Teaching Teachers To Use Computer Assisted Learning Effectively: Experimental and Quasi-Experimental Evidence

Philip Oreopoulos, University of Toronto, NBER, J-PAL
Chloe Gibbs, University of Notre Dame
Michael Jensen, University of Notre Dame
Joseph Price, Brigham Young University, NBER

Abstract: Mastery learning – the process by which students must demonstrate proficiency with a single topic before moving on – is well recognized as one of the best ways to learn, yet many teachers struggle or remain unsure about how to implement it into a classroom setting. This study leverages two field experiments to test the efficacy of a program designed to encourage greater mastery learning through technology and proactive continuous teacher support. Focusing on elementary and middle school mathematics, teachers receive weekly coaching in how to use Computer Assisted Learning (CAL) for students to follow a customized roadmap of incremental progress. Results indicate significant intent-to-treat effects on math performance of 0.12-0.22 standard deviations. Further analysis shows that these gains are concentrated among students in classrooms with at least an average of 35 minutes of practice per week. Teachers able to achieve high-dosage practice have a high degree of initial buy-in, a clear implementation strategy for when practice occurs, and a willingness to closely monitor progress and follow-up with struggling students.

Key Words: Computer Assisted Learning; Khan Academy; Personalized Instruction; Randomized Controlled Trial

Acknowledgements: This research was supported through the Social Sciences and Humanities Research Council of Canada (Insights Grant #20010045), America Achieves, the Jameel Poverty Action Lab, and the Overdeck Family Foundation. We are especially grateful to Kelli Hill from Khan Academy, Barry Fox and many other administrators from the Arlington Independent School District and to Renita Perry, Sharon Griffin, Keri Randolph, and Healthier McMillin from the Metro Nashville Public Schools district for their support and partnership. We also thank Nina Low, Dhanya Ashley, Dellannia Segreti, Anna Staud, and Brendan Perry for project management and research assistance, as well as the dozens of khoaches and hundreds of teachers who worked together with us to implement the KWiK program. Participants at numerous seminars and workshops provided helpful suggestions. Any errors or omissions are our own.

I. Introduction

A fundamental challenge in education is the fact that students are different. Students do not arrive at the start of a grade performing at the same level, and they do not progress at the same pace. In a classroom setting, teachers cannot easily provide individualized attention and, as a result, some students move onto other topics before clearly grasping earlier ones, while other students could be learning more advanced topics faster. Without mastering critical early topics and skills, foundations in subjects like mathematics, science, and reading can become weak. Students may become disengaged, discouraged, and resolved at being poor students, not because they lack the potential to excel, but because they were not given the opportunity to persist trying to learn material they did not initially understand.

One solution to this challenge would be to reduce class size to one. Students with their own private teacher would be able to progress at their own pace and teachers to provide continuous feedback, respond to questions, and address individual needs. This mastery approach was used in the nineteenth century by Oxford and Cambridge universities for teaching a small fraction of undergraduates (e.g., Beck 2007) and is used currently for teaching a small fraction of students at home (e.g., Ray 1988). But one-on-one instruction is too costly and impractical to scale. Willingness to pay is not enough to adopt this approach for every child.

An alternative solution is to supplement classroom instruction with one-on-one or small group tutoring. Many studies demonstrate large benefits from providing regular tutoring. For example, a meta-analysis examining 96 high-dosage tutoring experiments in the last forty years found an average standardized test score effect size of 0.29 (Nickow et al. 2024). The consistency in finding large assessment gains is remarkable given the variety of tutoring programs examined (by subject, tutor type, grade, implementation), suggesting that tutoring is one of the most reliably effective tools education administrators have for improving learning outcomes. Unfortunately, tutoring is also expensive, often costing thousands of dollars per student per year. Schools may only be able to offer tutoring to the students in most dire need. Even with more money to support it, tutoring is not easy to implement. Schools face difficulty finding enough tutors. During school, administrators face difficulty determining when to provide it, and if it is offered after school, few students offered tutoring actually show up to receive it (White et al. 2023).

As a further alternative, Computer Assisted Learning (CAL) has the potential to assist scaling personalized learning. For example, one such platform, Khan Academy (KA), allows students to progress through Grades 3-12 mathematics topics incrementally. Students watch short videos, work through short exercises, receive immediate feedback, review mistakes, and try again until answering all questions correctly. Teachers (and parents) can observe progress and intervene with further assistance when students require additional interactive support. While administratively cheaper, CAL faces its own challenges in getting students to use it effectively. Students are not generally motivated to work through exercises on their own (Beg et al. 2022). Integrating CAL in the classroom often requires that teachers alter the method they deliver instruction and become more proactive in monitoring progress and providing individualized assistance. Sufficient curriculum time is required for practice to occur. Teachers may be hesitant about the additional effort and time costs to learn how to implement CAL, skeptical to adopt it compared to using their previous year's curriculum, or feel too overwhelmed and busy.

A review found that 12 of 19 randomized controlled trials of CAL in the last several years had substantial impact, ranging from 0.14 to 0.56 standard deviations in test score improvement by the end of the school year (Escueta et al. 2020). The overall average effect size was 0.18. Quality of implementation is an important factor determining variation in these outcomes (Pane et al. 2010, Pane et al. 2014, Ritter et al. 2007), with teachers playing a critical role in ensuring effective time use (Fancsali et al. 2016). Facilitation of practice, monitoring of progress, and intervention with students making slow progress are key.

In this paper, we propose a new solution for improving implementation quality which involves providing weekly proactive guidance to teachers for training them how to use CAL effectively. The main ingredient of the Khoaching with Khan Academy program (KWiK) is to provide teachers with assistants (we call them khoaches) to guide them in following a default recipe for using CAL as part of their curriculum. Khoaches meet one-on-one with teachers at the start of the year, and then weekly to help set goals, troubleshoot issues, and discuss best practices. They also help monitor and interpret data, set up weekly assignments, and identify students needing extra attention. The recipe can be adjusted, depending on teacher's preferences and existing curriculum constraints, with the main goal to generate adequate practice time and progress using CAL for every student. Training teachers to use CAL in a classroom setting provides a potentially more cost-effective approach for facilitating personalized learning. The

approach is highly scalable, leveraging mostly existing resources and the school curriculum for helping motivate students to practice. Khoaching costs are one-time. Their scaffolding can be taken down once teachers feel comfortable in managing CAL in their classrooms on their own.

We present three pieces of evidence showing that KWiK significantly boosts mathematics achievement scores for elementary and middle school students, but only when teachers facilitate sufficient CAL practice time. Our first experiment involves training 47 Metro Nashville Public School teachers to deliver a week-long CAL review exercise in preparation for a state standardized test. Students in Grades 6 to 8 are randomized into working on one of two topics before taking a quiz seven days later, allowing us to also test for differential impacts across classrooms with different degrees of average CAL practice time. Test score gains were insignificant for students in classrooms with less than 5 minutes of average practice, but 0.27 standard deviations higher in ones with more than 50 minutes of average practice. Our second experiment randomizes 216 teachers from the Arlington Independent School District to receive a khoach during one school year. The grade 3 to 6 curriculum allows for more time and flexibility to teach math compared to the grade 7 to 8 curriculum. Correspondingly, the experiment's intent-to-treat effect in Grades 3-6 classrooms was 0.17, where practice time averaged more than 25 minutes each week, while not statistically significant effects were found for grade 7-8 classrooms, where practice time averaged only 4 minutes. Lastly, we use plausibly exogenous assignment of observationally similar students within the same school to classrooms with treated teachers who facilitated different amounts of average CAL practice time. We find the same pattern of greater test score gains for students in classrooms with greater average CAL practice times.

A key conclusion from this paper concerns the importance of teacher buy-in and implementation fidelity. Based on teacher descriptive data and qualitative interviews, variation across classrooms in average CAL practice appears mostly associated with differences in teacher commitment to fidelity and whether teachers treated the CAL practice as an optional part of the curriculum. Classrooms with the highest practice time were often with teachers that set high expectations for exercise completion, monitored student progress and followed-up with those falling behind. In classrooms where little CAL practice occurred, we found little or no gains to performance. But when teachers were able to provide at least 25 minutes of practice each week, students saw gains comparable to high-dosage tutoring programs.

Our study makes four important contributions. First, it demonstrates the effectiveness of a program that leverages (mostly) existing resources to facilitate more personalized learning and immediate feedback. Once teachers become familiar with the KWiK setup (after receiving guidance from khoaches), they become equipped with the knowledge for how to implement the program in their classrooms going forward at zero additional financial cost. Second, no previous field experiment has tested the effectiveness of Khan Academy, one of the largest CAL platforms in the world, in a developed-country setting. Third, the study highlights how Intent To Treat effects depend critically on implementation fidelity, context, and training. Our experimental and quasi-experimental results suggest a large gradient in program effectiveness by CAL practice time. Average effects are watered down from treated teachers not facilitating sufficient dosage time. If these teachers could be motivated or incentivized to help students practice more, the program's effectiveness could be substantially higher. As such, CAL effectiveness depends on implementation quality more than it does on the CAL platform itself. Finally, the study provides insight about why some teachers facilitate more CAL practice than others using teacher surveys and qualitative interviews. Key factors for sufficiently high-practice time include principal support, focus on one CAL platform being used in classrooms, teacher buy-in to the program's theory of change, regular monitoring and follow-up of struggling students, and treating the program as a mandatory and important rather than optional and bonus material.

The rest of the paper is structured as follows: in section two, we discuss key components of the KWiK program and situate it in context with previous literature. In section three, we present the week-long Nashville experiment, its implementation details, and results. In section four, we cover the setup and results from the full-year Arlington experiment. Section five presents quasi-experimental evidence of students with similar test scores matched to classrooms with different average practice times. Using survey data and qualitative interviews, section six explores why some teachers implemented the program more successfully than others. Section seven concludes with a discussion about scaling KWiK and the potential for using artificially intelligent khoaches.

II. Background

A. Previous Relevant Studies

Many computer assisted learning platforms have been designed to help develop particular math skills, only some of which have been tested experimentally for effectiveness.[1] ASSISTments is one that has teachers find and assign math questions as homework aligned to class instruction. Students work on assignments online and receive immediate feedback and support. Teachers can monitor individual and class progress, and are encouraged to review questions that a significant fraction of students answered incorrectly.

Mendicino et al. (2009) ran a small-scale experiment with 28 Grade 5 students with internet access across four classrooms. Each class worked on two homework assignments over one week, randomly using ASSISTments for one assignment and pencil and paper for the other (problems assigned were identical). Test score gains were 0.61 standard deviations higher when working with ASSISTments, though the study conditions on students that at least started using the online platform. Kehrer et al. (2013) had 61 Grade 7 students doing two homework assignments, one with ASSISTments and receiving immediate feedback versus one with a worksheet and receiving the same feedback the day later. When using ASSISTments, test scores were 0.53 standard deviations higher.

While impressive, it is important to note that these studies test for effectiveness of CAL over a short period of time (about one week) and look at outcomes closely aligned with practice material soon after practice has occurred. Other CAL experiments with short durations find similarly large effects (e.g., Roschelle et al. 2010 and Wang and Woodworth 2011). Implementation challenges over one week, however, are likely fewer compared to those arising from implementing CAL over the entire school year, and short-term knowledge gains may not translate to longer-term gains.

In contrast, two experiments using ASSISTments were conducted over the entire school year. Roschelle et al. (2016) randomized 43 schools in Maine to receive training and support for implementing ASSISTments to selected Grade 7 classrooms either during year 1 and 2 of the study (when the experiment occurred) or during year 3 and 4. Test scores in year 2 (year 1 was focused on training) were 0.18 standard deviations higher for the ASSISTments schools. Finally,

[1] See also Escueta et al. (2020) for a further discussion and meta-analysis of CAL experiments in developed countries and Major et al. (2021) for a meta-analysis of CAL experiments in developing countries. Ran et al. (2022) provide a meta-analysis of non-experimental CAL studies.

Feng et al. (2023) conducted a similar experiment in North Carolina among 63 paired schools. Students in ASSISTments classrooms performed, on average, 0.10 standard deviations better a year later than those in control classrooms (researchers were unable to obtain test scores during the year of the actual experiment due to the COVID-19 pandemic).

Another program called Cognitive Tutor (sometimes known as Mathia) includes an entire course curriculum for teachers to follow. Students use it to work individually on math exercises on their computers about two days a week and receive related group-based activities and teacher instruction during the remaining days. Morgan and Ritter (2002) find that Grade 9 students at four schools randomly in classrooms assigned to Cognitive Tutor did 0.29 standard deviations better on an end-of-year algebra assessment than students assigned to teachers practicing business-as-usual classrooms. Pane et al. (2010), on the other hand, estimate a negative impact for high school students assigned. About 2,000 geometry students in Baltimore high schools were randomly assigned to sections in which teachers either taught using Cognitive Tutor or not. Of those who persisted to the end of the year and completed a posttest, students in CAL classrooms did 0.19 standard deviations worse than students in business-as-usual classrooms. The researchers cited many implementation challenges, including teachers having difficulty implementing the group-based lessons, students not completing CAL exercises at the same time material was being covered in class, and students progressing too slowly to cover material on the test. In a follow up experiment, Pane et al. (2014) randomly assigned 73 high schools and 74 middle schools to use a regular algebra curriculum or one with Cognitive Tutor, allowing teachers to interact more freely with colleagues. Effect size estimates were imprecise, but after conditioning on past performance, treated students had test score gains about 0.20 standard deviations higher than control schools for both middle and high schools.

A couple of other CAL platforms have been experimentally tested in the United States. Barrow et al. (2009) examined the Interactive Computer Aided Natural Learning (I Can Learn) program in three urban school districts. The software was designed to provide one-to-one instruction, emphasizing a mastery approach where students online take a pretest, watch an instructional video, then work on problems until demonstrating sufficient comprehension. Teachers were asked to provide targeted help to students not progressing satisfactorily. Classes were randomized to receive instruction for teaching with “I Can Learn” or business as usual. Fidelity was generally high, with 65 percent of students completing the expected number of

computer lessons over the school year. Students in the treated group scored 0.17 standard deviations higher on state test scores than did students in the control group, with larger effects for students with lower pretest scores. Copeland et al (2023) examine IXL, another CAL platform providing videos and interactive math problems. Twenty-five Grades 3-5 teachers from the Holland Public School district in Michigan, within the same grade and school, were randomized to receive training to implement IXL or continue with their regular curriculum. End of year math scores were 0.13 standard deviations higher for students with treated teachers compared to control teachers. Effects were higher among Grade 3 students, corresponding to higher practice time throughout the year. Most teachers reported being only somewhat prepared to integrate CAL into their classrooms. Phillips et al. (2020) test the impact of training high school teachers to use ALEKS, a CAL platform for providing a sequential set of online exercises for students to work on towards mastery. Approximately 2,500 students were randomly assigned to use ALEKS as a supplement to the district's algebra curriculum or a control group that used the regular curriculum. No differences in end of year test scores were found. Based on a pretest, students were poorly prepared entering the course. Teachers expressed a tension between wanting to allocate more remedial CAL practice time and wanting to keep up with regular classroom material. The researchers concluded that teachers sacrificed CAL practice over regular content material.

B. Khan Academy

The CAL platform we utilize in this study is Khan Academy (KA), one of the most popular and most recognized around the world. In 2004, Sal Khan began remotely tutoring his cousin – first by phone and an online whiteboard, then by posting online videos. The videos were kept public and 'went viral' from students seeking help with math and other subjects. Khan responded by creating a non-profit organization in 2008 and left his job as a hedge fund analyst the next year to focus on expanding his CAL program. KA now has more than 15 million monthly visitors from 190 countries. Its weekly usage more than doubled after the start of the COVID-19 pandemic. It offers more than 30 courses, including math instruction from preschool to university-level calculus, and over 150,000 interactive exercises. Any student, teacher, or parent can access the website for free.

A key feature of KA is its positive user experience. Much attention and resources have gone into making the program intuitive, interactive, and progressive. Students can watch and re-watch videos, practice, take quizzes, receive hints when they cannot solve a question, and receive recommendations for whether to advance to a new topic or continue to improve on a current one. Another feature is the ability to link student accounts to parents and teachers. Teachers can choose specific content and exercises that align with their own curriculum, assign these to students, monitor progress using a detailed dashboard, offer feedback, and grade performance. The website allows flexibility for how users want to use its content. Teachers can adopt Khan Academy either as a homework tool, an in-class supplement, or a review tool. Parents can also monitor progress. These features and efforts to continuously improve content (all free) have helped make Khan Academy one of the best-known online learning platforms in North America and the world.

Evidence on KA as a tool for improving academic achievement is surprisingly sparse.[2] Test score gains are correlated with KA practice time: students who practiced on KA for at least 30 minutes a week had gains 0.26 standard deviations higher, on average, than observationally similar students who had KA access but used it for fewer than 15 minutes per week (Weatherholtz et al. 2022). But even if this were causal, the relationship does not help understand from a policy perspective how to facilitate greater KA practice. Snipes et al. (2015) examine an intensive 4-week summer math intervention that included an hour of KA practice a day and three additional hours of other instruction and support, but the large gains found from the program cannot necessarily be attributable to KA practice alone, since the comparison group receives no program at all. Ferman et al. (2021) report results from a randomized controlled trial in Brazil, in which treated teachers took their students to the school's computer lab to practice KA once a week (for 50 minutes) instead of their standard math classes. They find no impact on end-of-year math scores, but offer evidence to conclude that the program had negative effects for students in classrooms that faced computer access challenges and positive effects for those without. Büchel et al (2022) conduct an experiment in El Salvador in which treated primary students receive supplemental math instruction either with a teacher without using KA (or any CAL), a teacher using KA, or a technical assistant supervising the use of KA. All students did

[2] Most KA studies, experimental or otherwise, are linked on the platform's website: https://support.khanacademy.org/hc/en-us/articles/17154113319437-What-efficacy-studies-does-Khan-Academy-have

better than the control group. Treated students using KA performed about 0.08 standard deviations than treated students with just a teacher.

The Bill & Melinda Gates Foundation contracted with SRI International with the intent to conduct a large US evaluation of Khan Academy in 2011. That effort, however, led to a report that focused instead on implementation (Murphy et al., 2014). Nine California schools participated in the pilot. Teachers chose how they wished to adopt the platform. Most expressed appreciation for having a tool for greater personalized instruction, but preferred to use it primarily as an occasional supplement to core teacher instruction. Videos were used mostly at the discretion of students. Practice time varied widely across sites, from an average of about 11 minutes per week to an average of 90 minutes per week. Few teachers expected KA to be used as homework.

C. Khoaching With Khan Academy Principles

The studies mentioned above suggest that CAL effectiveness may depend critically on program fidelity and dosage. The studies with the smallest or even negative effects were all ones where researchers reported implementation challenges. The studies with the largest effects were all short-term experiments of CAL usage over about a week. Specific topics were covered with specific instructions for leveraging the software. Teachers had less flexibility to decide how to teach the topics compared to the full-year experiments. Large average effects may still fail to reveal important differences across classrooms and students because of differences in CAL dosage. Key unresolved questions remain around the optimal CAL setup, duration, dosage, and context, as well as how to motivate administrators, teachers and students towards these optimums. Fancsali et al (2016) investigates these issues by examining implementation data from a CAL experiment and surveying teachers. They conclude that "learner efficiency is driven by whether teachers take an active role in turning engaged time into academic learning time by cognitively, behaviorally, and effectively supporting students and encouraging students to mindfully be "on-task", avoid waiting time and behavior that does not enhance learning."

To assist teachers taking more of an active role in facilitating high-dosage CAL, the goal in the present study is to provide more proactive and personalized scaffolding to teachers for facilitating at least hour of practice for each student each week, while leaving the program

flexible enough to allow for teacher adjustments for their specific classroom needs. We try to 'simplify the recipe' for implementing effective CAL by providing a "roadmap" of Khan Academy videos and exercises to incrementally follow. The activities proceed in the same order as the assigned district curriculum, but, importantly, they can be approached at a pace directed by the student's mastery. Teaching assistants (khoaches) proactively collaborate with teachers to simplify instructions for setting up the roadmap and help make adjustments, if needed, to individual students who would benefit from practicing on earlier grade materials. Khoaches advise teachers to provide sufficient time for students to work through material, make mistakes, and try again. The target level of practice each week is at least an hour, divided as needed among in-school and after-school time.

Khoaches meet weekly with teachers, initially by zoom, and check in more by email later in the year as the teacher becomes more familiar with the program. They help set goals, troubleshoot issues, and discuss best practices. They also monitor and interpret classroom data, and help teachers identify students who are struggling and need additional support. Khoaches also suggest that teachers demonstrate at least one activity in class with students each week. During in class practice time, teachers are asked to monitor progress and focus on helping students struggling the most in completing exercises. A teacher dashboard allows real time tracking of student progress and practice levels.

III. Nashville Experiment

A. Setup

We conducted a pilot experiment with the Metro Nashville Public School (MNPS) district in March and April of 2022 to test the impact of following the KWiK approach over a one-week period. Participants were students from grades 6 through 8 attending seven high-need, low performing schools. Prior to statewide testing in Spring 2022, we worked with administrators to select two topics (Topic A and Topic B) for each of the 3 grade levels in the study. These topics had been discussed in class previously and were likely to be part of the upcoming standardized

tests, so the exercise doubled as a review. Randomizing at the student level allowed us to explore differences across classrooms due to differences in implementation fidelity.

Figure 1 shows a diagram of the experiment design. At the start of one class period, teachers administered a six question test estimated to take approximately 15 minutes that included (for each grade) three questions related to Topic A and three questions related to Topic B. Then students were randomly assigned to either Topic A or B and tasked to work on a Khan Academy assignment that covered the selected topic. The assignment consisted of three short videos and three corresponding exercises with four or seven questions each. Teachers were instructed to provide approximately one hour for students to work through the assignment and provide additional time at school or at home to continue working on it until completion (attaining at least three out of four or five out of seven questions correct for each exercise). About one week after the first test, a second designated class period was given for students to finish the assignment (if they had not already done so) and then take a second test with six questions covering both topics again.[3]

Of the 3,183 eligible students, 1,806 completed the six question pre-test (some eligible students were not actually in a classroom where the experiment took place while others ignored or chose not to participate). Altogether, 1,130 students completed both pre and post tests. Table 1 shows general balance among this sample by whether they were selected to work on exercises covering Topic A or B. Students are roughly evenly split across grades. Pre-test scores, gender, race, and special education status are, on average, about the same whether assigned to Topic A or B. Average practice time on KA was 19 minutes on the first day of the review, and 38 minutes in total over the week.

Table 2 shows the actual fraction of students that watched each video and completed each exercise. Each of these fractions was intended to be 100 percent. Students were supposed to take as much time as needed to review mistakes, rewatch videos, and redo exercises until completion. Teachers were asked to provide additional support for students that appeared stuck and unable to complete on their own. Clearly, not every student demonstrated mastery in their selected topic. Among all students completing the pre and post tests, 29 percent did not even watch the first

[3] Topics were selected by educational leads of the schools we were working with. Topics were selected from material covered in class earlier in the year, but likely not fully mastered by students. Exercises and test questions were selected by researchers using Khan Academy's existing library of exercises and videos.

video in its entirety. Only sixty-seven percent started the first KA exercise, and only about half completed it. Half watched the third exercise, and only 25 percent completed it.

Attrition in assignment progress differs not only within classrooms but across classrooms. When we group assignment progress by average classroom KA practice time we observe distinct differences in overall progress. Among the top 5 percent of classrooms in terms of average KA practice time (the top 2 classes), more than 90 percent of students completed the first exercise, 71 percent completed the second, and 61 percent completed the third. In contrast, among the bottom 5 percent of KA average practice time classrooms (the bottom 2 classes), only 41 percent completed the first exercise, 9 percent the second, and 3 percent the third.[4] Figure 2 further shows the wide contrast in practice time, videos watched, exercises attempted, and exercises completed across the 45 participating classrooms. These differences are also reflected in practice time differences: some classrooms exhibited average student practice times of more than an hour that week, while others averaged less than 5 minutes.

This variation remains after conditioning on baseline pretest scores. Appendix Figure A1 shows kernel density estimates of classroom practice time (adjusted to have mean zero), before and after conditioning on this pre-assignment characteristic. The wide dispersion in classroom practice time differences remains. We interpret this variation to suggest that teachers themselves appear greatly influential in whether students end up practicing a lot or a little. We examine below the consequences from differences in classroom practice time and degree of mastery by estimating treatment effects across these different categories.

B. Empirical Framework & Data

We observe the number of questions scored correctly for each student's post test, both for Topic A and Topic B. We standardize these two outcome variables by subtracting from the student's total the mean total of the comparison group (the group not assigned to practice on the topic) and then divide by the standard deviation of the comparison group. The score a student

[4] The average pre-score is higher among the top 2 classes compared to the lower 2, in terms of practice time (1.32 out of 3 questions correct compared to 0.83 questions correct respectively). But whether we drop the top 24 pre-score grades from the high-practice classes or the 28 low pre-score grades from the low-practice classes to ensure that the average pre-scores are the same, the large gap in average videos watched, exercises attempted and completed remains virtually unchanged (See Appendix Table A2).

received on the topic they were assigned to practice using KA is the treated score, and vice-versa for the control score. Thus, we estimate the following model:

$$(1) \qquad Y_{it} = \beta_0 + \beta_1 T_{it} + \beta_2 X_{it} + \alpha_i + \boldsymbol{\varepsilon}_{it}$$

Where Y_{it} is the standardized post test score for student i on topic t, T_{it} is the treatment status of topic t for student i, X_{it} is the standardized pretest score for student i on topic t, α_i is a student level fixed effect, and $\boldsymbol{\varepsilon}_{it}$ is the error term. The treatment effect, β_1 is the average difference in student's practice and unpracticed scores, averaged across the six topics and grades. Standard errors are clustered by student, since each student has two observations.

C. Results

Table 3 shows the estimated average gain in standardized test score from being given the KA mastery assignment to practice for a week compared to being given a different topic to work on. For the full sample of middle school students, the estimated effect is 0.22 standard deviations, not much lower compared to effect sizes found for more traditional tutoring programs.[5] Regressions 2-4 divide results by grade level. Since each grade practices different topics, we expect to see variation in impact. Grades 6 and 8 see similar effects, around 0.30 standard deviations. The effect on Grade 7 students is weakly significant and lower at 0.08 standard deviations, perhaps because the topics covered were more difficult overall. This is suggested in Table 2, where we see that Grade 7 also has the lowest attempt percentages on exercises and were the least likely to achieve familiar status, particularly for exercises 2 and 3. Thirty-three percent of Grade 6 students and 43 percent of Grade 8 students achieved familiarity (at least 70 percent of questions correct) with exercise 2, compared to only 16 percent of Grade 7 students. This led to fewer Grade 7 students even reaching the third video and exercise.

Regressions 5-8 look separately at classes based on the quantity of practice among students (students are categorized according to the average practice time among their classmates). In classes where students did not practice for more than five minutes, on average, we find no significant impact from the weekly review activity. The average impact on post test

[5] Estimated effects are similar across subgroups: 0.23*** (0.07) for males, 0.20*** (0.09) for females, 0.17* (0.09) for Black students, 0.22*** (0.07) for Hispanic students, 0.33* (0.15) for white students, and 0.15 (0.13) for special education students.

scores increases monotonically with average practice time in the class. Classrooms which averaged 25-50 minutes of practice saw improvements of 0.24 SDs and classrooms which averaged 50-100 minutes saw improvements of 0.27 SDs.

Regressions 9-12 similarly break out classes by the average number of "level-ups" per student. A level-up represents receiving at least 70 percent correct on an exercise initially, or redoing the exercise and receiving at least 70 percent compared to less than that earlier, or attaining 100 percent compared to less than that earlier. If every student in a class completed the weekly review as instructed, the average number of level-ups would be at least three, since every student was tasked to attain at least 70 percent correct on each of three exercises. Classes in which students averaged fewer than 0.8 level-ups saw insignificant improvements to scores. Classes that averaged between 0.8 and 1.2 level-ups saw improvements of 0.31 SDs, and classes above 1.2 level-ups saw an estimated average improvement of 0.35 SDs.

A concern with interpreting the variation in classroom practice time and mastery progress due to variation in teacher instruction is that they could be due to variation in student ability instead. For example, the F-test that student pre-score tests are balanced across classrooms takes on a value of 6.34 and therefore can be rejected. The coefficient from regressing class average practice time on class average pre-score is 5.22 (SE=1.29), so a classroom with an average one standard deviation higher pretest score would predict having an extra 5 minutes in average practice time. Yet as we saw above in section A, the distribution of classroom average practice time and level-ups remains practically unchanged after conditioning on pre-scores and other observable student characteristics. As an additional check and to examine classroom effects in more detail, Figures 3 and 4 plot average treatment effects for each classroom after conditioning on pre-score by average classroom practice time and level-ups respectively. The average of these treatment effects is 0.198, similar to the estimated effects without pre-score controls. Significant gains begin to appear for classrooms with at least 35 minutes of average practice time. Figure 3 also shows the estimated regression line for this relationship with and without conditioning on class average pretest score. The relationship remains virtually unchanged after accounting for differences in average prescore differences across classrooms.

A similar pattern of increasing treatment effect sizes arises when looking at average classroom level-ups. Significant effect sizes appear once classrooms have students leveling up an average of about 0.75 times over the week, with effects rising steadily with even more

average level-ups in the class. The outlier class with the highest average level-ups of 2.30 corresponds to having the largest effect size of 0.91. The linear trend in increasing treatment effects from increasing practice time remains virtually identical after conditioning on pre-test scores. We therefore interpret the positive slope in average classroom treatment effects as likely due to teacher differences in how they conducted the week review, with some able to facilitate greater quality practice than others.

To explore how much larger effect sizes might be if more students practiced and leveled up, regressions 13-15 in Table 3 show results conditional on the sample completing at least one, two, or three exercises. Among the sample of students who leveled up at least on the first exercise, post-test score performance increased an average of 0.38 SDs. For students that completed all three exercises, their overall post test score performance was more than half a standard deviation higher for the topic they worked on compared to the one they did not. While these samples condition on a select sample of more motivated or able students, they raise the possibility that average classroom learning gains could be substantially greater from helping more students persist towards mastery.

The main caveat with this experiment is that we are not comparing the benefits of using CAL with the benefits of providing regular classroom instruction (covering the same topic). Rather, we are comparing using CAL to help students' understanding of one particular math topic compared to using it to help with their understanding of another topic. The Nashville experiment demonstrates that CAL can be very effective for improving learning, but only when teachers facilitate sufficient practice time to master material. We turn next to how this kind of CAL activity compares to regular classroom instruction by randomizing whether teachers are given assistance in adopting CAL or not.

IV. Arlington Experiment

A. Setup

We conducted a year-long randomized controlled trial in the Arlington Independent School District (AISD) in Arlington, Texas. AISD is the 11th largest school district in the state

and covers a region just outside the Fort Worth area. Serving nearly 60,000 students in 77 schools, it primarily serves Hispanic (47.1%) and Black (25.8%) students.

Prior to the end of the 2020 school year, AISD administrators invited all math teachers in grades 3-8 to participate in our program. Teachers were told that they would be selected to receive personal assistance for implementing Khan Academy in their classrooms either in the 2021-22 or 2022-23 school year. Teachers were offered $300 for signing up, and another $300 for completing the onboarding process and starting to work with their assigned khoach. We had 312 teachers (about half of those eligible) initially registered to participate. Of these, 253 were actually rostered with regular classrooms during the 2021-22 school year. We removed a further 29 who left the district before the end of the school year, moved to teaching a non-math subject, or moved to a grade outside of our 3-8 sample range. Our final sample included 224 teachers, 112 in the treatment group and 112 in the control group. Teachers in the same grade and school were grouped for randomization to limit spillover effects between colleagues. Randomization was stratified by grade level and took place during the summer of 2021. All control teachers (those selected for assistance in the second year) were informed that they would be eligible to participate during the 2022-23 school year and we had no contact with the control teachers until the end of the 2021-22 school year.

Treatment teachers were given information for a professional development session that would take place during the district wide professional development days in August 2021. They were given the option of attending this session or watching a series of videos totalling about an hour, or doing both to help with training before meeting with their khoach prior to the beginning of school. The training involved motivating the program's mastery approach and describing the suggested recipe for facilitating at least an hour a week of regular CAL practice.

Table 4 describes average student characteristics among our 156 randomized teacher-grade groups. This represents 10,979 students among 224 teachers. These students generally match the district population in race, ethnicity, and gender. They are disproportionately concentrated in the elementary level (grades 3-6). In both treatment and control, nearly 50 percent of students are non-white, and 50 percent are ethnically Hispanic. Roughly 50 percent qualify for free lunch. The average number of days of school missed is about 20. We have a 4 percentage point difference in the percentage of female students and 3 percentage point difference in special education status between treatment and control group, which are statistically

significant at the 5 percent level. We condition on both of these variables in our primary specification. Note that student assignment to classes happened after randomization occurred, with no contact between researchers and those in the district determining class assignments.

Khoaches were hired and assigned to teachers during July 2022. 33 coaches supported our 112 treatment teachers, with coaches serving 2 to 6 teachers. The khoaches were primarily undergraduate and graduate students with a background in economics or education. They were trained over two sessions in both teacher interactions and Khan Academy use, and assigned based primarily on availability to meet during teachers' preferred times.

A high percentage of teachers attended the virtual professional development sessions (78.6 percent). More than half of the treated teachers watched a series of training videos and over 90 percent of teachers attended their first scheduled meeting with their khoach. There was some drop off in participation, with 76.8 percent participating in a second meeting with their khoach by the end of the first month of school (79.8 percent for elementary school teachers, 67.9 percent for middle school teachers).

Khoaches focused on attempting to meet with teachers each week virtually for about 30 minutes. As the year progressed they would often switch more towards communicating by email because meetings were becoming redundant or because of teachers' expressed preferences. Weekly emails were automatically sent to teachers to update them on student progress.

Like the Nashville experiment, average weekly practice time differed dramatically by classroom. Figure 5 shows the distribution of this variable, as well as the distribution of average weekly 'Level-Ups' across the 112 treated teachers' classes.[6] Some classes exhibited more than an hour of weekly practice time, on average, throughout the entire school year, which was the suggested target. Many other classes, however, exhibited zero or very little weekly average practice time. Notably, practice time was distinctly different by elementary and middle school status. Most elementary school classes (Grades 3-6) had average weekly practice times of 30 minutes or more (a mean of 34.7 minutes), whereas most middle school classes (Grades 7-8) had average weekly practice times of less than 10 minutes (a mean of 7.8 minutes). The contrast is likely due to three factors. First, elementary school classes included 70 minutes per day for

[6] Appendix Figure A2 shows these distributions for the classrooms of control classrooms, verifying that virtually no Khan Academy practice occurred among students with teachers in our control group. Appendix Figures A3 and A4 show the distributions of individual level practice times and level-ups among students with treated and control teachers respectively. An analysis of variance reveals that 24.6 percent of the individual variance in practice time among treated students is accounted for by differences across classrooms.

mathematics, while middle school classrooms included only 50 minutes per day. Second, Elementary school teachers taught the same students throughout the day and therefore had additional flexibility to adjust their class schedules or allow for additional Khan Academy practice time outside of the regular allocated time. Middle school mathematics teachers saw students only during their 50 minutes of allocated time. Third, elementary classrooms had additional enrichment time provided throughout the school week that teachers could use to assign reinforcing activities to students, including Khan Academy practice. In addition, AISD teachers often avoided assigning homework to students.[7] For these reasons, we present treatment effect estimates separately for elementary and secondary school students.

B. Empirical Framework & Data

Because treatment is randomly assigned, we use a simple regression for our main specification:

$$(2) \qquad Y_{igs} = \beta_0 + \beta_1 T_{igs} + \beta_2 X_{igs} + \gamma_g + \varepsilon_{igs}$$

Here, Y_{igs} is the standardized 2022 STAAR Math score for student i in grade g, school s. T_{igs} is the treatment status of student i in grade g in school s. X_{igs} is a matrix of student level characteristics including age, sex, race, ethnicity, days of school missed, english learner status, special ed status, and free lunch eligibility. γ_g is a grade level fixed effect and ε_{igs} is the error term, clustered at the grade-school level (the level of randomization).

C. Results

Table 5 shows our main results from the AISD experiment. Column 1 presents differences in the mean test score outcomes between the treatment and control groups. Column 2 shows estimated treatment effects from including a set of grade fixed effects, and Column 3 aligns with equation 2 above and includes grade fixed effects as well as added linear controls for

[7] Reasons given for not assigning homework were that parents preferred this setup, that more advantaged students received more help and support from their parents, and that some homes did not have adequate computer or internet access. Khoaches nevertheless advised encouraging at least some Khan Academy practice at home in order to avoid class time substitution and to provide additional opportunity for students to practice even more than what they could accomplish only at school.

some observable student demographics. All specifications group standard errors at the grade-school level (the level of randomization).

We see no statistically significant effect for our full sample of Grade 3 to 8 students. However, the finding masks distinct differences between elementary and middle schools, in line with distinct differences observed in Khan Academy practice time between the two groups. When we examine students from Grades 3 to 6 students, who practiced, on average, five times more than middle school students, we find a difference in end-of-year test score outcomes of 0.171 standard deviations. Adding grade fixed effects does not alter the estimate, while adding our demographic controls lowers the point estimate to 0.122, statistically significant at the five percent level. The point estimates for the effect of the program on middle school students are negative, though not statistically significant. One possibility is that middle school students were not provided enough time to continue working on exercises that they initially failed to complete. Without the opportunity to keep working on material until demonstrating mastery, practice time for middle school students may have been less productive compared to activities control group students were working on.

For elementary school classrooms, we also performed subgroup analyses to examine possible heterogeneous effects. Table 6 shows the estimates and statistics for these tests. Estimated effects for subgroups are less precise, and generally do not reveal any strong patterns for suggesting one group benefits more than another. For our primary specification (grade fixed effects and student characteristic controls), we fail to reject the null hypothesis that estimates are similar along any gender, race, ethnicity, learning ability, or economic divide. Rather, all elementary school groups appear to benefit from the program with estimated effects ranging from 0.082 to 0.215.

IV. Arlington Quasi-Experiment

To further investigate the importance of program fidelity, we explore the relationship between average classroom practice time and student performance. The analysis is similar to that from the teacher value-added literature in that it assumes students are "as-good-as-randomly" assigned to teachers, conditional on past performance. Chetty et al.

(2014a, 2014b), Kane et al. (2013), and Bacher-Hicks et al. (2019) have concluded that this approach can produce unbiased estimates of causal teacher fixed effects (overall teacher influences on student test scores). Rather than estimate these teacher fixed effects, however, here we estimate whether treated teachers who facilitated more Khan Academy practice time also ended up facilitating higher end-of-year test scores for students that entered each class with similar previous performance. We conduct this analysis across schools and within schools to see if a relationship exists even for students in the same school and grade but with different teachers.

Of course, a teacher's value added might also be correlated with their ability to generate higher Khan Academy practice time under the KWiK program, leading to an upward bias in the interpretation of the correlation between test scores and class practice time as causal. For this reason, we should be cautious, but we believe the analysis is nevertheless useful because a positive relationship is likely a necessary condition for a causal relationship to exist, and a positive relationship would reinforce the conclusions from the two previous experiments that a sufficient amount of practice time is needed in order to generate gains from the program's mastery approach to learning.

For a main illustrative example, Figure 6 plots each Grade 4 to 6 treated classroom's average weekly practice time on Khan Academy (average of other students in the class) and the corresponding average students' standardized test score after conditioning for previous year's STAAR score (we cannot look at Grade 3 because we do not have their previous year's test score).[8] Classrooms with less than 35 minutes of weekly practice do not exhibit any clear pattern, but average test scores become positively related by the time we include classrooms with at least 50 minutes of practice per week. The slope from the OLS regression for all 68 classrooms is 0.0035 (standard error 0.0011), implying about a 0.25 standard deviation increase in test score performance from being in the classroom with average practice time of 50 weekly minutes compared to none. Appendix Figures A5 and A6 show similar patterns using the full Grades 4 to 8 treated sample, and for the treated and control samples combined.

The more direct relationship between classroom progress (in terms of the weekly level ups averaged among other students in the class) and the corresponding average students' standardized test score is shown in Figure 7 for students with treated teachers in Grades 4 to 6.

[8] We first regress the standardized 2022 STAAR score on a student's 2021 score for the sample of students with treated teachers in Grades 4 to 6. We then plot the mean of these residuals by classroom and the corresponding classroom's average weekly Khan Academy practice time.

The relationship is similar to practice time. Classrooms with average weekly level-up activity of 2 or more exhibit significantly higher test-score gains. The slope of this relationship is almost 0.1 standard deviations, so being in a class with an average of 4 level-ups a week is associated with scoring 0.4 standard deviations higher than a class with no level-ups. Being in a class with an average of 5 level-ups a week is associated with scoring 0.5 standard deviations higher, and so on. Appendix Figures A7 and A8 show the same type of relationship for the Grades 4 to 8 sample and for the combined treated and control samples.

To estimate a possible critical level of weekly classroom practice time or number of level ups, we use the following model:

(3) $$Y_{igs} = B_0 + B_1 P_{i,}^{Z} + B_2 Y_{ig\text{-}1s} + \beta_3 X_{igs} + \varepsilon_{igs}$$

Here Y_{igs} is the standardized Math STAAR score in 2022 for student i in grade g and school s. P_i^Z is an indicator for whether the classroom practice among i's classmates was above Z minutes per week. We leave out a student's individual practice time in calculating this average. In an alternate specification, P_i^Z indicates whether a student's classmates averaged above Z "level-ups" per week. $Y_{ig\text{-}1s}$ is the student's standardized Math STAAR score in 2021, X_{igs} is a matrix of the student's demographic data, and ε_{igs} is the error term. B_1 is our coefficient of interest, and represents the effect of a student being assigned to a teacher who facilitates high practice. We estimate B_1 for the overall full sample, and for only the treated sample of teachers (to focus on the within treatment group variation in practice time). We also examine results after adding school and grade fixed effects, which leads to comparing classroom differences within the same school and grade level.

Table 7 reports the association of being in classrooms with higher practice times on STAAR scores. Average weekly practice among a student's classmates is calculated leaving out an individual's practice. 2022 STAAR scores are regressed on dummy variables for cutoffs in this classmate practice variable, controlling for the individual's prior test scores. Among the general sample and in the particular grades of interest, 4 to 6, we see an increasingly strong relationship with practice time, up to roughly 0.21 SDs at 35 minutes of practice. Above this, the relationship appears stable. Lack of significant practice in middle school prevents us from obtaining estimates among higher cutoffs for grades 7 and 8, as no student had classmates

average above 30 minutes throughout the year. We do see weakly significant negative associations to practice time of 10 minutes or more for students in Grades 7 and 8. Since most middle school KWiK teachers have practice times concentrated at low levels, this suggests the possibility that, if practice is not performed in quantities high enough to allow students to work through mistakes, try again, and cover a sufficiently large number of topics, this small time taken away from alternative classroom activities could have negative repercussions.

Table 8 shows similar results, but after including school-grade fixed effects to compare students from the same school and grade but with different teachers with different class practice times. The point estimates of these associations are less precise since there are fewer teachers in the sample from the same school and grade and less variation in performance between them. Nevertheless, we estimate statistically large associations between test scores and class practice, even for lower average practice times as low as 15 minutes per week or more. A student with two KWiK teachers in her grade and school does substantially better on her STAAR test score after spending her year with the teacher that facilitated more practice time and level-ups in the class.

Table 9 uses a similar specification to Table 7, but uses average classmate level-ups per week as a measure of class practice, rather than practice time.[9] Here, we see a similar increase in 2022 test scores associated with assignment to classes with higher average level-ups. In the full sample and grades 3-6, three or more level-ups per week is comparable to 35 minutes of practice, with estimates of 0.242 and 0.206 SDs, respectively. Interestingly, point estimates do not stagnate with higher level-up cutoffs in the same way they do with practice time. Achieving five or more level-ups per week is associated with 0.35 standard deviations higher test scores for the overall sample and 0.306 standard deviations for the grades 4-6 sample. This relationship suggests that an emphasis on mastering activities may improve outcomes more than an emphasis on a specific quantity of practice time. We also see higher scores for students in classes with higher level ups for middle school students. This suggests that the program could be effective in higher grades if students were given more opportunity to practice and progress on problems.

Finally, it is worth noting the within class relationships between a student's practice time, level-ups, and their subsequent test score performance. Table 10 shows coefficient estimates

[9] Appendix Table A2 shows similar but more imprecise results after also including gradexschool fixed effects (so as to compare students with similar previous performance in the same school and in the same grade).

from regressing a Grade 4-6 student's 2002 standardized Math STAAR score on their own average weekly Khan Academy practice time and average weekly level-up progress, and including teacher fixed effects and the same student level controls as before. Within treated classrooms, there exists a strong positive relationship between more practice, more level-ups, and higher test scores. When both practice time and level-ups are included in the same regression, Column 3 of Table 9 makes clear that it is leveling up activity that matters and not practice time. The coefficients for leveling up 1-2, 2-5, and 5+ times are almost identical with and without adding practice time in the regression, but the coefficients for practice time fall from highly significant to zero, or possibly negative for practice time of 50 minutes or more, with level-ups held constant. Thus, the pattern is consistent with students of similar background improving substantially from level-up progress in Khan Academy, regardless of how long it takes to level-up.

V. Teacher Adoption Differences

Our findings above support the conclusion that Computer Assisted Learning can be integrated into the classroom to produce significant mathematics learning gains over a school year at lower cost and with fewer implementation challenges compared to many tutoring programs. Facilitating student practice time and progress appears key. Therefore, of critical importance for scaling purposes is understanding what kinds of teachers achieved especially high quality practice and whether implementation improvements could be made to help more teachers attain high quality practice. To help address these questions, we explore below what observable teacher characteristics correlate most with CAL classroom practice time, and conduct qualitative interviews with those teachers who facilitated the most and least practice times in our study.

Prior to participation, teachers in our Texas experiment took two surveys. One, given in May 2021 during enrollment, was given to all teachers and collected data on the way classrooms were set up prior to the study. The second, administered during summer 2021, was given only to treatment teachers and asked about strategies the teacher anticipated employing during the 2021-22 school year. For the variables we observe, Table 11 shows estimated correlations between these teacher survey responses and the likelihood that a teacher's students average 35 minutes or more of practice throughout the school year (hereafter "high-dosage" classrooms).

There were 5% of teachers who reported teaching multiple grades in 2020-21, and these teachers were 48 percentage points less likely to have a high-dosage classroom. There were 25% of teachers who reported that homework was rarely assigned in their classrooms, and these teachers were 28 percentage points less likely to have a high-dosage classroom (.05 significance). Interestingly, teachers who reported having never used Khan Academy before (19% of teachers) and teachers who reported that less than three-quarters of their students had access to technology at home (23% of teachers) were not less likely to have high-practice. Conversely, teachers who reported expecting high amounts of in school and at home math practice (90+ minutes per week), were not more likely to have higher practice. A very small percentage of teachers reported that their students already worked on Khan Academy one or more times per week (6%). These teachers were less likely to have high practice, by over 50 percentage points. Teachers who engaged more with coaches and training at the start of the year were more likely to be successful. There were 77% of teachers who attended or watched professional development relating to the program, and these teachers were 31 percentage points more likely to have high practice. Teachers who held their first meeting with a coach (92% of teachers) were also more likely to have high practice, as were teachers who had a second meeting in the first 6 weeks of school (49 and 36 percentage points more likely, respectively).

For qualitative analysis, we conducted a series of individual virtual interviews lasting approximately 30 minutes each. We reached out to nineteen teachers with the highest average weekly student practice time (over 70 minutes for elementary school teachers, over 30 minutes for high school teachers), twelve of which responded. We also reached out to a random sample of 10 teachers with average weekly practice time between 5 and 15 minutes, 5 which responded. Teachers were asked to candidly give their thoughts on the program, as well as react to specific questions about their implementation of KWiK, barriers to their success, and overall experiences.

High-practice teachers mentioned a strong sense of buy-in at the beginning of the program. They internalized the practice goal of at least 60 minutes a week as part of their own teaching goals, rather than as a 'nice to have' option or bonus. Low-practice teachers admitted to thinking of the program more as an option to try if time allowed and convenient. One remarked, "Even for me, it became just another task. And so I just remember being like, I'm being honest with you, like it's hard for me right now to get them to do it because I don't really want to do it.".

High-practice teachers, which in many cases still faced the challenges cited by low practice teachers, were much more likely to see them as temporary setbacks, more likely to suggest their own solutions to barriers, more proactive in keeping communication with their khoach, and generally seemed more committed to the program. They also reported high satisfaction with the program and many reported anecdotal evidence of learning gains, even prior to researcher access to end of year test scores. When one of the high-practice teachers was asked why she thinks her class was exceptional in attaining a high degree of practice, she responded, "Maybe I just have an expectation that it gets done.". Others remarked:

"I had a plan to make sure that I would be able to integrate the goals in class. I made sure to convince myself that this was not going to be just another program"

"For me, I didn't see it as just a program that the kids got on and then just spent some time on. For me, I saw it as an amazing tutorial to help kids get caught up if they were behind. But not only that, work at their own pace…I don't see it as me sitting at my desk wasting time while they're doing something. I am on my Khan looking at their results, checking what it is. I'm communicating with them because I want them to know how important this program is for them…Some people see these programs as just another thing that the kids are supposed to do, but when I first went through the program and went through the classes and everything, I thought this is what I've been waiting on."

Another prevalent theme of high-practice teachers was a deliberate plan for when to practice every week. Whether it was 45 minutes for two days a week during regular math instruction (out of 90 minutes a day) or 20 minutes a day, effective teachers who facilitated high-dosage practice stuck with a weekly routine. Many allocated time at the beginning of each school each day, when the first twenty minutes was allocated for arrival and announcements:

"What I did at the very beginning of each class is they had to devote 20 minutes on Khan. The idea was to get a level up every single day."

"As soon as they come in, I've already put the computers on the desk, so all they have to do is open [them] up…They're just going straight to Khan Academy and then to [their] assignments".

Others used independent "What-I-Need" time for scheduling practice, a flexible period of school time for teachers to decide how to structure. This was the case for the few high-dosage practice middle school math teachers who had more limited time with students. A challenge for this period was that students might be in another classroom with a different teacher that was not as clear how to supervise practice.

High-practice elementary school teachers often used candy, stickers, free time, leaderboards and sometimes grades as incentives. One teacher offered a single "smartie or dum-dum" for each level-up. Another said if students get "five level ups in one week they get a prize like a candy or a bag of chips". Another uses "these little boxes of candy…and if they get 90 minutes, these kids will work for candy. I know it sounds crazy and I do this out of my own pocket because kids will work for incentive…[And] if you have, like the most level ups in the most minutes, you get a little trophy". Another offered "Chick-fil-A gift cards and big candy bars for the top three in minutes". Another offered stickers: "I just bought all these big stickers like you can put on your water bottles or your binder. They really look forward to that. So I give those up for 60 minutes…by Friday". Some, but not all, teachers used grades for KA practice and progress, especially middle school students. Grade schemes varied:

"They got what I call the buffer 100. They get 100 for getting everything completed…Like for a week, I try to say that I want something around…10-12 level ups a week, which I think is reasonable considering they have a lot of time to do it. They can finish it in any class if they want to . They have a 45 minute tutorial time as well as time in my class."

"If they get their 60 minutes in, then they would get 100. If they got 30 minutes in, it was 50.

"[Students are expected to go look at their grades…They all know that it is a percentage of 60. So if you did 55 minutes you're gonna get 55 out of 60. And there's another column that says level up for the week."

Teachers that facilitated high-dosage practice also closely monitored student activity. They became comfortable using KA tools for assessing progress and regularly intervened when observing inadequate progress:

"I would pull them into my room…and they would sit at the horseshoe table with me and they would do their level ups. And what I found out most of them who weren't progressing is that they weren't watching any of the videos. They were just starting to try the [exercises] and they didn't understand what they were doing. So when I sat there with them and I said this is what you're going to do, this is how you're going to do it, usually they got it really quick because they would watch the video and they were like oh that's what it is."

"I usually click into each student and I look at where they were practicing that day and their success or I also look at how many times or how many minutes they were on it. And I usually pull them aside at least once a week and then I talk to them about their progress."

"If I see that they haven't leveled up in a while…I try to go and kind of check in with them and see what I'm dealing with or tell them to have somebody near them help."

Most teachers did not give KA practice as homework, often because they were concerned not every student would have computer access:

"It wasn't fair for me to say, OK, this is what you have to do at home because there was some that just couldn't do it. So I just changed it where I said if you have a computer, if you're able to get in, then this is your option. If you want to receive a participation, they just come to school early before class and then I'll let you get in and do some work that you can accumulate. But I couldn't really assign anything because I had like out of the 17 students I had last year, probably 5 didn't have either. They didn't have technology, they didn't have internet service at home. So it was hard to make them do things at home when it was, they just couldn't do it."

"The problem with doing it at home was we don't have school issued computers that the students take home. We have one to one at school, but they're not allowed to take them home. So I had to

make sure that the kids had time to do it at school to get their minutes in. Couldn't require them to do them at home."

"I sent out frequent reminders saying Khan Academy…[is] available for them to use at home. You know, encourage your students to practice. I don't really know how much they do. I have one or two that do at home because their parents really push that. But I wouldn't say the majority of them do."

This lack of KA practice at home is unfortunate because the platform lends itself well as a homework tool. Students have more time to work on exercises compared to being restricted to working on Khan Academy during school. Unlike paper and pencil assignments, students can receive immediate feedback working through problems, review mistakes, and try again. Teachers can monitor progress and identify common errors to review in class. If computer access concerns could be alleviated, perhaps teachers would be more willing to adopt a homework-focused KWiK model, in which students practice exercises related to their teacher's current curriculum.

Low-practice teachers gave several potential explanations for low practice among their students. In grades 7 and 8, short classes were the most cited reason. In lower grades, explanations for low practice included: lack of class time, lack of support (or opposition to the program) from school administrators, competing school interventions, lack of access to technology in school, and serious learning gaps in classrooms that required more of a teacher's time and energy.

This contrast between high and low-practice teachers raises the question of whether a teacher's ability to facilitate high-dosage practice in their classroom depends on their innate characteristics, or whether practice time can be significantly influenced by program adjustments. We estimate average program effects and find evidence that teachers who facilitated more practice also saw students with the most gains, but this research cannot objectively establish whether low practicing teachers can be encouraged to become high-practice teachers. Our qualitative findings at least suggest potential areas for improvement for future iterations or similar programs. Firstly, designated program time during the school day, or improved scaffolding for at home practice. This could take the form of an elective class period in higher grades or curriculum designated days for this type of practice in lower grades. For out-of-school

practice, grading or other student incentives seem important, as well as creating a higher degree of homework completion expectations and addressing tech access and resources for students with less home support. Relatedly, it seems important that the program has sufficient buy-in, not only from the teacher, but from administrators, coaches, department leads, and those others with the school who have the power to remove barriers to teacher participation.

VI. CONCLUSIONS

A mastery approach to learning - allowing students to progress incrementally at their own pace, receive immediate feedback, review mistakes, and only move on to the next topic after demonstrating understanding of the previous one - offers one of the most convincing approaches for educating children (Schunk, 2012). The key issue is not whether adopting mastery is worthwhile, but how to do it at scale. Supplementing classroom instruction with tutoring moves towards this goal, but tutoring is often too expensive or operationally difficult to implement for large numbers of students. Computer Assisted Learning has the potential to lower costs by leveraging existing education resources and rely less on person-to-person interactions. However, school administrators and teachers must master themselves how to effectively incorporate CAL into classroom instruction.

In this paper, we investigate benefits from offering teachers more proactive and continuous support for adopting CAL in their classrooms. The Khoaching With Khan Academy program involved each teacher being assigned a 'khoach' who met with at the start the school year and weekly thereafter until the teacher felt comfortable to proceed using CAL on their own. The approach was based on the theory that teachers require a high degree of scaffolding and encouragement when adopting a new approach to their curriculum delivery.

Even with KWiK's proactive assistance, teachers varied markedly by the extent to which they facilitated CAL practice and progress in their classrooms. In Nashville, teachers were trained to use Khan Academy to help students practice on one of two topics as a mathematics review. Test scores a week later were, on average, 20 percent of a standard deviation higher for the topic students were instructed to review with CAL compared to the topic they did not work on. Upon further inspection, average gains were only detected for classrooms with student

practice times averaging more than 35 minutes over the week. In Arlington, Grades 3-8 math teacher volunteers were randomly assigned a khoach to assist throughout the school year. The school district did not generally promote homework, so most CAL practice occurred during school. Elementary school teachers had more flexibility to schedule practice time, which occurred often at the start of school or during independent study time. State test scores were 0.12-0.17 standard deviations higher for students with treated teachers than control. A clear relationship occurs between treated teacher's average class practice times and test score gains, even among students with similar initial math abilities and from the same schools, suggesting again that the more students are able to practice with CAL, the greater gains. This conclusion is reinforced from finding no significant treatment effects for Grades 7-8 students, in which their teachers were far less able to generate weekly CAL practice times more than 10 minutes per week.

Given evidence that positive program effects occur only after a classroom attains a threshold of average practice time, it may be worth imputing treatment-on-the-treated (TOT) effects under alternative assumptions. Table 12 compares our ITT with TOT estimates. Assuming no effects when a treated teacher did not meet more than twice, the average improvement to STARR math scores for students with elementary teachers who met this threshold, rises slightly from our ITT effect of 0.122 (with controls included) to 0.150. Assuming no effects for teachers with less than 20 minutes of average practice time, the TOT effect increases to 0.236, and using the threshold of 35 minutes of average weekly practice time, the TOT becomes 0.306 with controls, and 0.428 without controls. The TOT estimates are about the same or slightly higher when including elementary and middle school teachers, since few middle school teachers met these practice thresholds. The estimates are also about the same when instrumenting actual student practice time with random assignment rather than average student practice time within a classroom.

Overall, teachers whose students practiced the most were very committed to the program from the start and worked closely with khoaches. They carefully planned when weekly practice would occur, but did not often rely on homework. They monitored progress closely. They incorporated practice as part of their math curriculum. When students were not progressing, they intervened. They often took extra time customizing practice goals for different students. Relative

accomplishments were praised and rewarded, sometimes with stickers, candy, or free time. Teachers developed a classroom culture of students that tolerated or even enjoyed CAL practice.

The development of artificially intelligent tutors using large language models offers a further potential for scaling personalized learning using CAL. Khan Academy's 'Khanmigo', for example, uses Chat-GPT as its engine for offering students a virtual tutor, available any time and without stigma. Virtual tutors have the potential to further personalize instruction while using CAL. The same technology can be used to assist teachers for designing assignments more closely aligned with their curricula and for efficient monitoring of student progress. Further research is needed for determining how technology can best be used for helping scale mastery learning but, so far, the potential for substantially improving learning outcomes – at least in math – seems promising.

References

Bacher-Hicks, Andrew, Mark J. Chin, Thomas J. Kane, and Douglas O. Staiger. "An Experimental Evaluation of Three Teacher Quality Measures: Value-Added, Classroom Observations, and Student Surveys." *Economics of Education Review* 73 (2019). https://doi.org/10.1016/j.econedurev.2019.101919

Barrow, Lisa, Lisa Markman, and Cecilia Elena Rouse. "Technology's Edge: The Educational Benefits of Computer-Aided Instruction." *American Economic Journal: Economic Policy* 1, no. 1 (2009): 52-74.

Beck, Robert J. "Towards a Pedagogy of the Oxford Tutorial." Conference on the Oxford tutorial, Lawrence University, 2007. Retrieved from https://www2.lawrence.edu/fast/beckr/pdfs/OxfordTutorial_7_05_06.pdf

Beg, Sabrin, Waqas Halim, Adrienne M. Lucas, and Umar Saif. "Engaging Teachers with Technology Increased Achievement, Bypassing Teachers Did Not." *American Economic Journal: Economic Policy* 14, no. 2 (2022): 61-90. https://doi.org/10.1257/pol.20200713

Büchel, Konstantin, Martina Jakob, Christoph Kühnhanss, Daniel Steffen, and Aymo Brunetti. "The Relative Effectiveness of Teachers and Learning Software: Evidence from a Field Experiment in El Salvador." *Journal of Labor Economics* 40, no. 3 (2022): 543-777. https://doi.org/10.1086/717727

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review* 104, no. 9 (2014a): 2593-2632.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review* 104, no. 9 (2014b): 2633-79.

Copeland, Susan, Michael A. Cook, Ashley A. Grant, and Steven M. Ross. "Randomized-Control Efficacy Study of IXL Math in Holland Public Schools." Johns Hopkins Center for Research and Reform in Education, Baltimore, MD, 2023. Retrieved from: https://jscholarship.library.jhu.edu/handle/1774.2/69038

Escueta, Maya, Andre J. Nickow, Philip Oreopoulos, and Vincent Quan. "Upgrading Education with Technology: Insights from Experimental Research." *Journal of Economic Literature* 58, no. 4 (2020): 897-996.

Fancsali, Stephen E., Steven Ritter, Michael Yudelson, Michael Sandbothe, and Susan R. Berman. “Implementation Factors and Outcomes for Intelligent Tutoring Systems: A Case Study of Time and Efficiency with Cognitive Tutor Algebra.” *Proceedings of the Twenty-Ninth International Florida Artificial Intelligence Research Society Conference* (2016): 473-478.

Feng, Mingyu, Chunwei Huang, and Kelly Collins. “Technology-based Support Shows Promising Long-term Impact on Math Learning: Initial Results from a Randomized Controlled Trial in Middle Schools.” WestEd, San Francisco, CA, 2023. Retrieved from https://files.eric.ed.gov/fulltext/ED630781.pdf

Ferman, Bruno, Lucas Finamor, and Lycia Lima. “Are Public Schools in Developing Countries Ready to Integrate EdTech into Regular Instruction?” Munich Personal RePEc Archive paper no. 109063, Munich, Germany, 2021. Retrieved from https://mpra.ub.uni-muenchen.de/109063/

Kane, Thomas J., Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger. “Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment.” Bill & Melinda Gates Foundation technical report, Seattle, WA, 2013. Retrieved from https://files.eric.ed.gov/fulltext/ED540959.pdf

Kehrer, Paul, Kim Kelly, and Neil Heffernan. “Does Immediate Feedback While Doing Homework Improve Learning?” *Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference* (2013): 542-545.

Major, Louis, Gill A. Francis, and Maria Tsapali. “The Effectiveness of Technology-Supported Personalised Learning in Low- and Middle-Income Countries: A Meta-Analysis.” *British Journal of Educational Technology* 52, no. 5 (2021): 1935-1964. https://doi.org/10.1111/bjet.13116

Mendicino, Michael, Leena Razzaq, and Neil T. Heffernan. “A Comparison of Traditional Homework to Computer-Supported Homework.” *Journal of Research on Technology in Education* 41, no. 3 (2009): 331-359.

Morgan, Pat, and Steven Ritter. “An Experimental Study of the Effects of Cognitive Tutor Algebra I on Student Knowledge and Attitude.” Carnegie Learning, Inc., Pittsburgh, PA, 2002.

Murphy, Robert, Larry Gallagher, Andrew Krumm, Jessica Mislevy, and Amy Hafter. “Research

on the Use of Khan Academy in Schools." SRI Education, Menlo Park, CA, 2014. Retrieved from: https://www.sri.com/wp-content/uploads/2021/12/khan-academy-implementation-report-2014-04-15.pdf

Nickow, Andre, Philip Oreopoulos, and Vincent Quan. "The Promise of Tutoring for PreK–12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence." *American Educational Research Journal* 61, no. 1 (2024): 74-107. https://doi.org/10.3102/00028312231208687

Pane, John F., Beth Ann Griffin, Daniel F. McCaffrey, and Rita Karam. "Effectiveness of Cognitive Tutor Algebra I at Scale." *Educational Evaluation and Policy Analysis* 36, no. 2 (2014): 127-144. https://doi.org/10.3102/0162373713507480

Pane, John F., Daniel F. McCaffrey, Mary Ellen Slaughter, Jennifer L. Steele, and Gina S. Ikemoto. "An Experiment to Evaluate the Efficacy of Cognitive Tutor Geometry." *Journal of Research on Educational Effectiveness* 3 (2010): 254-281. https://doi.org/10.1080/19345741003681189

Phillips, Andrea, John F. Pane, Rebecca Reumann-Moore, and Oluwatosin Shenbanjo. "Implementing an Adaptive Intelligent Tutoring System as an Instructional Supplement." *Educational Technology Research and Development* 68 (2020): 1409-1437.

Ran, Hua, Nam Ju Kim, and Walter G. Secada. "A Meta-Analysis on the Effects of Technology's Functions and Roles on Students' Mathematics Achievement in K-12 Classrooms." *Journal of Computer Assisted Learning* 38, no. 1 (2022): 258-284. https://doi.org/10.1111/jcal.12611

Ray, Brian D. "Home Schools: A Synthesis of Research on Characteristics and Learner Outcomes." *Education and Urban Society* 21, no. 1 (1988): 16-31. https://doi.org/10.1177/0013124588021001003

Ritter, Steven, John R. Anderson, Kenneth R. Koedinger, and Albert Cortbett. "Cognitive Tutor: Applied Research in Mathematics Education." *Psychonomic Bulletin & Review* 14 (2007): 249-255. https://doi.org/10.3758/BF03194060

Roschelle, Jeremy, Mingyu Feng, Robert F. Murphy, and Craig A. Mason. "Online Mathematics Homework Increases Student Achievement." *AERA Open* 2, no. 4 (2016). https://doi.org/10.1177/2332858416673968

Roschelle, Jeremy, Nicole Shechtman, Deborah Tatar, Stephen Hegedus, Bill Hopkins, Susan Empson, Jennifer Knudsen, and Lawrence P. Gallagher. “Integration of Technology, Curriculum, and Professional Development for Advancing Middle School Mathematics: Three Large-Scale Studies.” *American Educational Research Journal* 47, no. 4 (2010): 833-878. https://doi.org/10.3102/0002831210367426

Schunk, Dale H. *Learning Theories: An Educational Perspective*. Boston, MA: Pearson, 2012.

Snipes, Jason, Chun-Wei Huang, Karina Jaquet, and Neal Finkelstein. “The Effects of the Elevate Math Summer Program on Math Achievement and Algebra Readiness.” U.S. Department of Education, Institute of Education Sciences, Regional Educational Laboratory West technical report no. REL2015–096, Washington, DC, 2015.

Wang, Haiwen, and Katrina Woodworth. “Evaluation of Rocketship Education’s Use of DreamBox Learning’s Online Mathematics Program.” SRI International, Center for Education Policy, Menlo Park, CA, 2011. Retrieved from: https://info.discoveryeducation.com/rs/063-SDC-839/images/ef-2011-08-SRI_Rocketship_Evaluation.pdf

Weatherholtz, Kodi, Phillip Grimaldi, and Kelli Millwood Hill. “Use of MAP Accelerator Associated with Better-than-Projected Gains in MAP Growth Scores.” Khan Academy technical report, Mountain View, CA, 2022. Retrieved from http://khan.co/MATechReport2022

White, Sara, Leiah Groom-Thomas, and Susanna Loeb. “A Systematic Review of Research on Tutoring Implementation: Considerations when Undertaking Complex Instructional Supports for Students.” Brown University’s Annenberg Institute EdWorkingPaper no. 22-652, Providence, RI, 2023. https://doi.org/10.26300/wztf-wj14

Figures and Tables

Table 1

Nashville Experiment, Balance and Student Characteristics

	N	Mean		Diff.
		Topic A	Topic B	
Grade 6	1130	0.33	0.34	0.011
Grade 7	1130	0.38	0.34	-0.005
Grade 8	1130	0.29	0.29	-0.006
Pre-test Score *(standardized)*	1130	-0.01	-0.00	0.010
Students per Teacher	1130	117.60	116.46	-1.134
Class Practice: Day 1 *(minutes, avg.)*	1126	18.99	18.71	-0.278
Class Practice: Total *(minutes, avg.)*	1126	38.40	38.12	-0.280
Class Minutes: after Day 1 *(minutes, avg.)*	1126	19.42	19.41	-0.002
Female	1118	0.46	0.42	-0.043**
Race: Black	1118	0.27	0.28	0.015
Race: White	1118	0.11	0.11	0.003
Ethnicity: Hispanic	1118	0.56	0.55	-0.013
Special Ed	1118	0.10	0.10	-0.005

Notes: **Balance table showing variable means and differences between students randomized into practice topic A or topic B. Randomization is based on which topic each student is assigned to work through with Khan Academy.**

Table 2

Nashville Experiment, Participant Progression through Practice Assignment

		Exercise 1			Exercise 2			Exercise 3		
	N	Watched	Attempted	Familiar	Watched	Attempted	Familiar	Watched	Attempted	Familiar
	(number)	*(fraction of total)*			*(fraction of total)*			*(fraction of total)*		
Grade 6										
Topic A	191	0.75	0.72	0.66	0.68	0.65	0.40	0.54	0.40	0.22
Topic B	184	0.59	0.56	0.41	0.49	0.43	0.27	0.69	0.49	0.22
Grade 7										
Topic A	222	0.66	0.67	0.45	0.52	0.37	0.14	0.36	0.30	0.22
Topic B	204	0.72	0.68	0.50	0.54	0.50	0.19	0.41	0.33	0.15
Grade 8										
Topic A	172	0.79	0.64	0.39	0.58	0.53	0.32	0.51	0.53	0.38
Topic B	157	0.80	0.75	0.54	0.64	0.69	0.55	0.62	0.48	0.32
Total	**1130**	**0.71**	**0.67**	**0.49**	**0.57**	**0.52**	**0.30**	**0.51**	**0.41**	**0.25**
By average progression among classmates										
Bottom 5%	75	0.53	0.41	0.20	0.28	0.25	0.09	0.19	0.15	0.03
Top 5%	79	0.95	0.92	0.78	0.81	0.78	0.71	0.75	0.68	0.61

Notes: **This table shows the percentage of participants who watched each video, attempted each exercise, and achieved "familiar" status (70% correct or higher) on each exercise; categorized by grade and topic. For each student, we also calculate the average number of exercises that their classmates became familiar with (not including a student's own practice).**

Table 3

Nashville Experiment, Main Results; also by Grade, Practice

	(1)	(2)	(3)	(4)
			By Grade	
	Full Sample	Grade 6	Grade 7	Grade 8
Treatment	0.22***	0.29***	0.08	0.30***
	(0.05)	(0.08)	(0.08)	(0.10)
N	*2260*	*750*	*852*	*658*
	(5)	**(6)**	**(7)**	**(8)**
		By Average Class Practice Time		
	<5 mins.	5-35 mins.	35-50 mins.	50-100 mins.
Treatment	0.08	0.16**	0.27***	0.27***
	(0.33)	(0.07)	(0.10)	(0.10)
N	*50*	*1000*	*636*	*566*
	(9)	**(10)**	**(11)**	**(12)**
		By Average Class Level-Ups		
	<0.4 skills	0.4 - 0.8 skills	0.8 - 1.2 skills	>1.2 skills
Treatment	0.06	0.09	0.31***	0.35***
	(0.11)	(0.09)	(0.10)	(0.09)
N	*386*	*622*	*586*	*666*
	(13)	**(14)**	**(15)**	
		By # of Activities Mastered		
	Exercise 1	Exercise 1-2	Exercise 1-3	
Treatment	0.38***	0.43***	0.52***	
	(0.07)	(0.10)	(0.12)	
N	*1110*	*640*	*436*	

Notes: **Treatment effects of practice on standardized post test scores are estimated using the model outlined in Equation (1). Each student provides 2 observations, their standardized performance on the topic they were asked to work on, and the one they did not work on. Regressions 2-4 divide effects by grade level. Regressions 5-8 divide effects by the average practice time in a given class. Regressions 9-12 are by average class level ups. Regressions 13-15 show results conditional on the sample completing at least one, two, or three exercises.**

Table 4

Arlington Experiment, Student Characteristics by Treatment Status

	Control	Treatment	Diff.
Age	10.05	10.09	0.034
Race: White	0.51	0.53	0.022
Race: Black	0.29	0.26	-0.030
Race: Native American	0.10	0.10	-0.005
Ethnicity: Hispanic	0.52	0.47	-0.051
Female	0.50	0.46	-0.35***
ESL	0.23	0.23	0.004
Special Ed	0.13	0.16	0.034**
Free Lunch Eligible	0.50	0.47	-0.023
Days Missed	20.53	20.19	-0.336
Grade 3	0.25	0.24	-0.013
Grade 4	0.17	0.21	0.036
Grade 5	0.20	0.17	-0.029
Grade 6	0.23	0.24	0.012
Grade 7	0.07	0.08	0.004
Grade 8	0.07	0.07	-0.009
N = 156			

Notes: **Averages for student characteristics across each grade-school randomization unit. Represents 10,979 students among 224 teachers. Teachers were randomized prior to the end of the 2020-2021 school year and treatment status was not communicated directly to school admins prior to the start of the 2021-2022 school year.**

Table 5

Arlington Experiment, Main Specification by Grade Level

	I No Control	II Grade FEs	III Grade FEs w/ Controls
Full Sample	0.036	0.044	0.025
	(0.092)	(0.083)	(0.076)
N	*10,979*	*10,979*	*10,979*
Grades 3-6	0.171**	0.172**	0.122**
	(0.069)	(0.070)	(0.058)
N	*7,234*	*7,234*	*7,234*
Grades 7-8	-0.201	-0.202	-0.173
	(0.206)	(0.194)	(0.202)
N	*3,745*	*3,745*	*3,745*

Notes: **OLS regressions of standardized 2022 Math STAAR scores on treatment. Standard errors clustered at the grade/school level. Controls for II include: age, sex, race, ethnicity, days missed, english learner status , special ed status, free lunch eligibility.**

Table 6

Arlington Experiment, Heterogeneous Treatment Effects for Students in Grades 3-6

	I No Controls	II Grade FEs	III Grade FEs w/ Controls
Full Sample (N=7234)	0.171**	0.172**	0.122**
	(0.069)	(0.070)	(0.058)
Gender			
Female (N=3507)	0.200***	0.201***	0.158**
	(0.072)	(0.071)	(0.061)
Male (N=3727)	0.139*	0.140*	0.085
	(0.075)	(0.076)	(0.063)
P-Value	*0.213*	*0.213*	*0.107*
Race/Ethnicity			
White (N=3794)	0.139*	0.139*	0.081
	(0.075)	(0.075)	(0.061)
Black (N=2015)	0.121	0.140*	0.163**
	(0.085)	(0.084)	(0.082)
Hispanic (N=3434)	0.065	0.068	0.098**
	(0.047)	(0.047)	(0.049)
Non-White (N=3440)	0.192**	0.202**	0.153**
	(0.085)	(0.085)	(0.085)
P-Value	0.527	0.487	0.383
Days Missed			
Days Missed < 5 (N=2018)	0.137	0.137	0.066
	(0.085)	(0.085)	(0.067)
Days Missed 6-25 (N=3742)	0.148**	0.152**	0.119**
	(0.064)	(0.066)	(0.057)
Days Missed > 25 (N=1474)	0.214**	0.215**	0.171**
	(0.084)	(0.082)	(0.070)
P-Value	0.484	0.477	0.432
Economic Indicator			
Free Lunch (N=3521)	0.128*	0.127**	0.110*
	(0.062)	(0.063)	(0.057)
No Free Lunch (N=3713)	0.189**	0.191**	0.133*
	(0.084)	(0.085)	(0.071)
P-Value	*0.349*	*0.368*	*0.369*
ESL Status			
English Second Language (N=1424)	0.093	0.087	0.044
	(0.097)	(0.097)	(0.084)
Non ESL (N=5810)	0.184**	0.188**	0.142**
	(0.072)	(0.072)	(0.058)
P-Value	*0.309*	*0.300*	*0.314*

Notes: **Treatment effects of practice on standardized post test scores are estimated using the model outlined in Equation (2). Treatment effects are calculated for the full sample and various subsamples of the data. Column I is a univariate OLS regression 2022 standardized test scores on treatment, column II adds controls to the regression, and column III is the full equation (2). Standard errors are clustered at the grade-school level. P-values values in table denote whether treatment effects are heterogeneous within subgroup categories.**

Table 7

Arlington Quasi-Experiment, Association with Being in a Classroom with Higher Average Weekly Practice Time and Math STAAR 2022 Test Score, by Classroom, Treated and Control Sample

	Full Sample	Grades 3-6	Grades 7-8
10+ minutes	0.001	0.006	-0.211**
	(0.053)	(0.061)	(0.097)
N	3,500	2,867	633
15+ minutes	0.047	0.033	-0.083
	(0.057)	(0.062)	(0.081)
N	2,755	2,436	319
20+ minutes	0.084	0.056	0.169
	(0.064)	(0.064)	(0.135)
N	2,488	2,309	179
25+ minutes	0.085	0.061	-0.138
	(0.077)	(0.072)	(0.214)
N	2,006	1,957	49
30+ minutes	0.173**	0.144**	-0.359
	(0.075)	(0.068)	(0.229)
N	1,661	1,657	4
35+ minutes	0.248***	0.213***	
	(0.074)	(0.067)	
N	1,387	1,387	
40+ minutes	0.241***	0.209***	
	(0.079)	(0.072)	
N	1,190	1,190	
45+ minutes	0.241***	0.209***	
	(0.079)	(0.079)	
N	1,138	1,138	
50+ minutes	0.202**	0.173**	
	(0.079)	(0.072)	
N	1,022	1,022	
55+ minutes	0.261***	0.229***	
	(0.090)	(0.084)	
N	792	792	
60+ minutes	0.265***	0.225**	
	(0.098)	(0.089)	
N	642	642	
N	*7,916*	*4,801*	*3,115*

Notes: **The independent variable is whether a student's classmates average above the given minutes in weekly Khan Academy practice. Dependent variable is a student's standardized 2022 STAAR score. No class in grades 7-8 averaged about 35 minutes per week. Controls for a student's 2021 STAAR scores and student demographics. Fixed effects for grade, standard errors clustered at the grade/school level.**

Table 8

Arlington Quasi-Experiment, Association with Being in a Classroom with Higher Average Weekly Practice Time and Math STAAR 2022 Test Score, by Classroom, Treated and Control Sample With Grade-School Fixed Effects

	Full Sample	Grades 3-6	Grades 7-8
10+ minutes	-0.047	0.092	-0.288**
	(0.088)	(0.114)	(0.104)
N	3,500	2,867	633
15+ minutes	0.224*	0.391**	0.185*
	(0.118)	(0.164)	(0.091)
N	2,755	2,436	319
20+ minutes	0.352***	0.394***	0.439*
	(0.125)	(0.102)	(0.220)
N	2,488	2,309	179
25+ minutes	0.142	0.190	-0.311**
	(0.190)	(0.170)	(0.137)
N	2,006	1,957	49
30+ minutes	0.0790	0.093	-0.204
	(0.178)	(0.157)	(0.150)
N	1,661	1,657	4
35+ minutes	0.364***	0.334***	
	(0.137)	(0.110)	
N	1,387	1,387	
40+ minutes	0.450**	0.391**	
	(0.212)	(0.164)	
N	1,190	1,190	
45+ minutes	0.450**	0.391**	
	(0.212)	(0.164)	
N	1,138	1,138	
50+ minutes	0.270	0.223	
	(0.189)	(0.162)	
N	1,022	1,022	
55+ minutes	0.387**	0.343**	
	(0.181)	(0.150)	
N	792	792	
60+ minutes	0.340**	0.294**	
	(0.150)	(0.119)	
N	642	642	
N	*7,916*	*4,801*	*3,115*

Notes: **The independent variable is whether a student's classmates average above the given minutes in weekly Khan Academy practice. Dependent variable is a student's standardized 2022 STAAR score. No class in grades 7-8 averaged about 35 minutes per week. Controls for a student's 2021 STAAR scores and student demographics. Fixed effects for grade-school, standard errors clustered at the grade/school level.**

Table 9

Arlington Quasi-Experiment, Association with Being in a Classroom with Higher Average Weekly Level Ups and Math STAAR 2022 Test Score, by Classroom, Treated and Control Sample

	Full Sample	Grades 3-6	Grades 7-8
1+ Level-ups	0.022	-0.015	0.048
	(0.061)	(0.063)	(0.112)
N	2,918	2,621	297
2+ Level-ups	0.215***	0.171***	0.297***
	(0.067)	(0.064)	(0.085)
N	1,789	1,729	60
3+ Level-ups	0.242***	0.206***	
	(0.071)	(0.064)	
N	1,441	1,441	
4+ Level-ups	0.284***	0.236***	
	(0.094)	(0.086)	
N	694	694	
5+ Level-ups	0.349***	0.306***	
	(0.121)	(0.114)	
N	467	467	
N	*7,916*	*4,801*	*3,115*

Notes: **The independent variable is whether a student's classmates average above the given number of Level-ups weekly in Khan Academy. Dependent variable is a student's standardized 2022 STAAR score. No class in grades 7-8 averaged about 3 level-ups per week. Controls for a student's 2021 STAAR scores and student demographics. Fixed effects for grade, standard errors clustered at the grade/school level.**

Table 10

Arlington Quasi-Experiment, OLS Regressions, Student Outcomes on Weekly Practice Time Time and Level-Ups, with Class Fixed Effects and Individual Controls

	Class FE w/ Controls	Class FE w/ Controls	Class FE w/ Controls
G4-6 Sample			
5-25 minutes	0.146***		0.022
	(0.056)		(0.055)
25-50 minutes	0.381***		-0.069
	(0.072)		(0.076)
50+ minutes	0.646***		-0.173*
	(0.084)		(0.094)
1-2 Level-ups		0.353***	0.366***
		(0.046)	(0.48)
2-5 Level-ups		0.654***	0.707***
		(0.049)	(0.055)
5+ Level-ups		1.163***	1.250***
		(0.060)	(0.070)
N = 4,754			

***Notes: O*LS regressions of individual weekly level-ups and weekly practice minutes on standardized 2022 Math STAAR scores. Level-ups is a measure provided by Khan Academy and represents moving from Unfamiliar to Familiar (70% or more correct), or Familiar to Mastered (100% correct) on one exercise. Students who practice between 0-5 minutes a week serve as comparison. Controls include student demographics and 2021 STAAR Math scores. Regressions include class fixed effects. Sample is Grades 4-6 students with KWiK treated teachers.**

Table 11

Correlation Between 35+ Minutes of Practice and Teacher Survey Answers

	Coeff.	Mean
Less than 75% of my students have access to technology at home	0.14	0.23
	(0.12)	
I teach multiple grades.	-0.48***	0.05
	(0.07)	
I have never used Khan Academy before	0.09	0.19
	(0.15)	
My students work on Khan Academy about once per week after school.	-0.57***	0.02
	(0.12)	
My students work on Khan Academy more than once per week after school.	-0.52***	0.04
	(0.12)	
I don't usually assign my students any homework.	-0.28**	0.25
	(0.13)	
My students are expected to independently practice math in school for 90+ minutes each week.	-0.01	0.27
	(0.13)	
My students are expected to practice math after school for 90+ minutes each week.	-0.22	0.04
	(0.26)	
Teacher attended or watched professional development session	0.31***	0.77
	(0.11)	
Teacher held first meeting with coach	0.49***	0.92
	(0.08)	
Teacher held second meeting prior to October 1st, 2021	0.36***	0.89
	(0.14)	

Notes: **This table shows correlation between a teacher's answers to several pre-programming survey questions and the likelihood that that teacher's class averaged 35 minutes or more of weekly practice the following year. The Mean column gives the proportion of teachers that responded in the affirmative to each question. Grade 3-6 teachers only, as no grade 7-8 classrooms average above 35 minutes of practice.**

Table 12

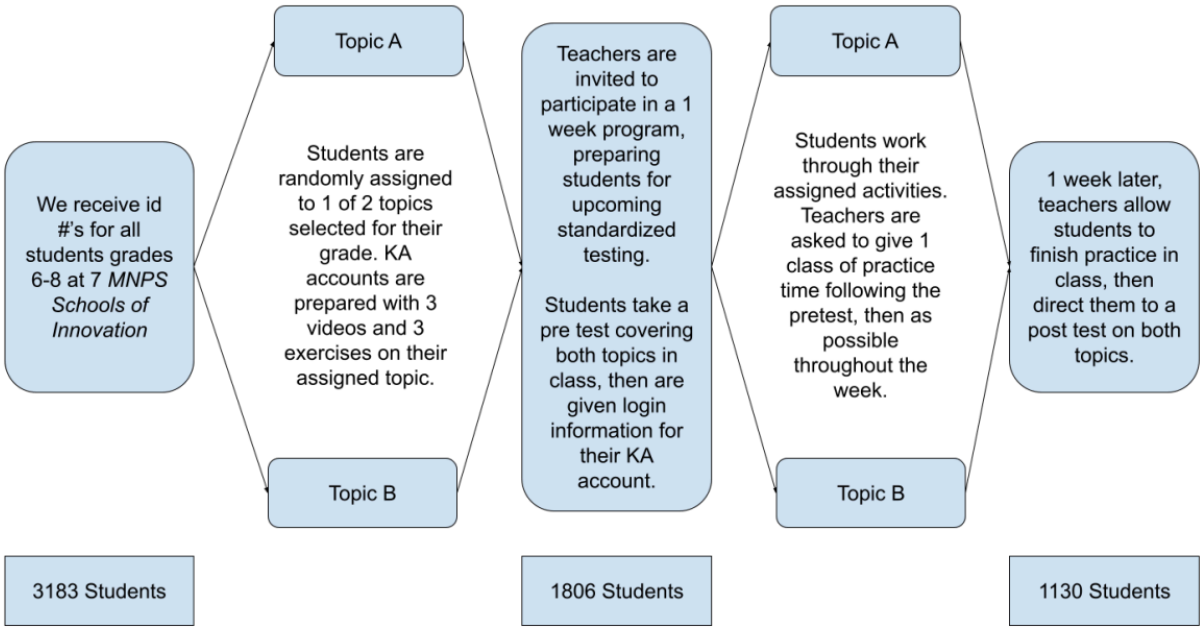
Arlington Experiment, Treatment on Treated (TOT) Estimates Under Alternative Assumptions for Treatment Effectiveness

	Grades 3-6 No control	Grades 3-6 Control	Full Sample No Control	Full Sample Control
ITT Results	0.171**	0.122**	0.036	0.025
	(0.069)	(0.058)	(0.092)	(0.076)
N	7,234	7,234	10,979	10,979
Teacher Met At Least Once	0.183**	0.131**	0.043	0.029
	(0.075)	(0.063)	(0.098)	(0.082)
N	7,234	7,234	10,979	10,979
Met Once and Scheduled Second Meeting	0.208**	0.150**	0.084	0.048
	(0.086)	(0.072)	(0.107)	(0.091)
N	7,234	7,234	10,979	10,979
Average Class Practice Time ≥ 20	0.289**	0.208**	0.356***	0.198*
	(0.122)	(0.103)	(0.131)	(0.106)
N	7,234	7,234	10,979	10,979
Average Class Practice Time ≥ 25	0.330**	0.236**	0.425***	0.247**
	(0.137)	(0.116)	(0.149)	(0.120)
N	7,234	7,234	10,979	10,979
Average Class Practice Time ≥ 30	0.359**	0.258**	0.469***	0.278**
	(0.141)	(0.121)	(0.159)	(0.125)
N	7,234	7,234	10,979	10,979
Average Class Practice Time ≥ 35	0.428***	0.306**	0.562***	0.331**
	(0.165)	(0.141)	(0.188)	(0.146)
N	7,234	7,234	10,979	10,979

Notes: **The table shows the 2SLS regressions of standardized 2022 Math STAAR scores on treatment instrumenting "received treatment" with "offered treatment". A classroom is considered to have been "offered treatment" if they were selected for the intervention. Two binary instruments corresponding to being offered treatment and being in grades 3-6 and being offered treatment and being in grades 7-8 are used. Rows 2-7 represent various definitions of having "received treatment" and row 1 shows the intent to treat effects. The control specification includes teacher fixed effects and controls for age, sex, race, ethnicity, days missed, english learner status, special ed status, and free lunch eligibility. Standard errors are clustered at the grade/school level.**

Figure 1

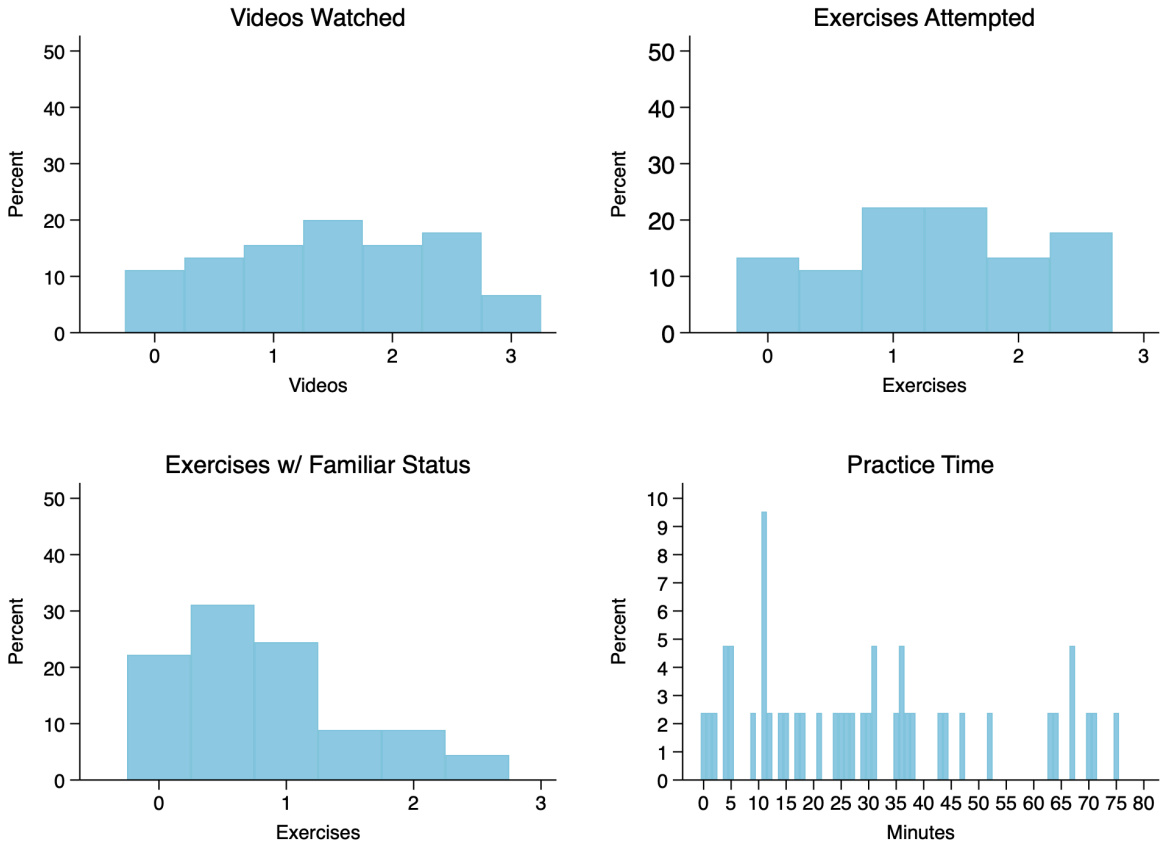
Nashville Experiment, Enrollment and Randomization Design



Notes: **Student scores on practiced topics act as treatment observations and scores on unpracticed scores serve as control. Pre and post tests contain 3 questions on each topic, with topics chosen by school administrators to be familiar but not mastered material. Randomized test banks ensured students had similar but different questions from classmates and between pre and post tests. Students are randomized at the individual level.**

Figure 2

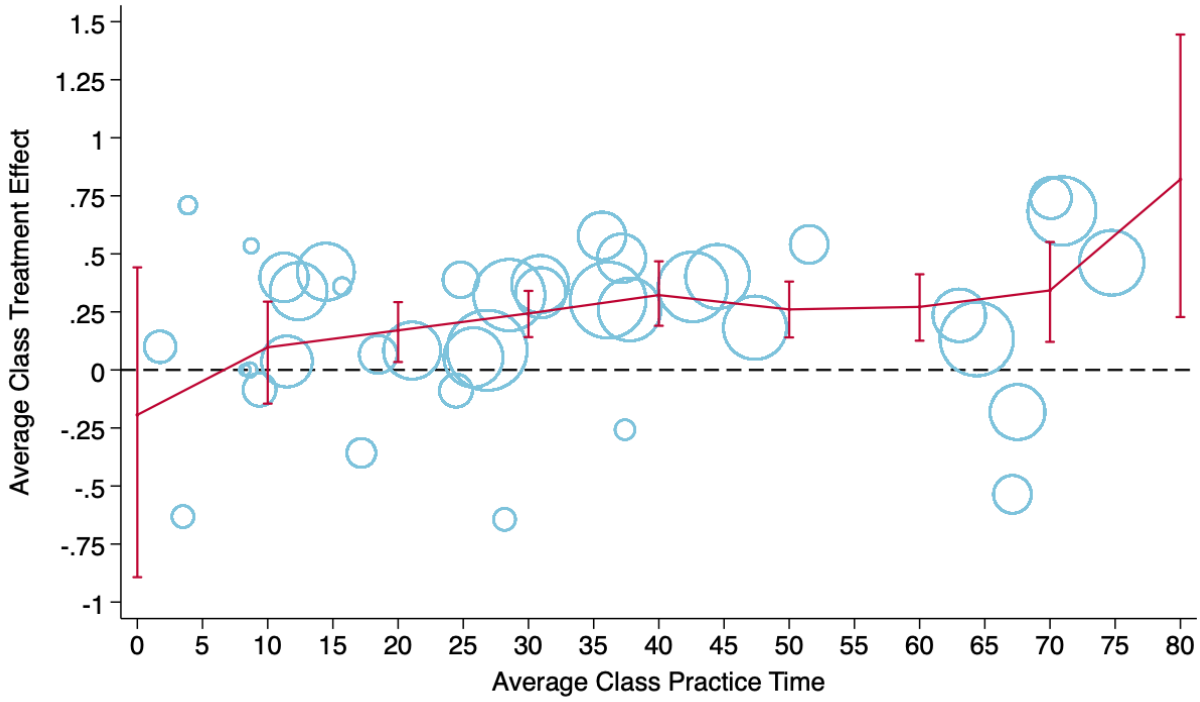
Nashville Experiment, Class Practice Distribution



Notes: **Distribution of class average number of videos watched, exercises attempted, exercises with familiar status, and minutes spent using Khan Academy. Represents 45 classrooms, without weighting by size. Familiar status represents a score above 70%.**

Figure 3

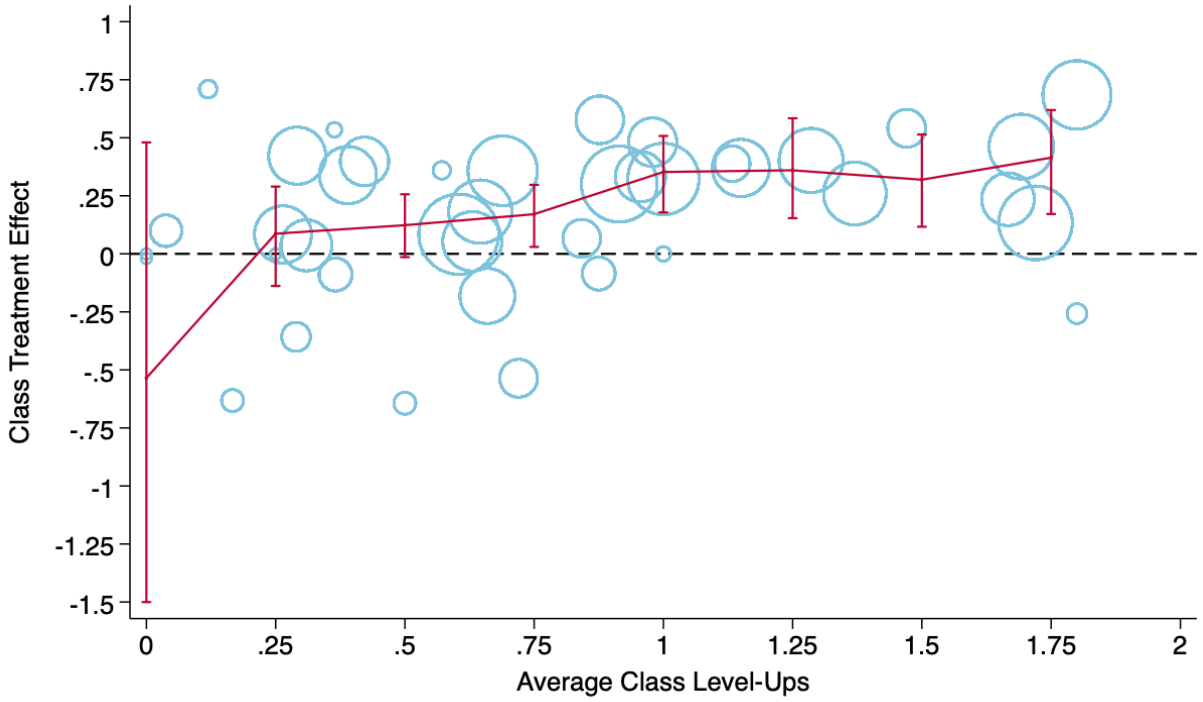
Nashville Experiment, Treatment Effects of Class Practice Time on Standardized Post Test Scores, by Classroom



Notes: **This figure plots the treatment effects of practice on standardized post test scores for each classroom, using the model outlined in Equation (1). These values are plotted against the average practice time for each classroom. Each hollow circle denotes the treatment effect for an individual classroom. Weighted by class size, circle size is relative to how many students there are in a given classroom that have a post test score. Outlier effects of ± 1 have been removed from the plot in addition to outlier classrooms with 110+ minutes of weekly practice time. The red line represents the nonparametric regression of class average residualized standardized math STARR score for 2022 regressed on weekly average class time. The mean classroom treatment effect calculated for the nonparametric regression is 0.25 with standard errors of 0.047. Both the nonparametric regression line and the confidence intervals are obtained via bootstrapping. The corresponding OLS regression slope is 0.027 with standard errors of 0.0020 and an intercept of 0.11.**

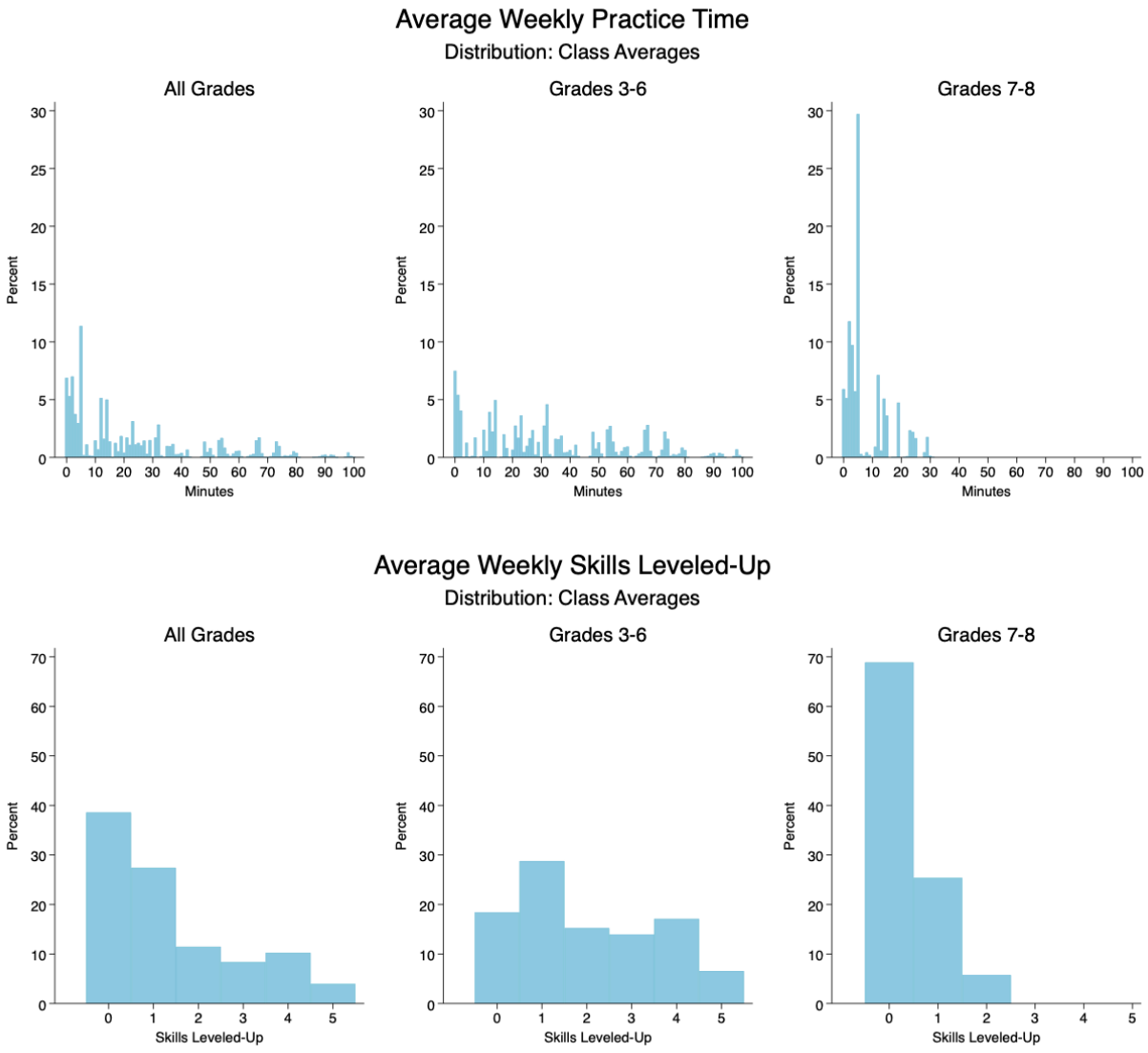
Figure 4

Nashville Experiment, Treatment Effects of Level-Ups on Standardized Post Test Scores, by Classroom



Notes: **The figure plots treatment effects of practice on standardized post test scores, conditional on classroom, using the model outlined in Equation (1). These values are plotted against average level-ups for each classroom. Each hollow circle denotes the average treatment effect for an individual classroom. Weighted by class size, circle size is relative to how many students there are in a given classroom that have a post test score. The red line represents the nonparametric regression of class treatment effects on weekly average class level-ups. The mean classroom treatment effect calculated for the nonparametric regression is 0.24 with standard errors of 0.044. Both the nonparametric regression line and the confidence intervals are obtained via bootstrapping. Outlier treatment effects of ± 2 were removed from the plot in addition to one outlier classroom with > 2 level-ups. The corresponding OLS regression slope is 0.219 with standard errors of 0.0847 and an intercept of -0.00116.**

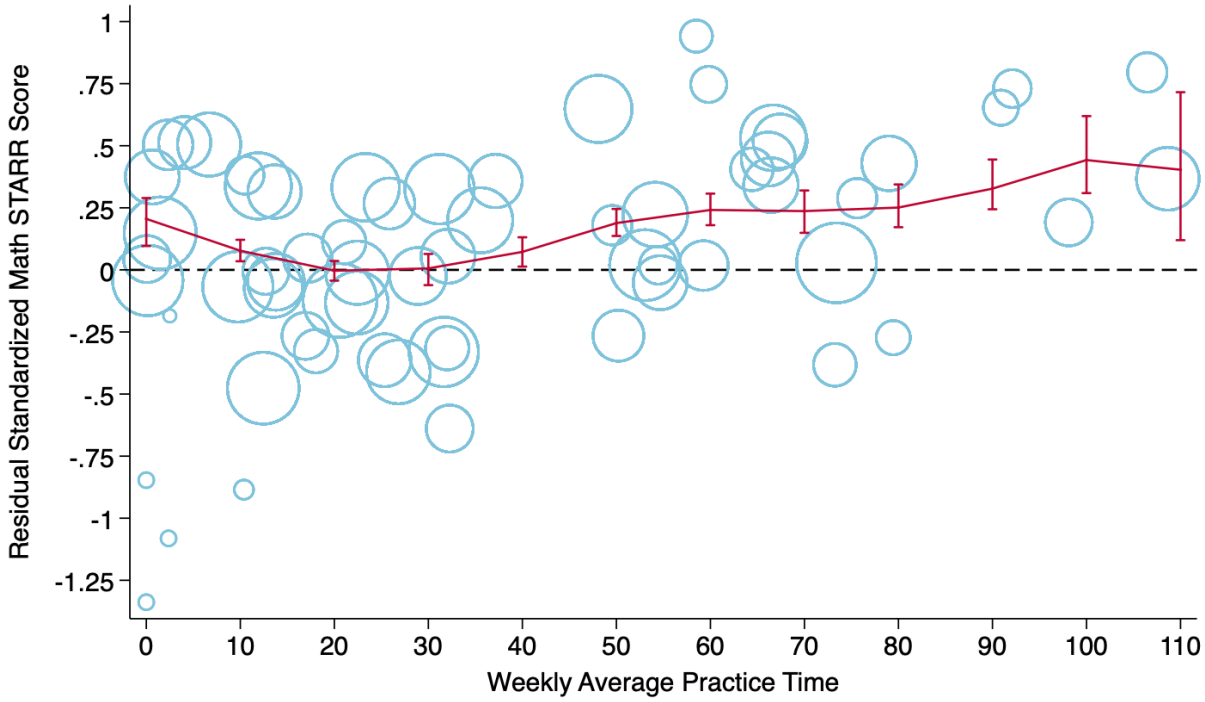
Figure 5
Arlington Experiment, Average Weekly Practice Time and Level-Ups by Treated Classroom



Notes: **The first three figures show the distribution of each treated classroom's average student weekly practice time on Khan Academy, averaged over the school year. The last three figures show the distribution of each classroom's average weekly number of Khan Academy exercise level-ups, averaged over the school year. The first row shows these distributions for the entire Grades 3-8 sample. The middle row and last row shows the distributions for Grades 3-6 and Grades 7-8 respectively.**

Figure 6

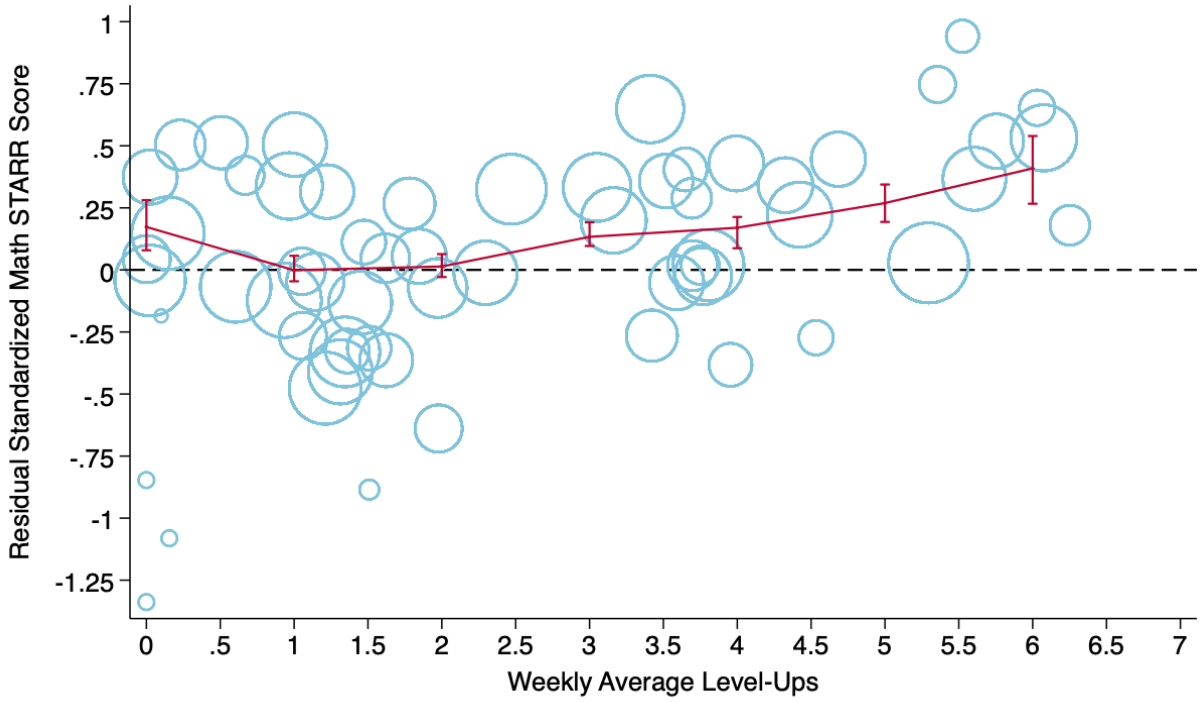
Arlington Quasi-Experiment, Average Weekly Practice Time and Math STAAR Residualized Standardized 2022 Test Score, by Classroom, Treated Sample, Grades 4-6



Notes: **This figure plots each Grade 4 to 6 treated classroom's average weekly practice time on Khan Academy and the corresponding average students' residual standardized Math STAAR 2022 test score after regressing a student's score on their previous year's score and obtaining predicted residuals. Outlier effects of ± 1 have been removed from the plot in addition to outlier classrooms with 110+ minutes of weekly practice time. Weighted by class size, circle size is relative to how many students there are in a given classroom that have a 2022 test score. The red line represents the nonparametric regression of class average residualized standardized math STARR score for 2022 regressed on weekly average class time. The mean classroom treatment effect calculated for the nonparametric regression is 0.13 with standard errors of 0.018. Both the nonparametric regression line and the confidence intervals are obtained via bootstrapping. The slope from the corresponding OLS regression of residual scores on weekly average class practice time is 0.0032 with standard errors of 0.0011 and an intercept of -0.12.**

Figure 7

Arlington Quasi-Experiment, Average Classroom Weekly Total Level Ups and Math STAAR Residualized Standardized 2022 Test Score, by Classroom, Treated Sample, Grades 4-6



Notes: **This figure plots each Grade 4 to 6 treated classroom's average weekly Level Ups on Khan Academy and the corresponding average students' residual standardized Math STAAR 2022 test score after regressing a student's score on their previous year's score and obtaining predicted residuals. Outlier effects of ± 1 in addition to classrooms with 6.5+ weekly average level ups have been removed from the plot. Weighted by class size, circle size is relative to how many students there are in a given classroom that have a 2022 test score. The red line represents the nonparametric regression of class average residualized standardized math STARR score for 2022 regressed on weekly average class level-ups. The mean classroom treatment effect calculated for the nonparametric regression is 0.13 with standard errors of 0.018. Both the nonparametric regression line and the confidence intervals are obtained via bootstrapping. The slope of the corresponding OLS regression of residual scores on weekly average class level-ups is 0.62 with standard errors of 0.016 and an intercept of -0.57.**

Appendix

Appendix Table A1

Nashville Experiment, Participant Assignment Progression Among Classrooms with the Highest and Lowest Average Practice Times, but With Samples Truncated to Equalize Prescore Means

		Exercise 1			Exercise 2			Exercise 3		
	N	Watched	Attempted	Familiar	Watched	Attempted	Familiar	Watched	Attempted	Familiar
	(number)	*(fraction of total)*			*(fraction of total)*			*(fraction of total)*		
By average progression among classmates										
Bottom 5%	75	0.53	0.41	0.20	0.28	0.25	0.09	0.19	0.15	0.03
Top 5%	79	0.95	0.92	0.78	0.81	0.78	0.71	0.75	0.68	0.61
By average progression among classmates: Top Matched to Bottom Mean Prescore										
Bottom 5%	75	0.53	0.41	0.20	0.28	0.25	0.09	0.19	0.15	0.03
Top 5%	55	0.96	0.91	0.76	0.78	0.71	0.65	0.65	0.62	0.53
By average progression among classmates: Bottom Matched to Top Mean Prescore										
Bottom 5%	47	0.53	0.40	0.23	0.28	0.23	0.09	0.17	0.13	0.04
Top 5%	79	0.95	0.92	0.78	0.81	0.78	0.71	0.75	0.68	0.61

Notes: **This table shows the percentage of participants who watched each video, attempted each exercise, and achieved "familiar" status (70% correct or higher) on each exercise. For each student, we also calculate the average number of exercises that their classmates became familiar with (not factoring in a student's own practice). By looking at practice completion for students in the top and bottom 5% of this measure, we can see that practice differs drastically by a student's classmates (i.e. class assignment). The first row shows the results from Table 2. The second row shows the results when the top 24 pre-score grades from the top 5% group are dropped to match the mean prescore value of the bottom 5% group. The third row shows the results when the bottom 28 pre-score grades are dropped from the bottom 5% group to match the mean prescore value of the top 5% group. When compared to Table 2, it can be seen the large gap in results remains even after pre-score mean values are matched for both groups.**

Appendix Table A2

Arlington Quasi-Experiment, Association with Being in a Classroom with Higher Average Weekly Level Ups and Math STAAR 2022 Test Score, by Classroom With Grade-School Fixed Effects, Treated and Control Sample

	Full Sample	Grades 3-6	Grades 7-8
1+ Level-ups	0.143	-0.066	0.405***
	(0.121)	(0.119)	(0.112)
N	2,918	2,621	297
2+ Level-ups	0.246**	0.155	0.571***
	(0.138)	(0.159)	(0.086)
N	1,789	1,729	60
3+ Level-ups	0.281**	0.270***	
	(0.120)	(0.100)	
N	1,441	1,441	
4+ Level-ups	0.259	0.245	
	(0.252)	(0.208)	
N	694	694	
5+ Level-ups	0.609***	0.575***	
	(0.157)	(0.109)	
N	467	467	
N	*7,916*	*4,801*	*3,115*

Notes: **The independent variable is whether a student's classmates average above the given number of Level-ups weekly in Khan Academy. Dependent variable is a student's standardized 2022 STAAR score. No class in grades 7-8 averaged about 3 level-ups per week. Controls for a student's 2021 STAAR scores and student demographics. Fixed effects for grade-school, standard errors clustered at the grade/school level.**

Appendix Table A3

Arlington Quasi-Experiment, Effects of Being in a Classroom with 35+ Minutes of Average Practice Heterogeneity by Prior Performance

	Full Sample		Grades 3-6	
	Controls w/ Grade FEs	Controls w/ Grade-school FEs	Controls w/ Grade FEs	Controls w/ Grade-school FEs
Full Sample	0.248***	0.364***	0.213***	0.334***
	(0.074)	(0.137)	(0.067)	(0.110)
N	*7,916*	*7,916*	*4,801*	*4,801*
>50% STAAR 2021	0.207***	0.193**	0.167***	0.193***
	(0.066)	(0.089)	(0.059)	(0.061)
N	*4,076*	*4,076*	*2,489*	*2,489*
<50% STAAR 2021	0.193***	0.373**	0.191**	0.342**
	(0.074)	(0.152)	(0.074)	(0.148)
N	*3,840*	*3,480*	*2,312*	*2,312*

Notes: **Treatment effects of practice on standardized post test scores are estimated using the model outlined in Equation (3). Equation (3) utilizes grade-school fixed effects and the regression with only grade fixed effects represents a simplified equation. Standard errors are clustered at the grade-school level. This table breaks down the Arlington Quasi-Experiment treatment effects by prior performance, which is represented by a student's STARR 2021 score. Students were divided into two STARR 2021 groups based on the median value. Columns 1 and 2 show the results for the full sample and columns 3 and 4 show the results for the sample only containing students in grades 3-6.**

Appendix Table A4

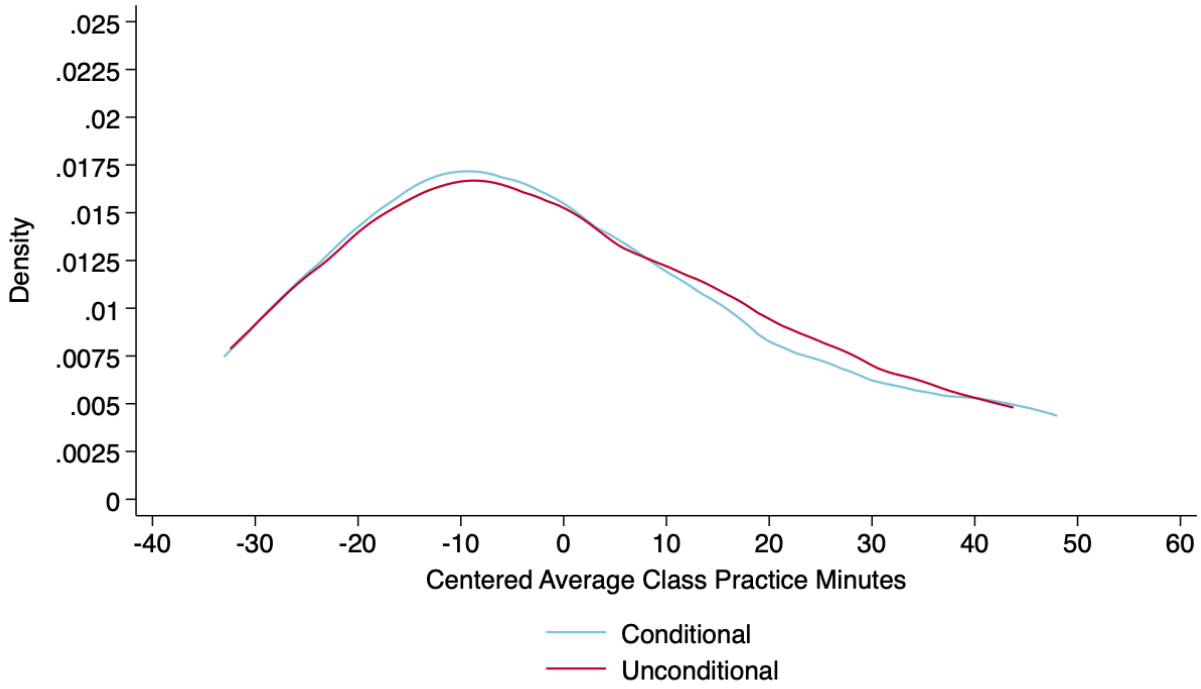
Arlington Quasi-Experiment, Summary of 2022 practice, cut by 2021 test scores (Grades 4-6)

	2021 Math STAAR Scores		
2022 Weekly Averages	Below Median	Above Median	Difference
Individual Practice Time	16.37	22.56	6.191***
Individual Level-ups	0.90	1.96	1.060***
Classmate Practice Time	18.11	20.72	2.608***
Classmate Average Practice >35 Minutes	0.17	0.22	0.055***
Classmate Level-ups	1.27	1.57	0.301***
Classmate Average Level-ups >3	0.19	0.25	0.058***

Notes: **This figure shows the average of various student characteristics categorized by being above or below the 2021 math STAAR test score median. Above median students recorded more practice and had classmates with higher practice times and level-ups. N=4801.**

Appendix Figure A1

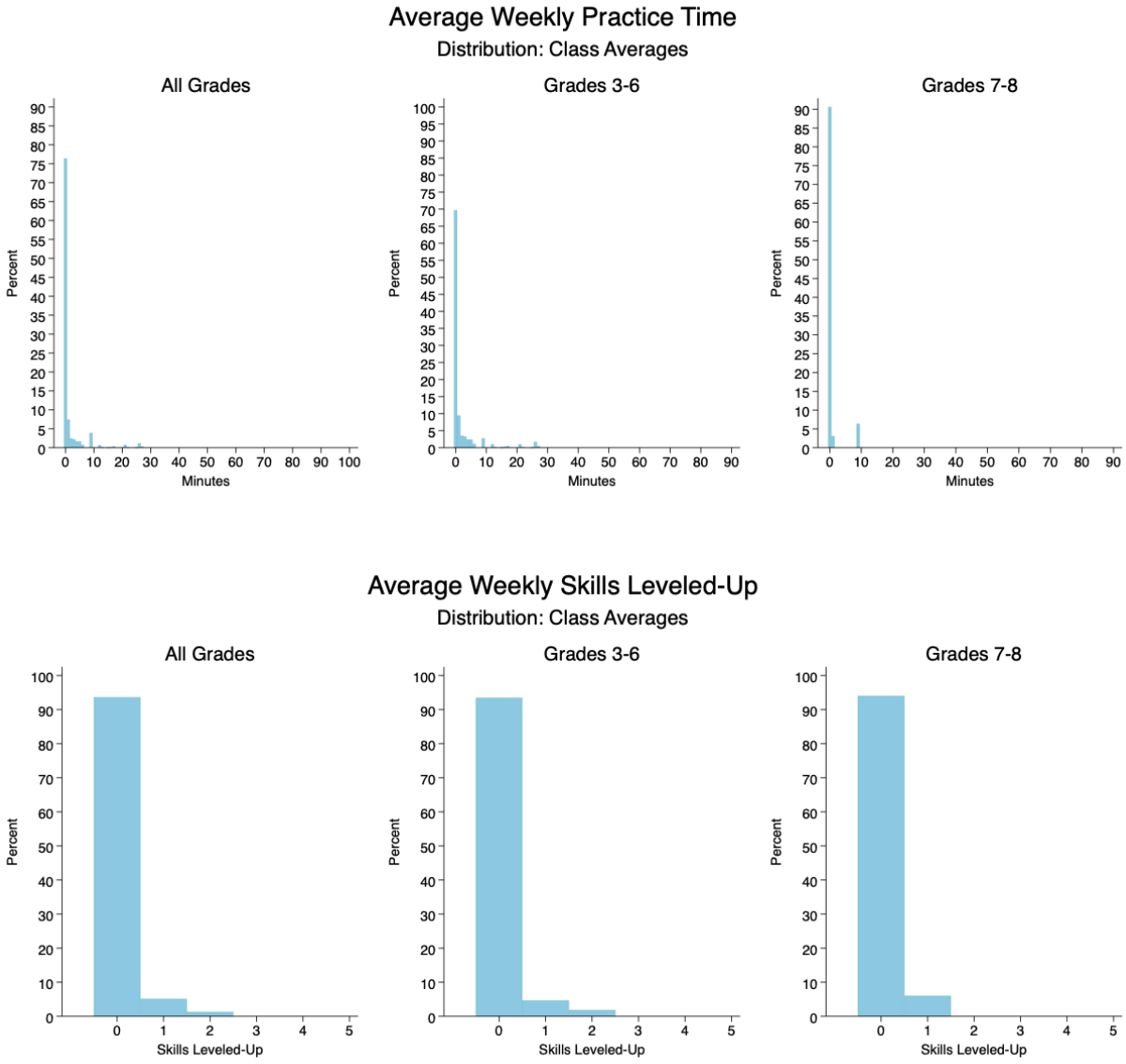
Nashville Experiment, Kernel Density of Demeaned Average Classroom Practice Time, With and Without Conditioning on Prior Student Background Characteristics



Notes: **The figure plots the kernel density of demeaned average classroom practice time, with and without conditioning prior on student background characteristics. The red line represents the unconditional kernel density plot and the blue line represents the conditional plot. Conditions include standardized pretest scores, gender, race, and special education status.**

Appendix Figure A2

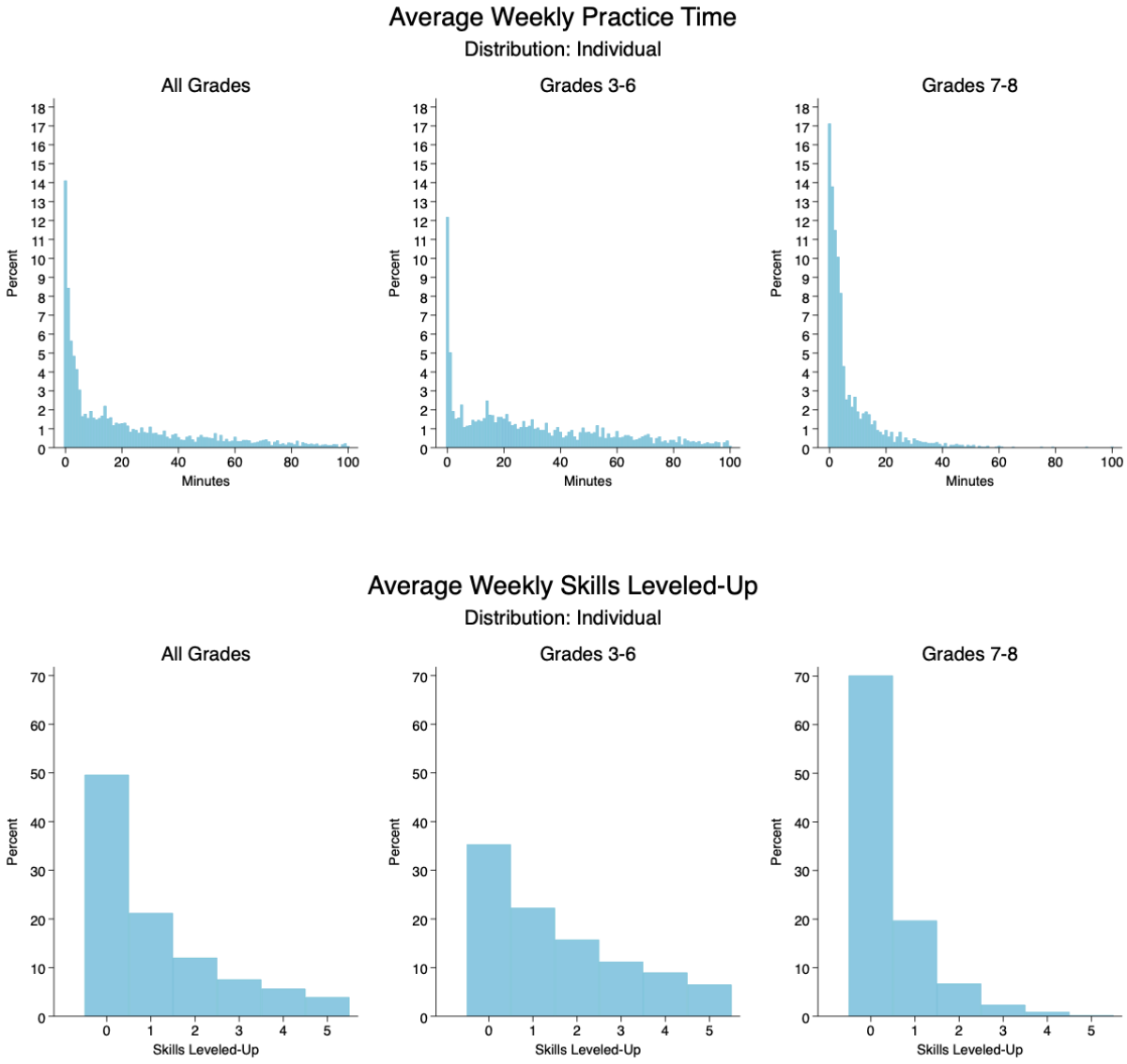
Arlington Experiment, Average Weekly Practice Time and Level-Ups by Control Classroom



Notes: **The first three figures show the distribution of each control classroom's average student weekly practice time on Khan Academy, averaged over the school year. The last three figures show the distribution of each classroom's average weekly number of Khan Academy exercise level-ups, averaged over the school year. The first row shows these distributions for the entire Grades 3-8 sample. The middle row and last row shows the distributions for Grades 3-6 and Grades 7-8 respectively.**

Appendix Figure A3

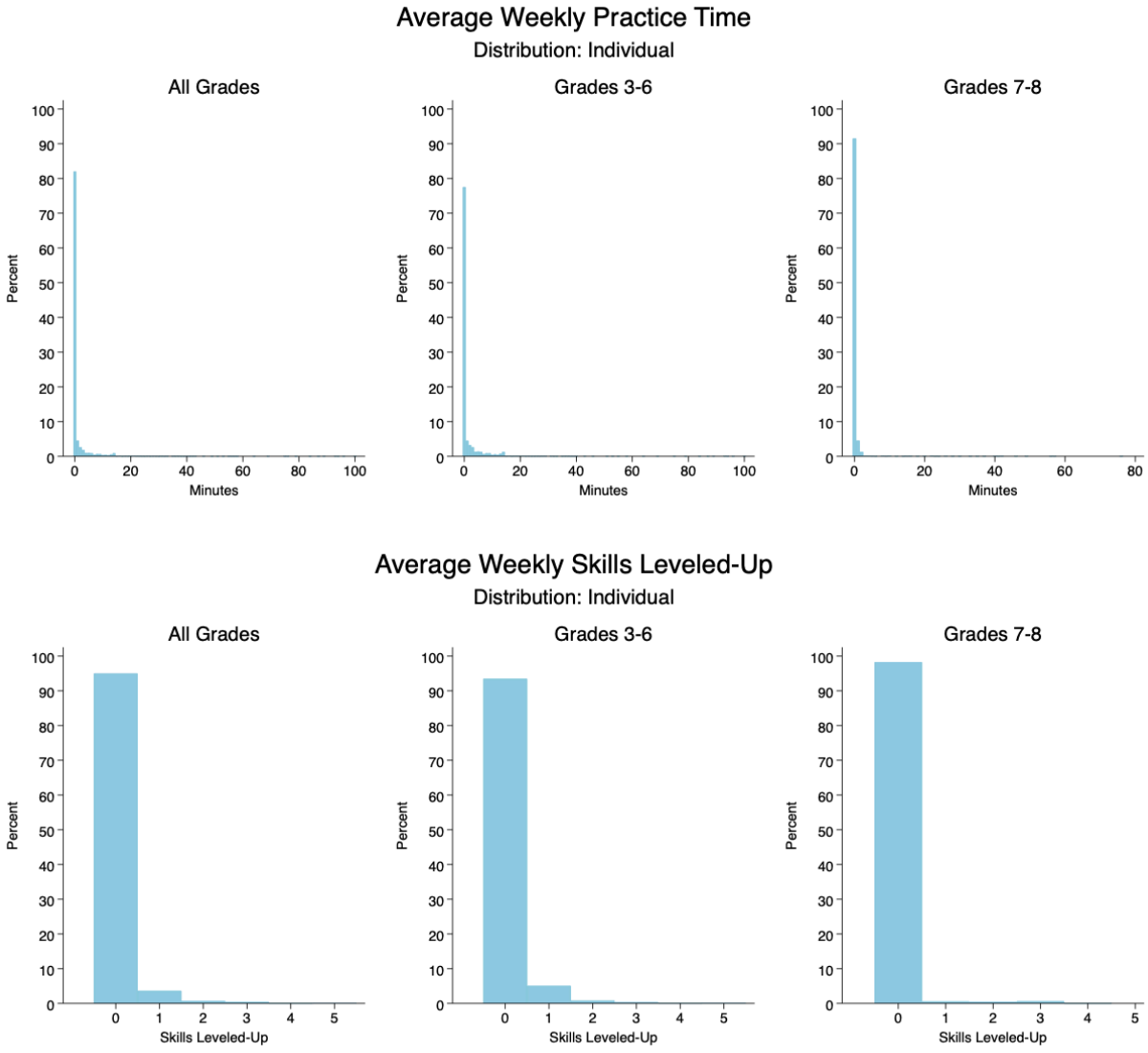
Arlington Experiment, Average Weekly Practice Time and Level-Ups by Treated Student



Notes: **The first three figures show the distribution of students' average weekly practice time on Khan Academy, averaged over the school year. The last three figures show the distribution of students' average weekly number of Khan Academy exercise level-ups, averaged over the school year. The first row shows these distributions for the entire Grades 3-8 treated sample. The middle row and last row shows the distributions for the Grades 3-6 and Grades 7-8 treated samples respectively.**

Appendix Figure A4

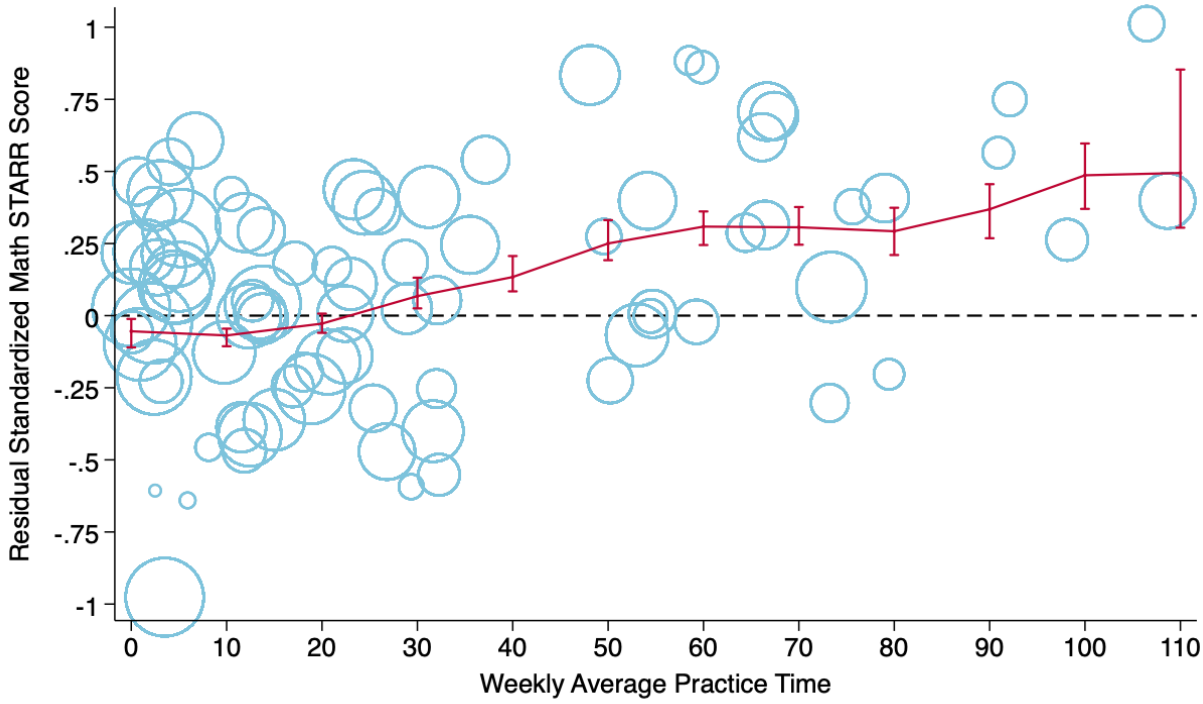
Arlington Experiment, Average Weekly Practice Time and Level-Ups by Control Student



Notes: **The first three figures show the distribution of students' average weekly practice time on Khan Academy, averaged over the school year. The last three figures show the distribution of students' average weekly number of Khan Academy exercise level-ups, averaged over the school year. The first column shows these distributions for the entire Grades 3-8 treated sample. The middle column and last column shows the distributions for the Grades 3-6 and Grades 7-8 treated samples respectively.**

Appendix Figure A5

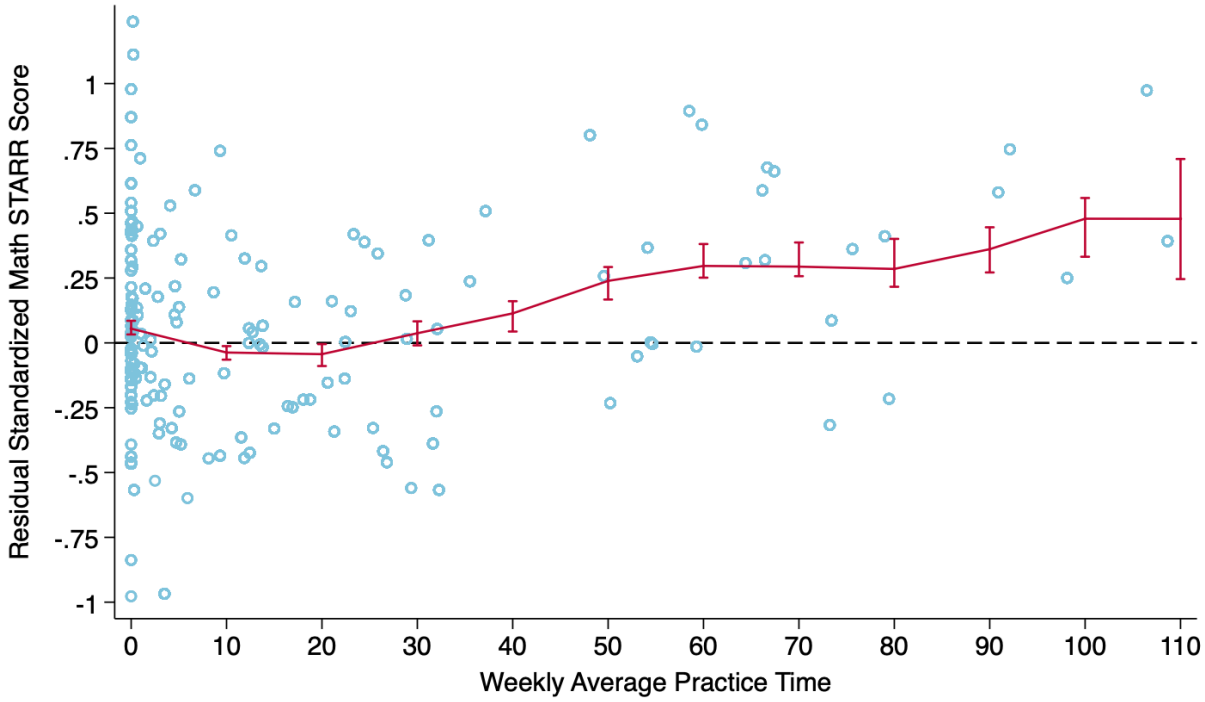
Arlington Quasi-Experiment, Average Weekly Practice Time and Math STAAR Residualized Standardized 2022 Test Score, by Classroom, Treated Sample, Grades 4-8



Notes: **This figure plots each Grade 4 to 8 treated classroom's average weekly practice time on Khan Academy and the corresponding average students' residual standardized Math STAAR 2022 test score after regressing a student's score on their previous year's score and obtaining predicted residuals. Outlier effects of ± 1 have been removed from the plot in addition to outlier classrooms with 110+ minutes of weekly practice time. Weighted by class size, circle size is relative to how many students there are in a given classroom that have a 2022 test score. The red line represents the nonparametric regression of class average residualized standardized math STARR score for 2022 regressed on weekly average class time. The mean classroom treatment effect calculated for the nonparametric regression is 0.035 with standard errors of 0.016. Both the nonparametric regression line and the confidence intervals are obtained via bootstrapping. The slope from the corresponding OLS regression of residual scores on weekly average class practice time is 0.0058 with standard errors of 0.0016 and an intercept of -0.11.**

Appendix Figure A6

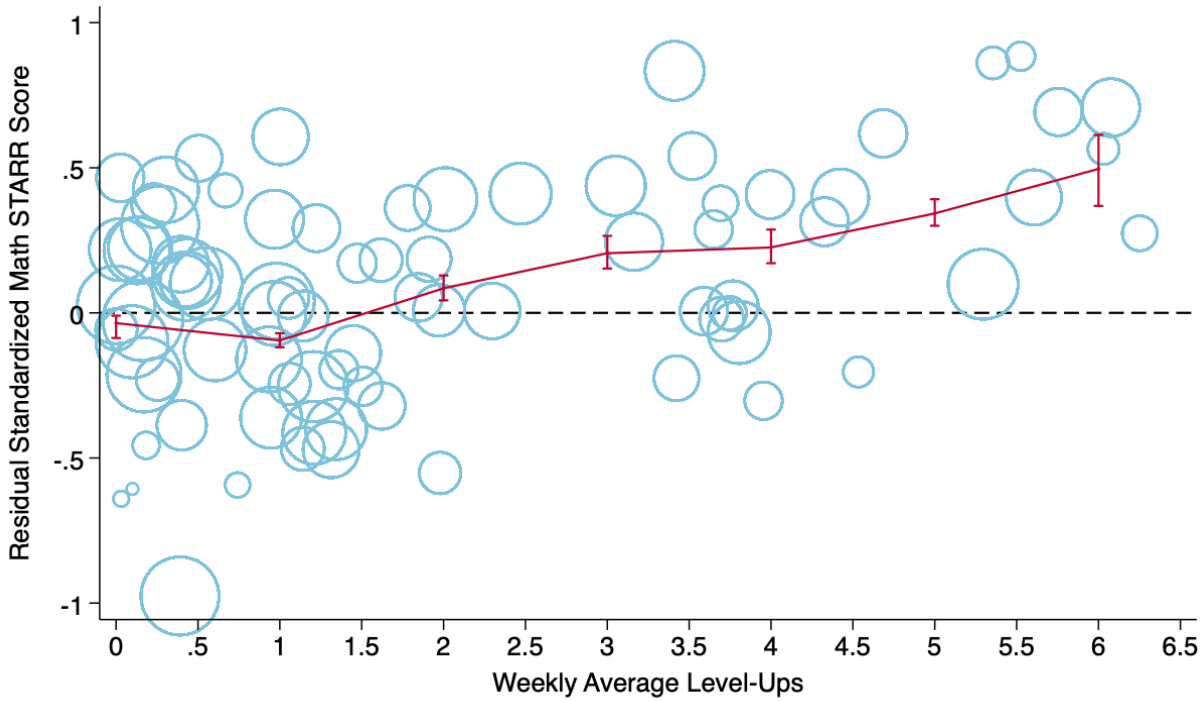
Arlington Quasi-Experiment, Average Weekly Practice Time and Math STAAR Residualized Standardized 2022 Test Score, by Classroom, Treated and Control Sample, Grades 4-8



Notes: **This figure plots each Grade 4 to 8 treated and control classroom's average weekly practice time on Khan Academy and the corresponding average students' residual standardized Math STAAR 2022 test score after regressing a student's score on their previous year's score and obtaining predicted residuals. Outlier effects of ± 1 have been removed from the plot in addition to outlier classrooms with 110+ minutes of weekly practice time. The red line represents the nonparametric regression of class average residualized standardized math STARR score for 2022 regressed on weekly average class time. The mean classroom treatment effect calculated for the nonparametric regression is 0.051 with standard errors of 0.010. Both the nonparametric regression line and the confidence intervals are obtained via bootstrapping. The slope from the corresponding OLS regression of residual scores on weekly average class practice time is 0.0037 with standard errors of 0.0012 and an intercept of -0.017.**

Appendix Figure A7

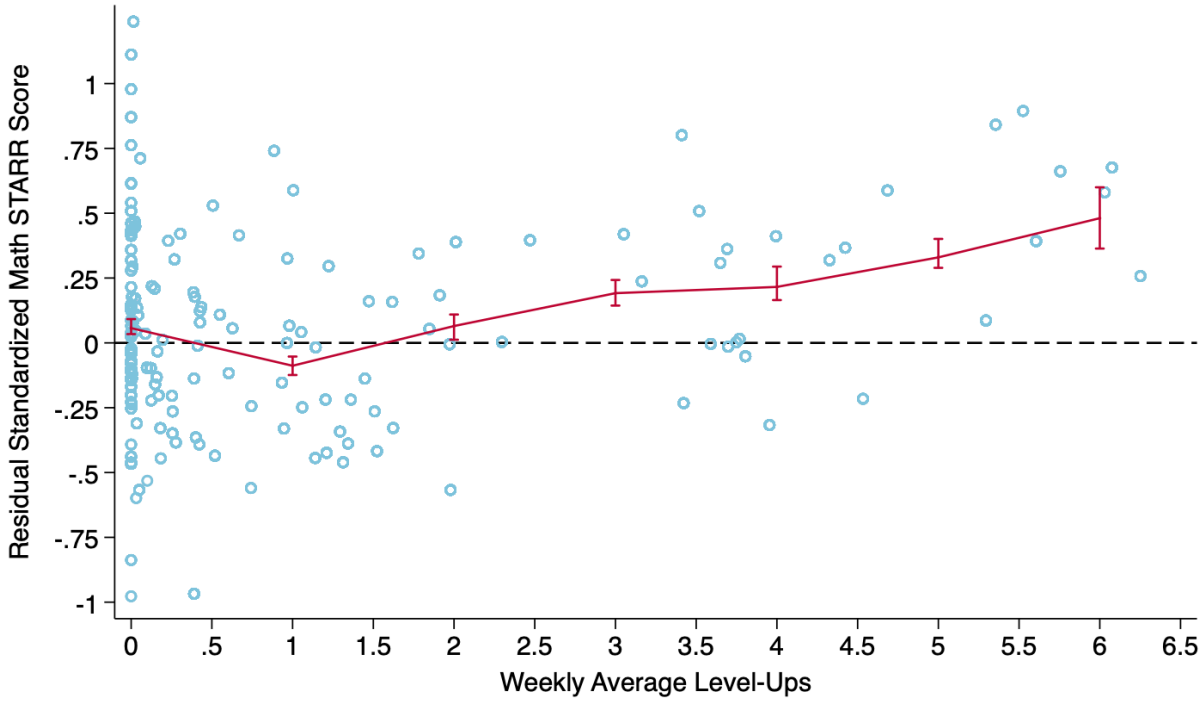
Arlington Quasi-Experiment, Average Classroom Weekly Total Level Ups and Math STAAR Residualized Standardized 2022 Test Score, by Classroom, Treated Sample, Grades 4-8



Notes: **This figure plots each Grade 4 to 8 treated classroom's average weekly Level Ups on Khan Academy and the corresponding average students' residual standardized Math STAAR 2022 test score after regressing a student's score on their previous year's score and obtaining predicted residuals. Outlier effects of ± 1 in addition to classrooms with 6.5+ weekly average level ups have been removed from the plot. Weighted by class size, circle size is relative to how many students there are in a given classroom that have a 2022 test score. The red line represents the nonparametric regression of class average residualized standardized math STARR score for 2022 regressed on weekly average class level-ups. The mean classroom treatment effect calculated for the nonparametric regression is 0.041 with standard errors of 0.015. Both the nonparametric regression line and the confidence intervals are obtained via bootstrapping. The slope from the corresponding OLS regression of residual scores on weekly average class level-ups is 0.093 with standard errors of 0.020 and an intercept of -0.14.**

Appendix Figure A8

Arlington Quasi-Experiment, Average Classroom Weekly Total Level Ups and Math STAAR Residualized Standardized 2002 Test Score, by Classroom, Treated and Control Sample, Grades 4-8



Notes: **This figure plots each Grade 4 to 8 treated and control classroom's average weekly Level Ups on Khan Academy and the corresponding average students' residual standardized Math STAAR 2022 test score after regressing a student's score on their previous year's score and obtaining predicted residuals. Outlier effects of ± 1 in addition to classrooms with 6.5+ weekly average level ups have been removed from the plot. The red line represents the nonparametric regression of class average residualized standardized math STARR score for 2022 regressed on weekly average class level-ups. The mean classroom treatment effect calculated for the nonparametric regression is 0.051 with standard errors of 0.010. Both the nonparametric regression line and the confidence intervals are obtained via bootstrapping. The slope from the corresponding OLS regression of residual scores on weekly average class level-ups is 0.061 with standard errors of 0.016 and an intercept of -0.027.**