort 2, approximately 33 percent were Black and 50 percent were Latinx. The year before the study, students in both cohorts had been absent approximately 15 days on average. Cohort 1 students had GPAs from the prior year of about C+ in both their math and core non-math classes. Cohort 2 students had slightly lower math grades in the year

from each elective period into either the Saga Technology tutoring program (treatment), or the status quo elective course (control).

prior to randomization – equivalent to about a C on average – but similar C+ average grades in non-math classes.

Tables 2 and 3 also present baseline balance tests. To test for baseline balance, we ran a regression of an indicator for random assignment to the treatment group on the baseline control variables and randomization block fixed effects. We then ran an F-test of the null hypothesis that the coefficients on all of the baseline control variables, excluding the randomization block fixed effects, were zero. The F-test rejects the null of baseline balance in both years, suggesting there were some differences in average characteristics at baseline. For cohort 1, the F-statistic is 2.179 and the p-value is 0.004, and for cohort 2, the F-statistic is 1.714 and the p-value is 0.038. Tables 2 and 3 also show the pairwise treatment-control comparisons for each baseline variable, which allow us to assess which baseline variables appear imbalanced and the magnitude of the differences. In cohort 1, the treatment group had a higher percentage of Black students (24.4 versus 23.3 percent), more days of suspensions (0.51 versus 0.24 days), and fewer failing grades (5.2 versus 7.2 percent in math courses and 4.4 versus 5.4 percent in non-math courses). In cohort 2, the treatment group had a higher percentage of Black students (35.2 versus 30.8 percent), and more days of suspensions (0.62 versus 0.27 days).

We address potential concerns related to this baseline imbalance in two ways. First, we present treatment effects both with and without baseline controls. The estimates with baseline controls address any bias that would result from differences in observables at baseline.

Second, we assess whether the differences in baseline observables were large in magnitude, separate from statistical significance. Since we can control for observable

baseline variables, the real concern is whether unobservables are imbalanced. However, a common assumption is that unobservables are positively correlated with observables (see e.g. Murphy and Topel (1990) , Altonji et al. (2005), Oster (2019)). It is thus reassuring that some of the baseline differences in observables would predict lower treatment test scores (more days of suspensions) and some would predict higher treatment test scores (fewer failing math grades).

To assess whether the baseline imbalances on net went in the direction of positive or negative bias, we created a weighted average of the baseline variables where the weights were chosen based on their correlation with the dependent variable. To do so we ran a regression of the primary outcome (standardized math test scores for cohort 1, mid-year math GPA for cohort 2) on all the baseline variables and the randomization block fixed effects. We then used this regression to calculate predicted outcomes based on the baseline variables and the randomization blocks. To isolate treatment-control differences in the baseline controls from differences in which randomization block students were in, we then subtracted the randomization block fixed effect from the predicted values. This left a weighted sum of the baseline controls, where the weights were drawn from the regression that predicts the math test score or mid-year GPA.

For cohort 1, we find that the treatment group's adjusted predicted test scores were lower than the control group's by 0.015 standard deviations. Thus, the baseline imbalance, while statistically significant, does not appear to be large in magnitude, and goes in the direction of predicting worse outcomes for the treatment group, which would suggest a negative bias of the estimated treatment effect. For cohort 2, we find that the average adjusted predicted mid-year GPAs of the treatment and control group were 2.161 and

2.158, a difference of only 0.003 grade points, on a scale where 1 point represents the difference between an A and a B, or a B and a C. Thus, it appears that while there were statistically significant differences in baseline characteristics, the differences were small in magnitude as measured by their likely influence on post-treatment outcomes.

We also assess whether there was balance in missingness of the main outcomes for both cohorts. We do not observe differential missingness, or attrition, in either cohort. These results are shown in Table 4. In cohort 1, standardized math test scores are missing for approximately 16 percent of students, though rates are equal across treatment and control groups. Because reading test scores are unavailable for all students in New York City, reading scores are missing at a significantly higher rate (63 and 67 percent for the control and treatment groups, respectively). Grade data is missing for approximately 7-8 percent of students in cohort 1 and for 11-13 percent of students in cohort 2. Attendance data is missing for 3-4 percent of students in cohort 1 and for 8 percent of students in cohort 2. None of the differences between treatment and control missingness rates are statistically significant.

5 Analysis Plan

The study is designed as a randomized controlled trial (RCT), in which randomization was stratified in randomization blocks defined by school, gender, and "randomization round" (i.e., the timing of randomization). We estimate both intent-to-treat (ITT) and treatment-on-the-treated (TOT) impacts of the Saga Technology model.

Our primary model (Model 1) estimates the ITT effect using the following equation,

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 X_i + \pi_1 B_i + \varepsilon_i \quad (1)$$

where Y_i is the outcome of interest for student i, T_i indicates whether student i was offered the chance to participate in the tutoring program, X is a set of baseline characteristics,[15] B_i is a set of randomization block indicators, ε is a random error term, and β_0, β_1, and β_2 are parameters to be estimated, and π_1 is a vector of the estimated coefficients on the randomization block fixed effects. The random assignment of T within each block assures that under standard assumptions, Ordinary Least Squares (OLS) estimation yields an unbiased estimate of the ITT effect, β_1, or the effect of being offered participation in the Saga Technology tutoring program. Students assigned to treatment are offered the opportunity but are not required to participate in tutoring so the ITT may not measure the effect of participation.

To measure the effect of participation we use random assignment of T as an instrument for participation in a two-stage least squares (2SLS) framework. If all of the tutoring participants were randomly selected to the treatment group (i.e., no control students participated in tutoring, or equivalently there was one-sided non-compliance), then under the identifying assumption that treatment assignment has no effect on outcomes of those assigned to treatment but who do not participate in tutoring, this method calculates the

[15]The baseline controls include age, race/ethnicity, English language learner status, diverse learner status, free or reduced-price lunch status, math GPA, non-math GPA, proportion of courses failed, proportion of math courses failed, proportion of non-math courses failed, standardized math and reading test scores, days absent from school, number of in-school disciplinary incidents, days of out of school suspensions, and a set of indicator variables for missing baseline data. If a student has any missing baseline value, we impute it using mean values in the control group.

effect of the treatment-on-the-treated (TOT), or the effect of participation for the group of students who choose to participate. We define participation as any student that received Saga Technology tutoring for at least one school day. This is a conservative definition of participation. Definitions based on higher thresholds of participation (e.g., having received at least one week of tutoring or having participated in at least one-quarter of tutoring sessions) would yield lower participation rates, and consequently higher TOT effects. The first stage equation is:

$$D_i = \gamma_0 + \gamma_1 T_i + \gamma_2 X_i + \pi_2 B_i + \mu_i \tag{2}$$

where D is an indicator for having participated in Saga tutoring for at least one day, the $\gamma' s$ are parameters to be estimated, μ is a random error term, and all other variables are defined as above. We then use 2SLS, instrumenting for participation with the random assignment indicator, to estimate the following relationship of interest:

$$Y_i = \delta_0 + \delta_1 D_i + \delta_2 X_i + \pi_3 B_i + \vartheta_i \tag{3}$$

and where δ_1 is the TOT, the average effect of participating in tutoring among the participants.

We assess the robustness of our results by estimating the ITT and TOT effects in models that exclude covariate controls (which we refer to in the tables as Model 2), and by calculating effects for the subset of students for whom we have complete baseline covariates, i.e., excluding students for whom we have missing baseline covariates (referred to in the tables as Model 3 (with baseline controls) and Model 4 (without baseline

controls)).

6 Results

6.1 Main Results

Table 5 shows estimated effects of Saga Technology on math learning for cohort 1, as measured by end-of-year state-administered standardized test scores. The estimated ITT effect of being assigned to treatment on end-of-year math achievement tests is 0.19 SD, with a TOT effect of 0.23 SD. These effects are statistically significant and are similar in magnitude to TOT effects of 2-to-1 in-person Saga tutoring in Guryan et al. (2023), which reported estimated ITT effects of 0.12 SD and TOT effects of 0.28. Given that existing literature suggests that students on average learn approximately 0.15-0.175 SD per year in high school (Reardon, 2011), these effect sizes are equivalent to an additional one to one-and-a-half years of additional math learning for treatment students[16]. These results suggest that the 4-to-1 hybrid model delivers a quantitatively similar treatment effect as the 2-to-1 every-day in-person tutoring model, despite having a per-pupil cost that is 30 percent lower (Guryan et al., 2023).

We also estimate an increase in math grades of 0.24 points, or one quarter of a letter grade, and a 22 percent reduction in math course failures (16.6 percent for the treatment group compared to the control complier mean of 21.2 percent), both of which are about half as large as the effects reported in Guryan et al. (2023). Given that the control group

[16]Alternatively, Bloom et al. (2008) suggests that students on average learn approximately 0.25 SD per year between grades 9 and 10 for math test scores. Under this more conservative approach, our effect sizes suggest learning gains of an additional three quarters of a year.

mean is around 1.9, the GPA effect translates into an improvement in math grades from about a C- to a C.[17] These impacts on grades in addition to test scores corroborate evidence of math learning and suggest that tutors are not simply "teaching to the test" (Guryan et al., 2023).

Unlike in Guryan et al. (2023), we do not find significant effects on non-math outcomes such as standardized reading scores, non-math GPA, or proportion of non-math courses failed for cohort 1. Similar to Guryan et al. (2023), we also do not observe significant effects on absences or on behavioral outcomes such as out-of-school suspensions for cohort 1 (Table 5).

The COVID-induced school closures in March 2020 led to the cancellation of standardized testing that would have taken place at the end of the 2019-20 school year. We are therefore not able to estimate treatment effects of the Saga Technology tutoring model on standardized test scores for cohort 2. We are, however, able to estimate effects on math and non-math GPA, course failures, and behavioral outcomes such as attendance and number of days suspended (out-of-school). These results for cohort 2 are consistent with findings for cohort 1 and are reported in Appendix Table A1. However, because the tutoring program was cut short in the middle of the 2019-20 school year, we focus instead on mid-year outcomes for cohort 2.

[17] According to the College Board, a GPA of 1.7 is equivalent to a C-, while a 2.0 GPA is a C, and 2.3 GPA is a 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).

6.1.1 Mid-Year Outcome Comparison Across Cohorts

The results for Cohort 2, served in AY2019-20, in principle could serve as a sort of 'out-of-sample' validation of the Cohort 1 (AY2018-19) results to ensure the main results are not false positives. In practice a perfect replication is complicated by the onset of the global COVID-19 pandemic that closed schools in March 2020.

Given that standardized test scores were not measured in cohort 2 (AY2019-20), we compare the mid-year outcomes on math GPA by cohort to see if we observe similar effects at a similar point in the school year, which would have been prior to the onset of pandemic-induced school closures in Cohort 2 (see Table 6). We find similar positive and significant effects for ITT and TOT across both cohorts for math GPA. For example, we find a TOT effect of 0.21 and 0.20 points on mid-year math grades for cohort 1 (AY2018-19) and cohort 2 (AY2019-20), respectively. For math course failures at mid-year, we find TOT effects of negative 4.5 percentage points and negative 7.1 percentage points for Cohorts 1 and 2, respectively. These effect sizes translate to 25% and 32% reductions in math course failures relative the control complier means.

We also find positive and significant effects on overall GPA when measured at middle of the year for both study years. The TOT estimates are 0.09 grade points for cohort 1 (AY2018-19) and 0.11 grade points for cohort 2 (AY2019-20). Whereas we do not find any significant effect of participation on overall course failures for cohort 1, we do find a significant decrease of 17 percent in overall course failures for the treatment group in cohort 2 relative to the control complier mean. Consistent with the end-of-year program effects for cohort 1, we do not find evidence of spillover effects on non-math GPA, course failures or attendance for mid-year outcomes. The consistency of results at mid-year across

the two study years suggests that we may have seen similar program effects on end-of-year standardized test scores in the absence of the COVID-19 pandemic, though we can never know for certain.

6.2 Robustness Checks

Our preferred specification, as reported in section 5.1, include controls for baseline covariates, X. In this section, we report estimates that test the sensitivity of these estimates to the baseline imbalance reported in section 4.3 and to the specific way we impute baseline covariates when they are missing. In the main specification, tor observations with missing baseline covariates, we impute missing values with the control group mean and include indicators for missingness as additional control variables.

To test the sensitivity of our estimates to the baseline imbalance reported in section 4.3 and to the specific way we impute baseline covariates, we estimate the ITT and TOT effect under three alternative models as specified in the pre-analysis plan. In particular, we include all students but drop baseline controls (Model 2), restrict analysis to the subset of students for whom we observe all baseline covariates (while continuing to control for X) (Model 3), and include only the subset of students for whom we observe all baseline covariates and drop baseline covariates (Model 4).

Table A2 reports the results for these imputation robustness checks for end-of-year outcomes for cohort 1. As seen in Table A2, we find ITT effects on standardized test scores that range from 0.19-0.23 SD and TOT effects that range from 0.23-0.28 SD. These effect sizes are similar to the ITT and TOT effect size of 0.19 and 0.23 SD we observed for standardized test scores under our main specification (See Table 5). As seen in Table A2,

effect sizes on other outcomes[18] under the alternative models are also similar to estimates for the main specification. For instance, we find effect sizes on math GPA in the range of 0.19-0.22 points (for ITT) and 0.23-0.27 points (for TOT) which are comparable to the main speficiation estimates of 0.20 points (ITT) and 0.24 points (TOT) that are reported in in Table 5. Additionally, we find qualitatively and quantitatively similar effect sizes when we analyze middle of year outcomes for cohort 1 (AY2018-19) under alternative specifications (See Table A3 in appendix) as compared to our results reported for middle of year outcomes for cohort 1 in Table 6.

Moreover, when we analyze results under alternative model specifications for cohort 2, we find qualitatively similar effect sizes as we estimate in our preferred specification. Table A4 reports these results for middle of the year outcomes for cohort 2. More specifically, we find effect sizes in the range of 0.13-0.14 points for the ITT estimates and 0.18-0.19 points for the TOT estimates for math GPA. For comparison, we find an ITT effect size of 0.15 points and a TOT effect size of 0.20 SD in our preferred specification, as reported in Table 6. Effect sizes on other outcomes such as math course failures and overall course failures are also quantitatively similar under alternative model specifications.

Overall, we find that the estimated effects on the primary outcome variable of interest, standardized math test scores for cohort 1, and secondary outcomes such as math GPA, math and overall course failures, are robust to a range of estimation decisions for both cohort 1 and cohort 2 and are not likely to be driven by variable missingness, imbalance at baseline, or imputation exercise.

[18]Secondary outcome variables as defined in the pre-analysis plan include outcomes other than standardized math test scores such as math GPA, non-math GPA, math course failures and non-math course failures.

6.3 Post-Intervention Year Program Effects

To understand longer term effects we analyze available data in the school year following treatment implementation with a focus on cohort 1, which (unlike cohort 2) received a full year of tutoring (in AY 2018-19). Because the pandemic caused so many disruptions in spring 2020, including the cancellation of end-of-year standardized testing, we measure longer-term effects using mid-year grades in AY2019-20 for cohort 1.

Estimated effects of Saga Tech on mid-year grades during the school year following the tutoring program are presented in Table 7. Despite some signs of treatment effect fade out, the results remain statistically significant. We estimate statistically significant TOT effects of approximately 0.145 points on math GPA approximately six months after treatment ends, indicating preservation of about 61 percent of the TOT estimate from the end of the program year of 0.238 points. The effects one year after tutoring on math course failures was a 20% reduction, indicating little to no fade out relative to the impacts observed during the tutoring year itself.[19]

We also check the robustness of the post-intervention year results to alternative model specifications as outlined in section 5.2. Tables A5 and A6 present these results for middle of the year outcomes in the year following treatment for cohort 1 and cohort 2 respectively. Similar to the results presented in Table 7 for program effects on mid-year outcomes in the year following treatment, we find statistically significant effects for the ITT and TOT estimate for math GPA and math course failures for cohort 1 but not for cohort 2.

[19]It is challenging to interpret longer-term treatment effects for cohort 2 because cohort 2 received tutoring for only part of the year due to the pandemic. For reference, we estimate TOT effects on mid-year math GPA, approximately one year after the tutoring program ended for cohort 2, of 0.07 points on math GPA, but this estimate is not statisically sigificant. The point estiamte would indicate a preservation of 34% of the EOY TOT effects of 0.204 points, a more significant fade out than we observe for cohort 1.

Moreover, Table A7 reports end of year outcomes for cohort 2 under alternative model specifications. Similar to our results in Table 7 under our main specification, we find quantitatively similar effect sizes under alternative model specifications.[20]

6.4 Sub-group Effects

In Tables 9 and A14, we report differential effects for students by gender, race/ethnicity, baseline math GPA quartiles, and baseline math test scores quartiles. We report effects on the primary outcome, standardized math test scores, as well as secondary outcomes such as math GPA, math course failures, and standardized reading test scores.

Gender: As can be seen in Table 9, treatment effects on standardized math scores are quite similar for boys and girls. The TOT estimate for girls and boys are 0.23 and 0.22, respectively. Effects on math GPA, non-math GPA, overall GPA, and course failures are also quantitatively similar and not statistically different between boys and girls.

Race/Ethnicity: As seen in Table 2, approximately 23 percent of students in cohort 1 are identified as Black and 59 percent as Latinx. Approximately 31 percent of the students in cohort 2 are identified as Black and 51 percent as Latinx. In models that allow the effect of tutoring to vary by race/ethnicity, which can be found in Tables 9, A8, A9, and A10, we do not find any statistical differences between Latinx and Black students in treatment effects on standardized math test scores or other outcomes, with the exception of math course failures, where we find a slightly larger effect for Latinx students.

Baseline Math Performance: In Tables 9, A8, A9, and A10, we show results separately for students broken out by quartiles of baseline standardized math test scores. We find that

[20]For instance, the ITT estimate is 0.05 points for math GPA as reported in Table 7 under Model 1 and ranges between 0.035-0.042 for alternative specification.

students in all four quartiles of baseline test scores have similar treatment effects of around 0.22 to 0.24 SD (TOT) on standardized math test scores. These estimates are individually statistically significant for each of the quartiles, and a joint test fails to reject the null hypothesis that the effects are the same across the quartiles. When we divide students by quartiles of baseline math grades, we see larger differences in the point estimates across quartiles, ranging from 0.12 to 0.27 SD, as shown in Table 9, although statistical tests do not reject the null hypothesis that these effects are the same for each quartile.[21]

7 How Much Does Technology Contribute to Student Learning in a Tutoring Model that Combines Tutors and Computer Aided Instruction?

In the analyses presented thus far, we have shown that the Saga Technology program – a combination of every-other-day tutoring and every-other-day usage of the ALEKS computer-assisted learning program – led to gains in student learning. This finding raises obvious questions: Why does Saga Technology work? Would every-other-day tutoring without the technology component have had the same effect on learning? Was the time spent on ALEKS important to student learning? To answer these questions, ideally we would randomly assign some students to work with a tutor on alternate days with no technology added, and others to work with a tutor on alternate days along with CAL access on the intervening days, and compare both conditions to a business-as-usual condition.

[21]To test if the difference between quartiles is statistically significant, we perform a series of pair-wise t-tests and a joint F-test. These results are reported in Table 9).

The experimental design of the present study is not constructed to separately identify the effect of each intervention component—tutoring and technology. However, we have usage data from the ALEKS platform that can shed light on how the use of CAL contributed to student learning. Motivating these analyses is perhaps the obvious assumption that students cannot benefit from learning technology if they do not use it. With that motivation in mind, we present evidence of: 1) how much the students in the study used ALEKS on average, 2) which students used ALEKS more and which used it less, and 3) a quasi-experimental analysis to measure the effect of ALEKS use on test score gains.

We begin by presenting evidence on how much students used the CAL platform. Table 10 provides an overview of ALEKS usage for cohort 1 (AY2018-2019).[22]. We see that ALEKS was used almost exclusively by treatment students who participated in the Saga Technology intervention (97 percent of treatment takers are observed in the ALEKS data). Students who participated in Saga Technology spent on average 1,756 minutes, or 29 hours, on ALEKS. If the Saga Technology program had been implemented with fidelity as designed, each student participating in the program would have used ALEKS for approximately 75 minutes per week, or 45 hours over the course of the school year.[23]

[22]We also have ALEKS data for cohort 2 (AY2019-2020). We focus here on describing ALEKS usage in cohort 1 because that is the cohort where we can relate ALEKS usage to end of year standardized test scores. We provide an overview of ALEKS data in the appendix (see Table A17 and Figures A.1-A.3 We have slightly lower coverage in cohort 2 (90 percent compared with 97 percent of treatment takers in cohort 1) and the average number of hours is also slightly lower (21 hours compared to 29 hours). After March 2020, 26 percent of students used ALEKS remotely during COVID-related school closures.

[23]Since students should have spent every other day on ALEKS, they were expected to spend 2.5 days, on average, every week, with every session lasting for approximately 30 minutes. While a Math Lab tutoring session is approximately 50 minutes in total, students spend the first part of each session working on a "Do Now" problem, transition to working either on ALEKS or in-person with their tutor, and end with a "Ticket to Leave" problem.

Thus, the average student used ALEKS for about two-thirds (29/45) of the designed dosage. Figure 2 shows that there was considerable variation in usage. Some students in the program spent nearly no time at all using the CAL platform, while others spent close to 60 hours or more. The interquartile range of ALEKS usage was 21.4 hours. The usage data is inclusive of all time a student spent on ALEKS: during the tutoring session, at other times during the school day, or at home. However, students were not encouraged to work on the platform outside of "Math Lab" and anecdotal evidence we gathered suggests this was an infrequent occurrence.

We also observe the number of topics students attempted and the number of topics on which students demonstrated learning. On average in AY2018-19, students who participated in Saga Technology attempted 129 topics on ALEKS and demonstrated learning on about three-fourths of those, or 96 topics.[24] Given the districts in the study on average had a 35-week school year, the average student spent 50 minutes per week on ALEKS, attempted 3.2 topics per week, and demonstrated learning on 2.3 topics per week. We present cohort 2 usage in the Appendix (see Table A17 and Figures A.1-A.3). With the clear exception of the period after COVID-related school closures began, usage in cohort 2 followed similar patterns. Between March 2020 and the end of the 2019-20 school year only 26% of students used the platform remotely. While comparisons with the early days of the COVID pandemic are confounded for obvious reasons, this change in usage points lends to the possibility that the rates of usage we observed in cohort 1 and in cohort 2 prior to March 2020 were dependent on the school-based structures that encouraged students to

[24] ALEKS defines "topics attempted" as the number of topics that a student has attempted to learn but not yet successfully completed. ALEKS defines a "topic learned" as a student achieving a total of 5 points on questions answered on the topic.

use the platform systematically during the school day.

We next present evidence on how ALEKS usage was correlated with baseline student characteristics. Table 11 shows that students with higher math grades, higher reading test scores, higher attendance, and lower levels of misconduct at baseline used the computer-assisted learning platform more intensively. This subset of students spent more time on the ALEKS platform, attempted more topics, and demonstrated learning of more topics. For example, we see that a one-standard-deviation increase in baseline standardized reading test scores was correlated with students spending 1.6 more hours on the ALEKS platform, leading to approximately seven more topics attempted and six more learned. These findings, while not causal, may have implications for the potential role of technology in addressing long-standing inequities in education. In this controlled setting—in school and with an adult monitoring—the students who used the learning technology more were the students who were already performing better on tests and attending school more often. The students who were farther behind were less likely to use the learning technology. Returning to the idea that students can only benefit from learning technologies if they use them, these patterns are concerning if the goal is to help narrow existing gaps or to help students who are behind grade level to catch up. One might also reasonably conjecture that usage outside of the context of a structured tutorial—for example, outside of school hours or on a voluntary basis—may vary more widely and be more inequitably distributed.

Finally, we present a quasi-experimental analysis of the effect of ALEKS usage on test scores. This analysis investigates whether the CAL technology component of the tutoring program contributed to student learning by assessing whether students who used the ALEKS platform more had larger treatment effects on test scores. Before we describe

the quasi-experimental analysis, it is worth considering why an analysis based on the correlation between ALEKS usage and test score gains would be confounded. Two sources of bias are particularly important. First, the vast majority of the variation in ALEKS usage is the difference in average usage between treatment students and control students (almost all of whom did not use ALEKS at all). Thus, the correlation between ALEKS usage and test scores would confound the effect of ALEKS usage with the effects of the other components of the treatment, namely tutoring. Second, as we just showed, the variation in ALEKS usage among the treatment students is correlated with observables—baseline test scores, grades, and attendance—that are likely also correlated with unobservable predictors of test scores. Thus, even the correlation between test scores and ALEKS usage among treatment students would likely be subject to omitted variable bias.

To address these sources of bias, we conduct a quasi-experimental analysis similar to the analysis in Kling, Liebman and Katz (2007). Specifically, we estimate a two-stage least squares (2SLS) model that uses randomization-block-by-treatment indicators as instruments for ALEKS usage. The first-stage model predicts treatment-control differences in ALEKS usage rates at the randomization block level. The second stage estimates the relationship between randomization-block-level treatment effects on test scores and randomization-block-level differences in ALEKS usage. The quasi-experimental analysis essentially estimates whether the treatment effects on test scores were higher in the randomization blocks where ALEKS usage was higher than in the randomization blocks where ALEKS usage was lower. The key identifying assumption is that variation in ALEKS usage across randomization blocks was not correlated with any other factor that caused treatment effects of the tutoring program to vary.

We present the results of the 2SLS analysis in Table 12. The top and bottom panels of the table show estimates of the effect of ALEKS usage on end-of-year math test scores and math GPA, respectively. Each entry in the table is an estimate from a separate regression. For comparison purposes, the first column shows ordinary least squares estimates of the correlation between ALEKS usage and math outcomes. The latter two columns present the 2SLS estimates, one version without baseline controls and one that adds baseline controls. All of the estimates in the table are statistically significant.

The top row presents the 2SLS estimates of the effect of time spent on ALEKS on end-of-year math test scores. The coefficient is 0.007, which indicates that each additional hour of ALEKS usage increases math test scores by 0.007 standard deviations. To put this magnitude into perspective, consider that the interquartile range of ALEKS usage over the school year was 21.4 hours. The 0.007 coefficient estimate implies that a 21.4 hour increase in ALEKS usage would generate a 0.15 standard deviation increase in test scores. Thus, variation in ALEKS usage on the order of the interquartile range was associated with variation in test score effects two-thirds the size of the average TOT effect (0.15/0.23). The estimated effects of the other two measures of CAL use, total topics attempted and total topics learned, are 0.001 and 0.002 respectively. The interquartile ranges of topics attempted and learned were 133 and 102. Thus, an interquartile range increase in topics attempted on ALEKS was associated with a 0.13 standard deviation increase in test scores, and an interquartile range increase in topics learned on ALEKS was associated with a 0.20 standard deviation increase in test scores. These are 57 percent and 87 percent of the average TOT estimate, respectively.

The 2SLS estimates of the effect of CAL usage on end-of-year math GPA, shown

in the bottom panel of the table, are similar in magnitude. The coefficient estimates for total time spent on ALEKS, topics attempted, and topics learned are 0.008, 0.002, and 0.002, respectively. An increase in time spent on ALEKS equivalent to the interquartile range was associated with a 0.17 standard deviation increase in math GPA, which is 71 percent of the average TOT estimated effect of 0.24. An increase in topics attempted on ALEKS equivalent to the interquartile range was associated with a 0.27 standard deviation increase in math GPA, which is 113 percent of the average TOT estimated effect. An increase in topics learned on ALEKS equivalent to the interquartile range was associated with a 0.20 standard deviation increase in math GPA, which is 84 percent of the average TOT estimated effect.

While not definitive, these results suggest that the use of ALEKS played some role in the treatment effects we estimate for the Saga Technology intervention, and that every-other-day tutoring that is not paired with learning technology on the alternate days might generate substantially smaller learning effects.

8 Benefit-Cost Analysis

In this section, we present a benefit-cost analysis of the Saga Technology tutoring model. To facilitate comparisons with the Saga 2-to-1 tutoring model, we follow the methodology used to calculate benefit-cost ratios in Guryan et al. (2023).[25]

[25]We explored the possibility of a three-arm study where students would be randomly assigned to Saga's traditional model, Saga's technology model, or to a control group. Unfortunately, a combination of factors made implementing a three-arm study impossible. In the absence of the three-arm study, as a next-best analysis, we present comparisons of cost-benefit ratios of the Saga technology model with existing studies.

Guryan et al. (2023) calculated the cost per pupil of the 2-to-1 Saga tutoring model at approximately $3,500 per pupil (with a defensible range of $3,200 to $4,800). For the Saga Technology model, we calculate costs at $2,200 per pupil with a defensible range of $1,900 to $2,600 as shown in Table A18. The per-pupil costs of the Saga Technology model are approximately one-third lower than the costs of the 2-to-1 Saga model. The cost reductions derive from substituting tutor time for time spent on the ALEKS platform, which had a per-pupil cost of about $45 per student per year at the time of the program implementation.

To calculate the benefits on expected earnings as outlined in the pre-analysis plan, we combine findings from Chetty et al. (2011) with the study results to calculate the expected increase in participants' adult earnings. Chetty et al. (2011) finds that each one-percentile increase in 8th grade test scores is associated with approximately $150 in additional annual earnings. To estimate the present discounted value of lifetime earnings gains using this approach, we convert the standardized test scores into percentile scores following Kline and Walters (2016) and Guryan et al. (2023). We find that, on average, the Saga Technology model increases student's test scores by roughly 6 percentile points for students who are assigned to treatment (ITT estimate) and by 7 percentile points for students who participate in the program for at least one day (TOT estimate).[26] Using the approach outlined above and adjusting the gain in annual earnings for inflation using CPI-U, our best estimate is that the Saga Technology model increases adult yearly earnings by approximately $950 dollars for each student offered the chance to participate (ITT estimate) and $1,150 for each program participant (TOT estimate). We then estimate the

[26]We estimate benefits for cohort 1 only since we have don't have data on standardized test scores for cohort 2 due to impact of COVID-19 in AY2019-20.

present discounted value of earnings gains as a result of the Saga Technology model. Following Guryan et al. (2023), we assume these gains accrue from ages 25 to 59 and use a discount rate of 5% to discount benefits to age 14 which is the mean age for students attending 9th grade in our sample and is the age at which costs are being incurred. We find that the present discounted value of the earnings gains is about $10,800 using TOT estimates and around $8,900 under ITT estimates. With per-student costs ranging between $2,000 to $2,600 respectively under the ITT and TOT estimates[27], this implies the benefit-cost ratio ranges between 4.3 to 4.7 for students offered an opportunity to participate in the Saga Technology program (ITT) and 4.2 to 4.5 for students who participated in the Saga Technology program (TOT). The reason for the slight difference between benefit-cost ratios under the ITT and TOT approach is that the benefits are estimated using regressions based on the sample of students for whom we have standardized math test scores whereas costs are calculated using the original assignment (and take-up) to treatment.[28] The implied benefit-cost ratio range was 2.4-3.6 for study 1 and in the range of 5.4-8.0 for study 2 for the 2-to-1 Saga tutoring model using TOT estimates (Guryan et al., 2023). The benefit-cost ratios are therefore higher than those found in Guryan et al. (2023) for study 1 and only slightly lower than the range of study 2.

As means of comparison with other interventions, often cited early childhood programs such as Abecedarian Project (Masse and Barnett, 2002) and Perry Preschool Program (Heckman et al., 2010) have benefit-cost ratios that are on the order of 1.9-2.2 and 3.9-6.8

[27]The average total cost is around $2,050 and $2,590 per ITT and TOT estimates respectively. Similarly, as shown in Table A18, the average variable cost is $1,890 and $2,390 under ITT and TOT estimates.

[28]More specifically, the take-up rate for the assignment to treatment is around 79%, whereas it is around 82% for the sample of students for whom we have standardized test scores, which leads to the small difference in the benefit-cost ratios under the two approaches.

respectively when estimated using 5% discount rate as in our case. Moreover, Krueger (2003) yields a benefit-cost ratio of about 2 using a 4% discount rate for a 7-student reduction in class size in grades K-3.[29] Similarly, Borman and Hewes (2002) show that the Success for All model yields a 0.04 SD improvement in math per $1,000 spent. The corresponding measure for the Saga Technology model is 0.08 SD improvement in math test scores per $1,000 spent based on ITT estimates and 0.10 SD improvement in test scores based on TOT estimates.[30]

We also estimate benefits using the approach outlined in Hanushek and Woessmann (2008), who upon reviewing several studies find that a standard deviation increase in test scores is associated with a 12% increase in earnings. We apply this effect size, in combination with our estimated increase on standardized test scores (as reported in Table 5), on a quadratic earnings age trajectory estimated using the 2019 American Community Survey data on Black and Latinx individuals, between the ages of 25 and 59, from Chicago and New York City (discounted to age 14 at a 5% rate) to estimate a benefit-cost ratio of 3.6 which is higher than the benefit-cost ratios of 1.4 to 2.2 reported for the 2-to-1 model for study year 1 and is in the range of estimates (3.2-4.8) reported for study year 2 in Guryan et al. (2023).

This evidence indicates the reported benefit-cost ratios of the Saga Technology model compare favorably in magnitude not just to the 2-to-1 Saga tutoring model but also to some of the most successful early childhood programs such as the Abecederian Project, the Perry Preschool Program, the Tennessee STAR class size reduction experiment, as

[29]Using a 5% discount rate would lead to a benefit-cost ratio for the Tennessee Star class size reduction below 2.

[30]Success for All, however, also led to improvements in reading scores (a 0.09 SD improvement per $1,000) whereas we do not find significant impacts of math tutoring on reading test scores in this context.

well as the Success for All Model. In strategically substituting some time away from in-person tutors to a technology platform, the Saga Technology model was able to lower the cost of the intervention by approximately one-third with comparable effectiveness and cost-effectiveness.

9 Discussion

In the aftermath of the global COVID-19 pandemic, the U.S. Secretary of Education, Miguel Cardona, strongly 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 (U.S. Department of Education, 2022). This recommendation aligned with the advice of numerous experts who advocated for tutoring as a key strategy to mitigate the learning setbacks caused by the pandemic, based largely on the weight of the evidence (Robinson et al., 2021). However, school districts aiming to expand their tutoring initiatives are encountering various obstacles (Carbonari et al., 2022). One of the most significant challenges to scaling up the proven benefits of high-dosage tutoring is its cost. Even with the one-time influx of funds through the American Rescue Plan that allows school districts to use these resources towards providing tutoring, the per-student cost of tutoring limits the number of students that districts can serve (Nickow et al., 2024; Carbonari et al., 2022). Furthermore, the difficulty of attracting and hiring full-time teachers and tutors, possibly resulting from what has been dubbed the 'Great Resignation', is likely to further increase the costs associated with high-dosage tutoring.[31]

[31]Offering higher wages would attract more tutors but would increase the per-pupil cost and therefore would reduce the number of children who could be served within a fixed budget (Guryan et al., 2023).

This paper evaluates a lower-cost and more scalable tutoring model developed by Saga Education, called "Saga Technology." Instead of daily 2-to-1 tutoring, students work with tutors in groups of 4-to-1, alternating between days working with the tutor in student pairs and days on the computer-assisted learning (CAL) platform, ALEKS. Through an RCT implemented in partnership with Saga Education, the Chicago Public Schools (CPS), and the New York City Department of Education (NYC DOE) during the 2018-19 and 2019-20 academic years, we estimate learning gains comparable to those achieved through daily 2-to-1 tutoring. It is possible, in other words, through the selective incorporation of technology into tutoring, to cut costs by around 30% and reduce the number of tutors required to serve a given number of students by 50%, with comparable effects on learning.[32]

While the experiment was not designed to decompose the effects of the tutoring and technology components separately, a 2SLS analysis shows a positive and statistically significant relationship between ALEKS (CAL) usage and student learning. These results suggest the CAL component of the tutoring model contributed to the overall treatment effects. Given that we also see that student baseline characteristics are predictive of ALEKS usage – higher performing students spend more time on ALEKS and consequently attempt and learn more topics – this has important implications for understanding how to strategically incorporate technology in educational settings.

How much further could we go in incorporating technology to lower cost (and labor

[32]Our best estimate for the per-pupil cost of the Saga Technology model during the study years is around $2,250 with a defensible range of $1,900 to $2,600 which highlights the cost savings relative to Saga's 2-to-1 in-person tutoring model which had costs in the range of $3,200 to $4,800 (Guryan et al., 2023). Since this evaluation, Saga's per-pupil costs have continued to decline and current estimates indicate a per-pupil cost of around $1,800-$2,200 for the model evaluated in this study which would in turn allow the program to be delivered to even more students.

requirements), thereby improving scalability, before we begin to compromise learning effectiveness? Does the answer to that question depend on the specific student and their tolerance for time on CAL? These are questions we leave for future research.

The tutoring model we study in this paper incorporates elements of computer assisted learning (CAL) into tutoring in a way that maintains the factors that make high-dosage tutoring effective: a personal and sustained relationship between tutor and student; sessions during the school day to ensure attendance; and daily sessions to ensure engagement while alternating between an in-person tutor and CAL software. Future studies should consider whether other variations in the structure of high-dosage tutoring might further reduce costs while maintaining the substantial learning benefits. Still, the findings in this paper indicate the potential of incorporating CAL into high-dosage tutoring to facilitate scaling this promising intervention to more students.

References

Akpınar, Ezgin et al. (2021) "The effect of online learning on tertiary level students mental health during the COVID-19 lockdown," *The European Journal of Social & Behavioural Sciences*.

Altonji, Joseph G, Todd E Elder, and Christopher R Taber (2005) "Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools," *Journal of political economy*, 113 (1), 151–184.

Angrist, Joshua D, Susan M Dynarski, Thomas J Kane, Parag A Pathak, and Christopher R Walters (2012) "Who benefits from KIPP?" *Journal of policy Analysis and Management*, 31 (4), 837–860.

Angrist, Joshua D., Parag A. Pathak, and Christopher R. Walters (2013) "Explaining Charter School Effectiveness," *American Economic Journal: Applied Economics*, 5 (4), 1–27, 10.1257/app.5.4.1.

Angrist, Joshua and Victor Lavy (2002) "New evidence on classroom computers and pupil learning," *The Economic Journal*, 112 (482), 735–765.

Bettinger, Eric, Robert Fairlie, Anastasia Kapuza, Elena Kardanova, Prashant Loyalka, and Andrey Zakharov (2023) "Diminishing Marginal Returns to Computer-Assisted Learning," *Journal of Policy Analysis and Management*, 42 (2), 552–570.

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

Bloom, Howard S, Carolyn J Hill, Alison Rebeck Black, and Mark W Lipsey (2008) "Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions," *Journal of Research on Educational Effectiveness*, 1 (4), 289–328.

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), 243–266.

Canfield, Ward (2001) "ALEKS: A Web-based intelligent tutoring system," *Mathematics and Computer Education*, 35 (2), 152.

Carbonari, Maria V, Miles Davison, Michael DeArmond et al. (2022) "The Challenges of Implementing Academic COVID Recovery Interventions: Evidence from the Road to Recovery Project."

Chetty, Raj, John N Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan (2011) "How does your kindergarten classroom affect your earnings? Evidence from Project STAR," *The Quarterly journal of economics*, 126 (4), 1593–1660.

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), 655–681.

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), 133–152.

Davis, Jonathan MV, Jonathan Guryan, Kelly Hallberg, and Jens Ludwig (2017) "The economics of scale-up,"Technical report, National Bureau of Economic Research.

Dee, Thomas S (2023) "Where the kids went: Nonpublic schooling and demographic change during the pandemic exodus from public schools," *Teachers College Record*, 125 (6), 119–129.

——— (2024) "Higher chronic absenteeism threatens academic recovery from the COVID-19 pandemic," *Proceedings of the National Academy of Sciences*, 121 (3), e2312249121.

Dorn, Emma, Bryan Hancock, Jimmy Sarakatsannis, and Ellen Viruleg (2020) "COVID-19 and student learning in the United States: The hurt could last a lifetime," *McKinsey & Company*, 1, 1–9.

Drane, Catherine F, Lynette Vernon, and Sarah O'Shea (2021) "Vulnerable learners in the age of COVID-19: A scoping review," *The Australian Educational Researcher*, 48 (4), 585–604.

Escueta, Maya, Vincent Quan, Andre Joshua Nickow, and Philip Oreopoulos (2017) "Education technology: An evidence-based review."

Fang, Ying, Zhihong Ren, Xiangen Hu, and Arthur C Graesser (2019) "A meta-analysis of the effectiveness of ALEKS on learning," *Educational Psychology*, 39 (10), 1278–1292.

Fryer, RG (2012) "Learning from the successes and failures of charter schools," *The Hamilton Project*, 1–18.

García, Emma and Elaine Weiss (2020) "COVID-19 and Student Performance, Equity, and US Education Policy: Lessons from Pre-Pandemic Research to Inform Relief, Recovery, and Rebuilding.," *Economic Policy Institute*.

Guryan, Jonathan, Jens Ludwig, Monica P. Bhatt et al. (2023) "Not Too Late: Improving Academic Outcomes among Adolescents," *American Economic Review*, 113 (3), 738–65, 10.1257/aer.20210434.

Hanushek, Eric A and Ludger Woessmann (2008) "The role of cognitive skills in economic development," *Journal of economic literature*, 46 (3), 607–668.

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), 114–128.

Heffernan, Neil T and Cristina Lindquist Heffernan (2014) "The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching," *International Journal of Artificial Intelligence in Education*, 24, 470–497.

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

Huffman, Kevin (2020) "Homeschooling during the coronavirus will set back a generation of children," *The Washington Post*.

Kline, Patrick and Christopher R Walters (2016) "Evaluating public programs with close substitutes: The case of Head Start," *The Quarterly Journal of Economics*, 131 (4), 1795–1848.

Krueger, Alan B (2003) "Economic considerations and class size," *The economic journal*, 113 (485), F34–F63.

LaFave, Allison J, Joseph A Taylor, Amelia M Barter, and Ariel S Jacobs (2022) "Student Engagement on the National Assessment of Educational Progress (NAEP): A Systematic Review and Meta-Analysis of Extant Research," *Educational Assessment*, 1–24.

Limone, Pierpaolo and Giusi Antonia Toto (2021) "Psychological and emotional effects of Digital Technology on Children in Covid-19 Pandemic," *Brain Sciences*, 11 (9), 1126.

Malamud, Ofer and Cristian Pop-Eleches (2011) "Home computer use and the development of human capital," *The Quarterly journal of economics*, 126 (2), 987–1027.

Masse, Leonard N and W Steven Barnett (2002) "A Benefit Cost Analysis of the Abecedarian Early Childhood Intervention.."

Murphy, Kevin M and Robert H Topel (1990) "Efficiency wages reconsidered: Theory and evidence," in *Advances in the Theory and Measurement of Unemployment*, 204–240: Springer.

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," Working Paper 27476, National Bureau of Economic Research, 10.3386/w27476.

——— (2024) "The Promise of Tutoring for PreK–12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence," *American Educational Research Journal*, 61 (1), 74–107.

Oster, Emily (2019) "Unobservable selection and coefficient stability: Theory and evidence," *Journal of Business & Economic Statistics*, 37 (2), 187–204.

Reardon, Sean F (2011) "The widening academic achievement gap between the rich and the poor: New evidence and possible explanations," *Whither opportunity*, 1 (1), 91–116.

Ritter, Steven, John R Anderson, Kenneth R Koedinger, and Albert Corbett (2007) "Cognitive Tutor: Applied research in mathematics education," *Psychonomic bulletin & review*, 14 (2), 249–255.

Robinson, Carly D, Matthew A Kraft, Susanna Loeb, and Beth E Schueler (2021) "Accelerating Student Learning with High-Dosage Tutoring. EdResearch for Recovery Design Principles Series.," *EdResearch for Recovery Project*.

Rockoff, Jonah E (2004) "The impact of individual teachers on student achievement: Evidence from panel data," *American economic review*, 94 (2), 247–252.

U.S. Department of Education (2022) "Secretary Cardona's Vision for Education in America," `https://www.ed.gov/news/speeches/priorities-speech`.

Tables and Figures

Table 1: Randomization and Take-up rate, by cohort

Cohort	N	Randomization Rate (%)	Treatment Take-up Rate(%)	Control Cross-over Rate(%)
Cohort 1 (AY2018-2019)	2005	51.27	79.28	0.10
Cohort 2 (AY2019-2020)	1841	56.06	74.71	6.67

Notes. Randomization rate is the percent of students in the study who were assigned to the treatment condition. Take-up is defined as having participated in at least one tutoring session. Control crossover rate is the percent of the control group who participated in at least one tutoring session. For cohort 1, take-up rate across the 6 study schools varies from 65-98%. For cohort 2, take-up rate across the 7 study schools varies from 42-92%.

Table 2: Cohort 1 Baseline Characteristics (AY2018-2019)

	Control Group (N = 977)	Treatment Group (N = 1028)	T-Test (P-value)
Age	14.053	14.005	0.426
% Diverse Learner	14.841	18.482	0.166
% Free/Reduced-Price Lunch	91.709	90.272	0.262
% English Language Learner	11.668	13.035	0.144
% Black	23.337	24.416	0.043**
% Latinx	58.854	55.350	0.889
Number of Absences	13.355	14.035	0.850
Number of Days Suspended (Out-of-School)	0.244	0.507	0.078*
Number of Disciplinary Incidents	0.312	0.383	0.188
Math GPA	2.231	2.198	0.721
Non-Math GPA	2.483	2.501	0.126
Standardized Math Test Score	0.000	-0.010	0.485
Standardized Reading Test Score	0.000	0.015	0.491
Proportion of Math Course Failed	0.072	0.052	0.024**
Proportion of Non-Math Course Failed	0.054	0.044	0.079*
Proportion of Overall Courses Failed	0.058	0.046	0.036**
F(16, 1957) = 2.179			
F-test of Joint significance (P-value) = 0.004***			
F-test using Randomization Inference (P-value) = 0.026**			

Notes. *** p<0.01, ** p<0.05, * p <0.1 Students are identified as diverse learners when they receive accommodations such as Individualized Education Plans (IEP). The free/reduced-price lunch indicator is based on whether a student receives free, reduced, or at-cost food services. English language learners are students in an English as a Second Language program. Race/ethnicity identifiers (such as Black and Latinx) are based on administrative data. Behavioral outcomes such as number of absences, days suspended, and disciplinary incidents measure the number of times/days that a certain incident occurs. Math GPA is calculated based on a student's performance in their primary math course. Non-math GPA is calculated based on student performance on their core courses other than math (Science, Social Studies, English). To construct standardized test scores, we use individual test scores for the reading and math sections from the Spring test date and generate a z-score using the control mean and standard deviation of the scores. Proportion of course failures is determined by the quantity of Fs awarded during the school year and is presented based on overall, math, and non-math courses. Fixed effects using randomization blocks are included in all estimation regressions. We calculate the True F-statistic using the actual treatment assignment, and also by randomly re-assigning the treatment indicator within randomization blocks (fixing randomization rate within the block) and estimating the corresponding F-statistic and the associated p-value. We repeat this process 10,000 times. In the distribution of 10,000 observations, we see where the true F-statistic lies, and report the rank, e.g., the true F-statistic has a rank of 260 among 10,000 observations.

Table 3: Cohort 2 Baseline Characteristics (AY2019-2020)

	Control Group (N = 809)	Treatment Group (N = 1032)	T-Test (P-value)
Age	14.044	14.035	0.490
% Diverse Learner	20.148	18.992	0.775
% Free/Reduced-Price Lunch	87.515	88.857	0.425
% English Language Learner	14.091	13.081	0.397
% Black	30.779	35.174	0.096*
% Latinx	51.422	49.128	0.658
Number of Absences	16.033	14.781	0.141
Number of Days Suspended (Out-of-School)	0.274	0.623	0.005***
Number of Disciplinary Incidents	0.367	0.485	0.119
Math GPA	2.039	2.042	0.973
Non-Math GPA	2.322	2.324	0.880
Standardized Math Test Score	0.000	0.026	0.472
Standardized Reading Test Score	0.000	-0.028	0.628
Proportion of Math Courses Failed	0.088	0.070	0.183
Proportion of Non-Math Courses Failed	0.067	0.065	0.948
Proportion of Overall Courses Failed	0.071	0.066	0.630
F(16, 1799) = 1.714			
F-test of Joint Significance (P-value) = 0.038**			
F-test using Randomization Inference (P-value) = 0.089*			

Notes. *** p<0.01, ** p<0.05, * p <0.1. Students are identified as diverse learners when they receive accommodations such as Individualized Education Plans (IEP). The free/reduced-price lunch indicator is based on whether a student receives free, reduced, or at-cost food services. English language learners are students in an English as a Second Language program. Race/ethnicity identifiers (such as Black and Latinx) are based on administrative data. Behavioral outcomes such as number of absences, days suspended, and disciplinary incidents measure the number of times/days that a certain incident occurs. Math GPA is calculated based on a student's performance in their primary math course. Non-math GPA is calculated based on student performance on their core courses other than math (Science, Social Studies, English). To construct standardized test scores, we gather individual test scores for the reading and math sections from the Spring test date each year and generate a z-score using the control mean and standard deviation of the scores. Proportion of course failures is determined by the number of Fs awarded in courses taken during the school year and is presented based on overall, math, and non-math courses. Fixed effects using randomization blocks are included in all estimation regressions. We calculate the true F-statistic using the actual treatment assignment, and also by randomly re-assigning the treatment indicator within randomization blocks (fixing randomization rate within the block) and estimating the corresponding F-statistic and the associated p-value. We repeat this process 10,000 times. In the distribution of 10,000 observations, we see where the true F-statistic lies, and report the rank, e.g., the true F-statistic has a rank of 890 among 10,000 observations.

Table 4: Outcome Missingness by Cohort and Treatment Group

	Cohort 1			Cohort 2		
% Missing	Control Group (N = 977)	Treatment Group (N = 1028)	T-Test (P-value)	Control Group (N = 809)	Treatment Group (N = 1032)	T-Test (P-value)
Math Test Score	16.17	16.34	0.78			
Chicago	15.76	15.66	0.93			
NYC	16.49	16.77	0.66			
Reading Test Score	63.36	67.51	0.93			
Chicago	15.76	15.66	0.93			
EOY GPA	8.29	7.30	0.65	11.37	12.11	0.91
EOY Math GPA	8.50	7.39	0.67	12.11	12.98	0.98
EOY Non-Math GPA	8.29	7.30	0.65	11.37	12.11	0.91
EOY Attendance Data	4.30	3.02	0.23	8.41	7.85	0.47
MOY GPA	8.90	7.68	0.54	11.99	12.69	0.85
MOY Math GPA	9.21	7.88	0.56	12.61	13.37	0.90
MOY Non-Math GPA	8.90	7.68	0.54	12.11	12.69	0.78
MOY Attendance Data	60.59	64.11	0.18	68.97	63.57	0.54
Chicago	9.41	6.82	0.18	11.62	10.69	0.54

Notes. The value displayed for t-tests are p-values. Standard errors are robust. Fixed effects using variable block are included in all estimation regressions. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level. Math and reading test score variables are determined at the time of standardized testing administration. End-of-year (EOY) and middle-of-year (MOY) are presented across both cohorts. EOY outcomes are based on a student's performance at the end of the school year, while MOY outcomes are based on a student's performance after their first semester. GPA is calculated based on a student's performance across their core classes (Math, Science, Social Studies, and English). Math GPA is calculated based on a student's performance in their primary math course. Non-math GPA is calculated based on student performance on their core courses other than math (Science, Social Studies, English). Attendance is determined based on the number of days a student is absent, and includes both excused and unexcused absences.

Table 5: End of Year Program Effects with Baseline Covariates, Cohort 1 (AY2018-2019)

	N	Control Mean	ITT	Control Complier Mean	TOT
Standardized Math Score	1679	0.000	0.190***	-0.035	0.230***
			(0.034)		(0.041)
Math GPA	1846	1.931	0.197***	1.883	0.238***
			(0.046)		(0.054)
Proportion of Math Courses Failed	1846	0.199	-0.038***	0.212	-0.046***
			(0.014)		(0.016)
Overall GPA	1849	2.058	0.053	1.988	0.064
			(0.038)		(0.045)
Proportion of Overall Courses Failed	1849	0.175	-0.004	0.188	-0.005
			(0.011)		(0.013)
Standardized Reading Score	692	0.000	-0.049	-0.008	-0.070
			(0.056)		(0.078)
Non-Math GPA	1849	2.094	0.008	2.018	0.010
			(0.039)		(0.046)
Proportion of Non-Math Courses Failed	1849	0.168	0.006	0.181	0.007
			(0.011)		(0.013)
Number of Days Absent	1932	18.308	-0.051	19.408	-0.062
			(0.888)		(1.075)
Number of Days Suspended (Out of School)	2005	0.262	-0.026	0.359	-0.034
			(0.074)		(0.093)

Notes. Program effects estimated on sample of students who were randomized into Saga Technology in AY2018-2019, using the following baseline covariates: randomization block, age, race/ethnicity, English language learner status, diverse learner status, socioeconomic status, math GPA, non-math GPA, proportion of courses failed, proportion of math courses failed, proportion of non-math courses failed, standardized math and reading test scores, number of days absent from school, number of in-school disciplinary incidents, number of days of out-of-school suspensions, and a set of indicator variables for missing baseline data. To construct standardized test scores, we gather individual test scores for the reading and math sections from the Spring test date each year and generate a z-score using the control mean and standard deviation of the scores. Math GPA is determined by a student's grade in their primary math course. GPA is on a 0.00-4.00 point scale using only core classes (Math, Science, Social Studies and English). Note that we only have outcome standardized reading test scores for our Chicago sample. Proportion of math courses failed is calculated based on the proportion of math courses attempted and math courses failed. Non-math GPA is calculated based on a student's grade in each of their core classes other than math (Science, Social Studies, and English). Number of days absent include excused and unexcused absences. We measure out-of-school suspensions as the total number of days a student is suspended during the school year. Missing baseline data has been imputed using mean values of the control group.

Table 6: Mid-Year Program Effects with Baseline Covariates, for Cohort 1 (AY2018-2019) & Cohort 2 (AY2019-2020)

	Cohort 1					Cohort 2				
	N	Control Mean	ITT	Control Complier Mean	TOT	N	Control Mean	ITT	Control Complier Mean	TOT
Math GPA	1834	2.053	0.177***	2.028	0.214***	1601	1.884	0.154***	1.797	0.204***
			(0.049)		(0.058)			(0.053)		(0.069)
Proportion of Math Courses Failed	1834	0.172	-0.037**	0.183	-0.045**	1601	0.198	-0.054***	0.222	-0.071***
			(0.015)		(0.018)			(0.017)		(0.022)
Overall GPA	1839	2.138	0.074*	2.064	0.090**	1613	1.954	0.083**	1.893	0.110**
			(0.038)		(0.046)			(0.041)		(0.054)
Proportion of Overall Courses Failed	1839	0.155	-0.011	0.168	-0.013	1613	0.177	-0.024**	0.184	-0.032**
			(0.011)		(0.013)			(0.012)		(0.015)
Non-Math GPA	1839	2.165	0.040	2.078	0.048	1612	1.972	0.067	1.914	0.089
			(0.040)		(0.047)			(0.043)		(0.056)
Proportion of Non-Math Courses Failed	1839	0.149	-0.002	0.162	-0.003	1612	0.170	-0.016	0.174	-0.021
			(0.011)		(0.014)			(0.012)		(0.016)
Number of Days Absent	754	7.395	-0.213	7.498	-0.300	627	7.725	-0.297	7.636	-0.439
			(0.501)		(0.692)			(0.551)		(0.792)

Notes. Treatment effects estimated on sample of students who were randomized into Saga Tech in Cohort 1 (AY2018-19) and Cohort 2 (AY2019-20), using the following baseline covariates: randomization block, age, race/ethnicity, English language learner status, diverse learner status, socioeconomic status, math GPA, non-math GPA, proportion of courses failed, proportion of math courses failed, proportion of non-math courses failed, standardized math and reading test scores, number of days absent from school, number of in-school disciplinary incidents, number of days of out of school suspensions, and a set of indicator variables for missing baseline data. GPA is on a 0.00-4.00 point scale using only core classes (Math, Science, Social Studies and English). To construct standardized test scores, we gather individual test scores for the reading and math sections from the Spring test date each year and generate a z-score using the control mean and standard deviation of the scores. Number of days absent include excused and unexcused absences. We measure out of school suspensions as the total number of days a student is suspended during the school year. Missing baseline data has been imputed using mean values of the control group.

Table 7: Program Effects on Mid-Year Outcomes in Year Following Treatment, Controlling for Baseline Covariates, by Cohort

	Cohort 1					Cohort 2				
	N	Control Mean	ITT	Control Complier Mean	TOT	N	Control Mean	ITT	Control Complier Mean	TOT
Math GPA	1707	1.690	0.120**	1.563	0.145**	1361	2.107	0.052	2.054	0.070
			(0.055)		(0.065)			(0.060)		(0.080)
Proportion of Math Courses Failed	1707	0.262	-0.048**	0.293	-0.058**	1361	0.063	-0.014	0.067	-0.019
			(0.020)		(0.023)			(0.012)		(0.016)
Overall GPA	1726	1.919	0.056	1.831	0.067	1503	2.170	0.040	2.111	0.054
			(0.044)		(0.052)			(0.046)		(0.060)
Proportion of Overall Courses Failed	1726	0.206	-0.016	0.226	-0.020	1503	0.050	-0.007	0.053	-0.010
			(0.014)		(0.016)			(0.008)		(0.010)
Non-Math GPA	1724	1.992	0.034	1.921	0.041	1488	2.220	0.025	2.169	0.033
			(0.045)		(0.054)			(0.046)		(0.061)
Proportion of Non-Math Courses Failed	1724	0.188	-0.008	0.204	-0.009	1488	0.048	-0.005	0.050	-0.007
			(0.014)		(0.016)			(0.008)		(0.010)

Notes. Program effects estimated during AY2019-2020 and AY2020-2021 on sample of students who were randomized into Saga Technology in AY2018-2019 and AY2019-2020, respectively. Treatment effects estimated using the following baseline covariates: block, age, race/ethnicity, English language learner status, diverse learner status, socioeconomic status, math GPA, non-math GPA, proportion of courses failed, proportion of math courses failed, proportion of non-math courses failed, standardized math and reading test scores, days absent from school, number of in-school disciplinary incidents, days of out-of-school suspensions, and a set of indicator variables for missing baseline data. GPA is on a 0.00-4.00 point scale using only core classes (Math, Science, Social Studies and English). To construct standardized test scores, we gather individual test scores for the reading and math sections from the Spring test date each year and generate a z-score using the control mean and standard deviation of the scores. Number of days absent include excused and unexcused absences. We measure out-of-school suspensions as the total number of days a student is suspended during the school year.

Table 8: Program Effects on End of Year Outcomes in Year Following Treatment, Controlling for Baseline Covariates, Cohort 2 (AY2019-2020)

	N	Control Mean	ITT	Control Complier Mean	TOT
Math GPA	1434	2.021	0.036	1.954	0.048
			(0.054)		(0.071)
Proportion of Math Courses Failed	1434	0.068	-0.016	0.075	-0.022
			(0.011)		(0.014)
Overall GPA	1543	2.104	0.027	2.051	0.036
			(0.043)		(0.057)
Proportion of Overall Courses Failed	1543	0.055	-0.010	0.060	-0.013
			(0.008)		(0.010)
Non-Math GPA	1536	2.151	0.017	2.104	0.022
			(0.044)		(0.058)
Proportion of Non-Math Courses Failed	1536	0.053	-0.008	0.058	-0.011
			(0.008)		(0.010)
Number of Days Absent	1620	41.825	-2.685	42.205	-3.644
			(1.859)		(2.481)
Number of Disciplinary incidents	1841	0.017	0.011	0.025	0.015
			(0.009)		(0.013)

Notes. Program Effects on End-of-Year Outcomes estimated, during AY2020-2021, on sample of students who were randomized into Saga Technology in AY2019-2020. Treatment effects estimated using the following baseline covariates: randomization block, age, race/ethnicity, English language learner status, diverse learner status, socioeconomic status, math GPA, non-math GPA, proportion of courses failed, proportion of math courses failed, proportion of non-math courses failed, standardized math and reading test scores, number of days absent from school, number of in-school disciplinary incidents, number of days of out-of-school suspensions, and a set of indicator variables for missing baseline data. GPA is on a 0.00-4.00 point scale using only core classes (Math, Science, Social Studies and English). To construct standardized test scores, we gather individual test scores for the reading and math sections from the Spring test date each year and generate a z-score using the control mean and standard deviation of the scores. Number of days absent include excused and unexcused absences. We measure out-of-school suspensions as the total number of days a student is suspended during the school year. Missing baseline data has been imputed using mean values of the control group.

Table 9: Heterogeneous Treatment Effects on EOY Math Outcomes with Baseline Covariates, Cohort 1

	Math Test scores		Math GPA		Prop. of Math Failures	
	ITT	TOT	ITT	TOT	ITT	TOT
Gender						
Girls	0.191***	0.228***	0.191***	0.226***	-0.043**	-0.050**
	(0.044)	(0.052)	(0.063)	(0.072)	(0.019)	(0.021)
Boys	0.178***	0.221***	0.201***	0.249***	-0.033	-0.041
	(0.054)	(0.064)	(0.068)	(0.082)	(0.021)	(0.025)
T-test (Girls/Boys)	0.861	0.932	0.918	0.834	0.730	0.774
Race/Ethnicity						
Black	0.143*	0.157*	0.101	0.115	0.020	0.022
	(0.077)	(0.081)	(0.096)	(0.104)	(0.029)	(0.032)
Latinx	0.224***	0.276***	0.265***	0.320***	-0.055***	-0.067***
	(0.046)	(0.053)	(0.063)	(0.073)	(0.019)	(0.022)
Other	0.138*	0.184*	0.063	0.085	-0.034	-0.046
	(0.076)	(0.097)	(0.096)	(0.123)	(0.027)	(0.034)
T-test (Latinx/Black)	0.371	0.221	0.154	0.106	0.034**	0.023**
Baseline Math GPA						
Bottom Quartile (Q1)	0.124	0.149	0.166*	0.197*	-0.022	-0.026
	(0.093)	(0.105)	(0.092)	(0.103)	(0.039)	(0.044)
Quartile 2	0.269***	0.319***	0.305***	0.360***	-0.078***	-0.092***
	(0.064)	(0.072)	(0.088)	(0.098)	(0.026)	(0.029)
Quartile 3	0.141**	0.170**	0.150	0.180*	-0.011	-0.013
	(0.062)	(0.070)	(0.093)	(0.105)	(0.020)	(0.023)
Top Quartile (Q4)	0.227***	0.298***	0.041	0.054	-0.006	-0.008
	(0.074)	(0.090)	(0.104)	(0.127)	(0.011)	(0.014)
T-test (Q 1/2)	0.201	0.185	0.271	0.252	0.235	0.214
T-test (Q 1/3)	0.876	0.874	0.907	0.909	0.794	0.784
T-test (Q 1/4)	0.386	0.284	0.370	0.383	0.702	0.699
Joint F-test (Q 1,2,3,4)	0.420	0.339	0.268	0.269	0.091*	0.074*
Baseline Math Test Scores						
Bottom Quartile (Q1)	0.160**	0.195**	0.239***	0.295***	-0.049	-0.061
	(0.076)	(0.087)	(0.092)	(0.108)	(0.036)	(0.042)
Quartile 2	0.190***	0.225***	0.203**	0.243**	-0.045	-0.054*
	(0.072)	(0.079)	(0.099)	(0.112)	(0.029)	(0.033)
Quartile 3	0.182**	0.212**	0.159	0.179	-0.008	-0.009
	(0.080)	(0.087)	(0.111)	(0.117)	(0.025)	(0.027)
Top Quartile (Q4)	0.188***	0.244***	0.127	0.161	-0.037*	-0.046*
	(0.056)	(0.069)	(0.088)	(0.105)	(0.021)	(0.025)
T-test (Q 1/2)	0.772	0.795	0.789	0.738	0.934	0.904
T-test (Q 1/3)	0.841	0.889	0.579	0.465	0.355	0.305
T-test (Q 1/4)	0.765	0.658	0.382	0.374	0.764	0.770
Joint F-test (Q 1,2,3,4)	0.991	0.975	0.837	0.810	0.716	0.619

Notes. Quartiles 1-4 mentioned under Baseline Math GPA are quartiles created using students' baseline end-of-year (EOY) math GPA. Quartiles 1-4 mentioned under EOY Math Test Scores are quartiles created using students' baseline EOY standardized math test scores. The value displayed for T-test and F-tests are p-values. T-test (Q1/2), T-test (Q1/3), and T-test (Q1/4) signify T-tests between Bottom Quartile (1) and 2, Bottom Quartile (1) and 3, and Bottom Quartile (1) and 4, respectively. *** p<0.01, ** p<0.05, * p <0.1

Table 10: ALEKS Usage, Cohort 1

	N	Observed in ALEKS data (%)	ALEKS Usage	Mean	Median	25th %tile	75th %tile
Treatment Group							
Participants	815	97.06	Time Spent (hrs)	29.28	32.87	18.72	40.10
			# Topics Attempted	129.04	111.00	55.00	189.00
			# Topics Learned	96.60	80.00	36.00	138.00
			Time Spent per Week (hrs)	0.84	0.94	0.53	1.15
Non-Participants	213	0.94	Time Spent (hrs)	0.00	0.00	0.00	0.00
			# Topics Attempted	0.00	0.00	0.00	0.00
			# Topics Learned	0.00	0.00	0.00	0.00
			Time Spent per Week (hrs)	0.00	0.00	0.00	0.00
Control Group							
Participants	1	100.00	Time Spent (hrs)	1.43	1.43	1.43	1.43
			# Topics Attempted	7.00	7.00	7.00	7.00
			# Topics Learned	7.00	7.00	7.00	7.00
			Time Spent per Week (hrs)	0.04	0.04	0.04	0.04
Non-Participants	976	0.00	Time Spent (hrs)	.	.	.	.
			# Topics Attempted	.	.	.	.
			# Topics Learned	.	.	.	.
			Time Spent per Week (hrs)	.	.	.	.

Notes. N signifies the number of students assigned to treatment (or control) and who took up (or didn't take up) treatment. % Observed in ALEKS represents the percentage of students who had an account on ALEKS. This variable is comprised of students who spent non-zero time on ALEKS as well as students who spent zero time on the platform. We measure ALEKS usage in terms of time spent on ALEKS in hours, total topics learned on ALEKS, and total topics attempted on ALEKS. ALEKS defines "topics attempted" as the number of topics that a student has attempted to learn, but not yet successfully completed. Moreover, ALEKS defines a "topics learned" as when a student achieves a total of 5 points per topic, where a student receives one point for each correct answer, and one point subtracted for each incorrect answer. We calculate number of weeks after excluding holidays and weekends between the start and end date of tutoring. Cohort 1 received 35 weeks of tutoring on average.

Table 11: Baseline Characteristics that Predict ALEKS Usage, Cohort 1 (AY2018-2019)

	Total Time (hrs)	Total Topics Attempted	Total Topics Learned
Age	-0.743	-5.291	-4.995
	(0.863)	(4.686)	(3.547)
Diverse Learner	-1.617	-8.573	-9.053*
	(1.428)	(7.562)	(5.476)
Free/Reduced Price Lunch	1.445	9.853	9.749
	(1.660)	(9.318)	(6.944)
English Learner	-0.982	-0.017	1.437
	(1.898)	(11.164)	(8.782)
Black	1.808	7.109	3.174
	(1.705)	(9.391)	(7.269)
Latinx	1.068	3.108	0.189
	(1.570)	(9.259)	(7.373)
Number of Absences	-0.274***	-1.046***	-0.758***
	(0.044)	(0.239)	(0.183)
Number of Days Suspended (Out of School)	-0.002	-0.017	0.072
	(0.099)	(0.395)	(0.284)
Number of Disciplinary Incidents	-1.781***	-7.820***	-5.952***
	(0.381)	(1.778)	(1.319)
Math GPA	1.861**	11.817**	10.316***
	(0.849)	(4.788)	(3.720)
Non-Math GPA	0.947	6.940	4.583
	(1.000)	(5.728)	(4.446)
Standardized Math Test Score	-1.264	-2.212	2.317
	(0.775)	(4.096)	(3.225)
Standardized Reading Test Score	1.623**	7.073*	6.257*
	(0.699)	(4.188)	(3.281)
Proportion of Fs in Math Courses	9.147	55.675	49.396
	(17.553)	(101.295)	(75.620)
Proportion of Fs in Non-Math Courses	4.996	77.589	95.917
	(59.369)	(339.829)	(251.004)
Proportion of Fs	-14.335	-128.269	-133.873
	(76.456)	(438.725)	(324.768)
N	1,028	1,028	1,028

Notes. Baseline Characteristics of Treatment Students that Predict Higher ALEKS Usage. We regress different measures of ALEKS usage on baseline covariates, their corresponding missingness indicators, and randomization blocks. We measure ALEKS usage in terms of time spent on ALEKS in hours, total topics learned on ALEKS, and total topics attempted on ALEKS. ALEKS considers a topic learned when a student earns 5 points per topic, where a student receives one point for each correct answer, and one point is subtracted for each incorrect answer. *** p<0.01, ** p<0.05, * p <0.1

Table 12: 2SLS Estimates of the Effect of CAL Usage on End-of-Year Math Outcomes, Cohort 1 (AY2018-19)

Dependent Variable: Math Test Score			
	OLS	2SLS	2SLS
Total Time spent on ALEKS (hr.)	0.010***	0.007***	0.007***
	(0.0014)	(0.0013)	(0.0018)
Total Topics Attempted	0.002***	0.002***	0.001***
	(0.0003)	(0.0003)	(0.0004)
Total Topics Learned	0.003***	0.002***	0.002***
	(0.0003)	(0.0004)	(0.0005)
N	1679	1679	1679
Control for Baseline Covariates	No	No	Yes
Dependent Variable: Math GPA			
	OLS	2SLS	2SLS
Total Time spent on ALEKS (hr.)	0.017***	0.007***	0.008***
	(0.0017)	(0.0018)	(0.0024)
Total Topics Attempted	0.004***	0.002***	0.002***
	(0.0003)	(0.0004)	(0.0005)
Total Topics Learned	0.005***	0.003***	0.002***
	(0.0004)	(0.0005)	(0.0007)
N	1846	1846	1846
Control for Baseline Covariates	No	No	Yes

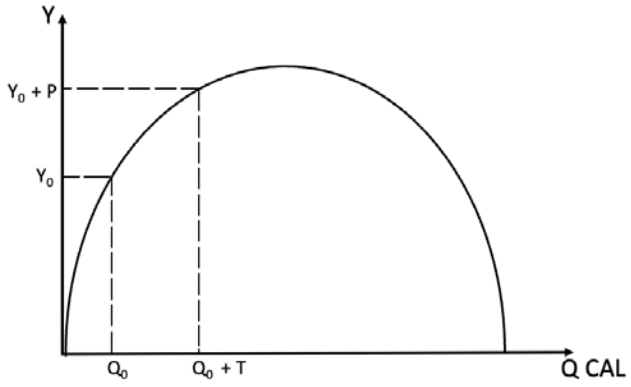
Notes. Estimating the relationship between ALEKS usage and end-of-year (EOY) Math outcomes using the Ordinary Least Squares (OLS) and Two Stage Least Squares (2SLS) model. In the OLS regression, we estimate the correlation between EOY math outcomes and ALEKS usage without controlling for randomization blocks and baseline covariates. We perform two 2SLS regressions, one with controlling for baseline controls and one without. We use the interaction between treatment status and randomization block indicators as instrument for the interaction between ALEKS usage and treatment status. We measure ALEKS usage in terms of Time spent on ALEKS in hours, number of topics attempted, and number of topics learned on ALEKS. ALEKS defines "topics attempted" as the number of topics that a student has attempted to learn. ALEKS considers a topic learned when a student earns 5 points per topic, where a student receives one point for each correct answer, and one point subtracted for each incorrect answer. For students who didn't use ALEKS, we impute value 0 in total time spent and total topics learned on ALEKS. We control for the following baseline covariates: age, race/ethnicity, English language learner status, diverse learner status, socioeconomic status, math GPA, non-math GPA, proportion of courses failed, proportion of math courses failed, proportion of non-math courses failed, standardized math and reading test scores, number of days absent from school, number of in-school disciplinary incidents, number of days of out-of-school suspensions, and a set of indicator variables for missing baseline data. GPA is on a 0.00-4.00 point scale using only core classes (Math, Science, Social Studies and English). To construct standardized test scores, we gather individual test scores for the reading and math sections from the Spring test date each year and generate a z-score using the control mean and standard deviation of the scores. Number of days absent include excused and unexcused absences. We measure out-of-school suspensions as the total number of days a student is suspended during the school year. Missing baseline data has been imputed using mean values of the control group. *** p<0.01, ** p<0.05, * p <0.1

Table 13: Average Total and Variable Costs of Program

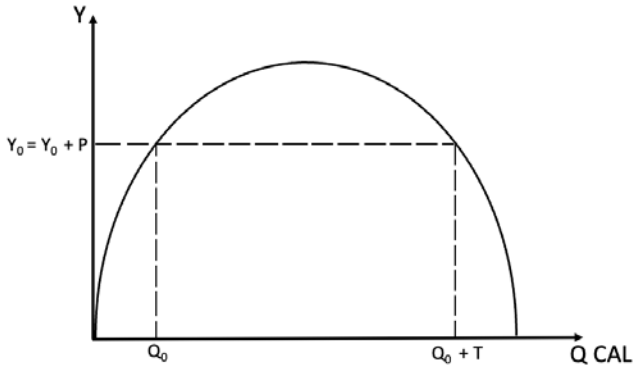
Input	Year One	Year Two	Total
A. Program Size			
Student Capacity	1028	1032	2060
Participants	815	771	1586
Tutors	35	37	72
Student/Tutor	23.29	20.84	22.03
Schools	6	7	13
B. Costs			
Total Cost	$2,107,000	$2,107,000	$4,214,000
Total Variable Cost	$1,947,000	$ 1,947,000	$3,894,000
C. Average Total Cost			
Per Treatment Slot	$2,049.61	$2,041.67	$2,045.63
Per Participant	$2,585	$2,732.81	$2,657.00
D. Average Variable Cost			
Per Treatment Slot	$1,893.97	$1,886.63	$1,890.29
Per Participant	$2,388.96	$2,525.29	$2,455.23

Notes. This table shows how we calculate average program costs. Panel A summarizes program size in each year. Panel B refers to cost implied in A18. Panel C and D use information from Panel A and B to calculate average total and variable cost per treatment slot and per participant.

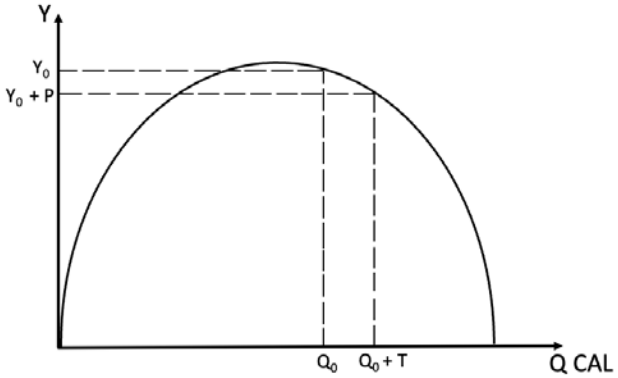
Figure 1: Can Technology be Productively Incorporated into Tutoring?



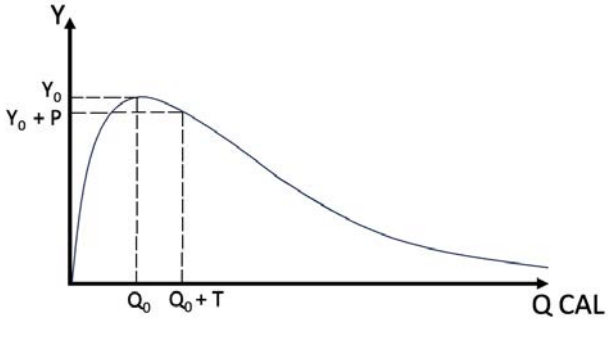
(a) Best-Case Scenario



(b) Bad Scenario: Adding too much tech into Tutoring



(c) Bad Scenario: Schools already doing a lot of Tech



(d) Bad Scenario: Kids having limited tolerance for Tech

Notes. Y denotes student learning, Q is time spent on CAL, Q_0 and Y_0 are baseline levels of CAL usage in schools and baseline level of students learning, respectively, T is incremental addition of CAL usage in schools, and P is change in students learning due to increase in CAL usage

Figure 2: Total Time Spent on ALEKS, by Cohort 1, in AY2018-2019

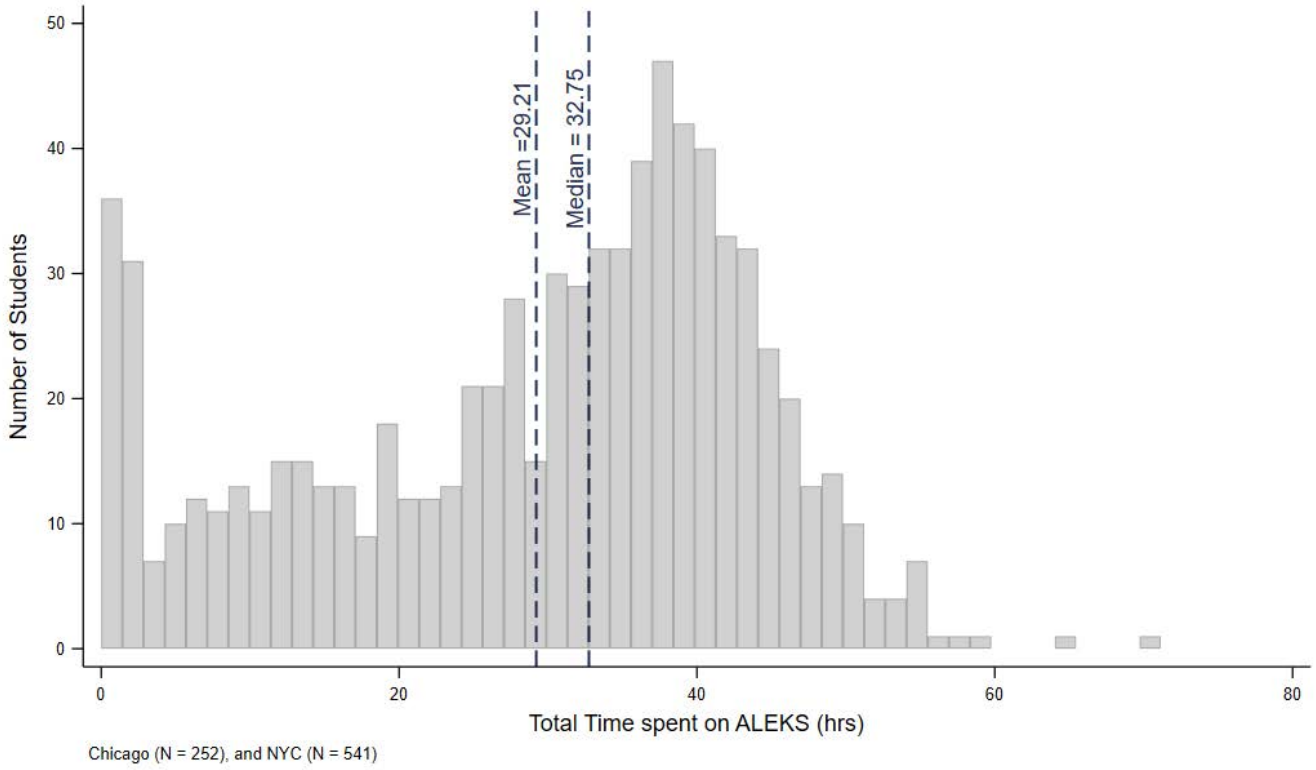


Figure 3: Total Topics Attempted on ALEKS, by Cohort 1, in AY2018-2019

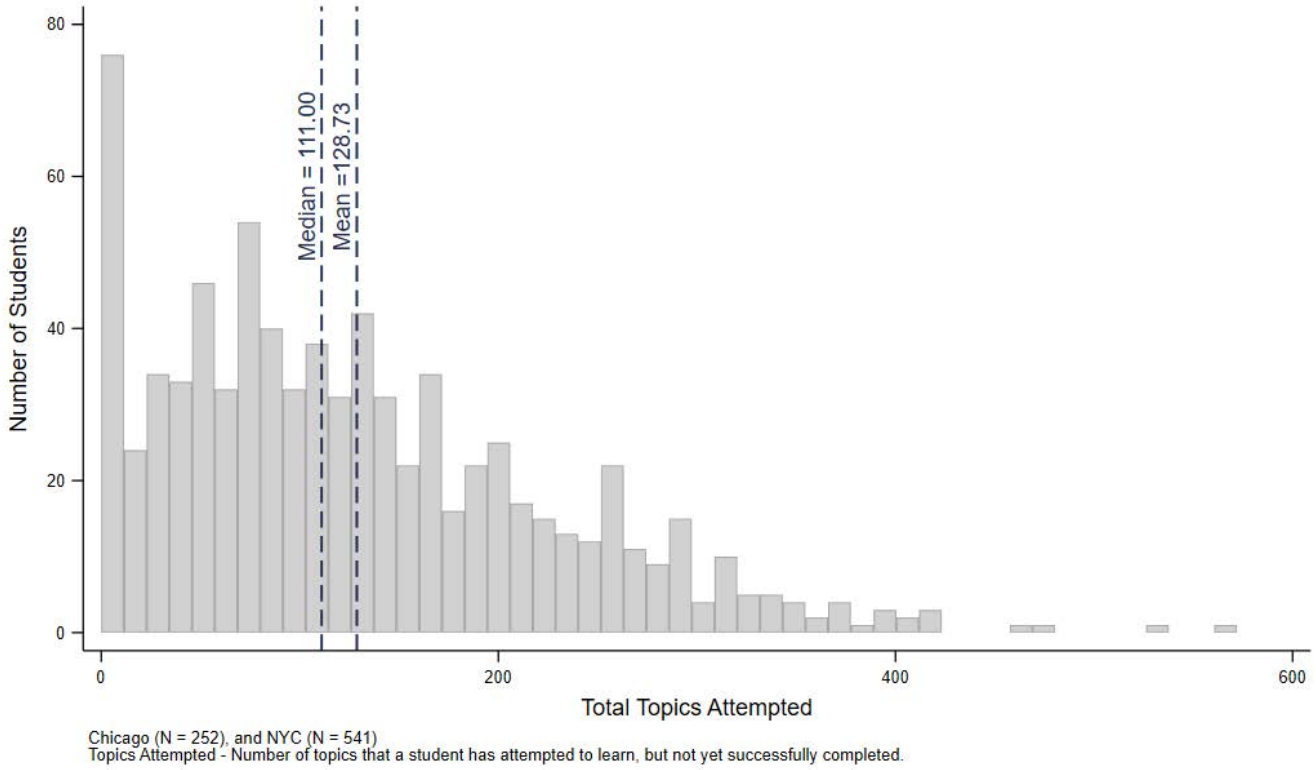
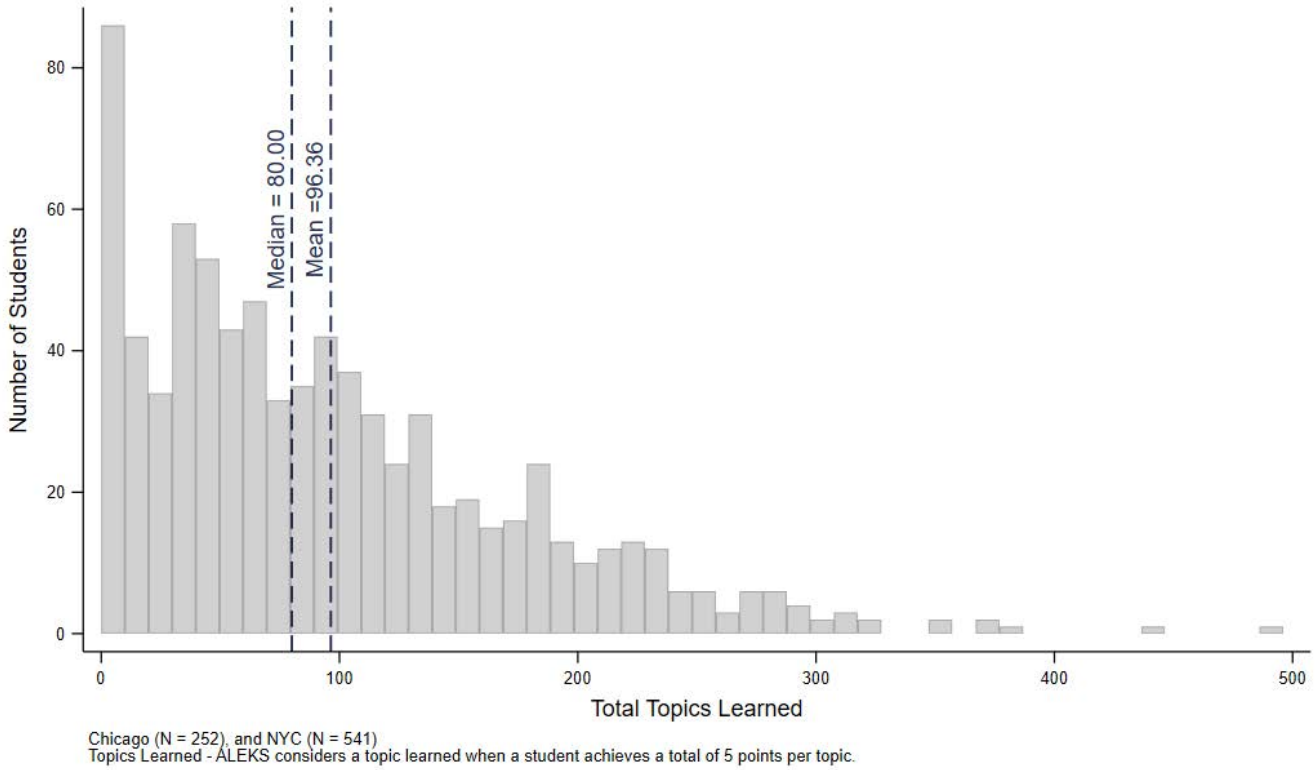


Figure 4: Total Topics Learned on ALEKS, by Cohort 1, in AY2018-2019



Chicago (N = 252), and NYC (N = 541)
Topics Learned - ALEKS considers a topic learned when a student achieves a total of 5 points per topic.