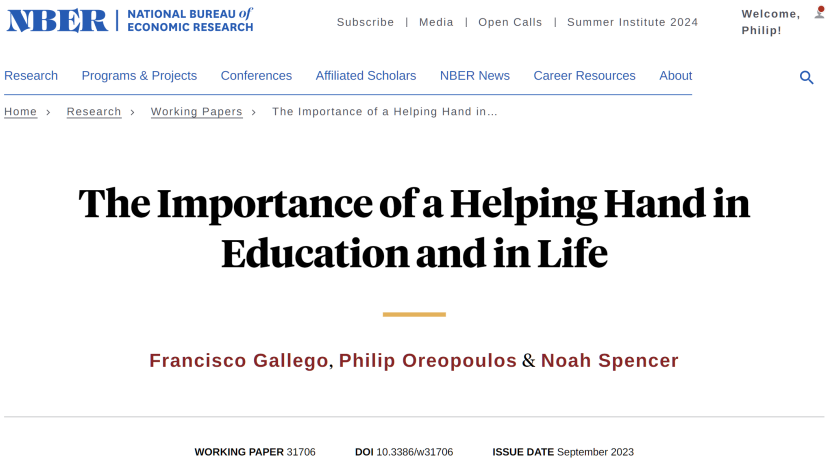
Personalized Support at Scale

2024 CLEF Keynote Lecture

Philip Oreopoulos
University of Toronto
National Bureau of Economic Research
Jameel Poverty Action Lab

Papers related to talk



NBER | NATIONAL BUREAU of ECONOMIC RESEARCH

Subscribe | Media | Open Calls | Summer Institute 2024

Welcome, Philip!

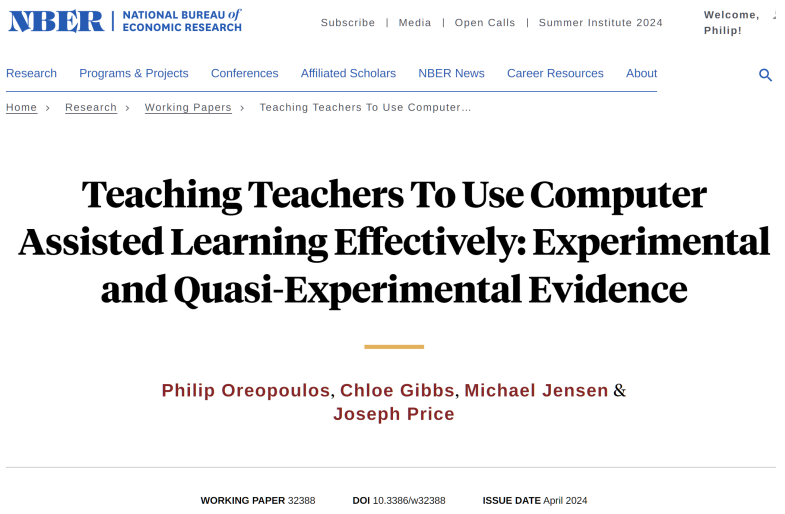
Research Programs & Projects Conferences Affiliated Scholars NBER News Career Resources About

Home › Research › Working Papers › The Importance of a Helping Hand in...

The Importance of a Helping Hand in Education and in Life

Francisco Gallego, Philip Oreopoulos & Noah Spencer

WORKING PAPER 31706 DOI 10.3386/w31706 ISSUE DATE September 2023



NBER | NATIONAL BUREAU of ECONOMIC RESEARCH

Subscribe | Media | Open Calls | Summer Institute 2024

Welcome, Philip!

Research Programs & Projects Conferences Affiliated Scholars NBER News Career Resources About

Home › Research › Working Papers › Teaching Teachers To Use Computer...

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

Philip Oreopoulos, Chloe Gibbs, Michael Jensen & Joseph Price

WORKING PAPER 32388 DOI 10.3386/w32388 ISSUE DATE April 2024

Cuhna Heckman (2007) Human Capital Production Function

$$S_{i,t} = F(S_{i,t-1}, I_{i,t})$$

- $S_{i,t}$ is a vector of current skills for student i at time t
- $S_{i,t-1}$ is a vector of earlier skills for student i at time $t-1$
- $I_{i,t}$ is skill investment: practice, support

Practice by individual (time, effort, focus)
Support from outside (assistance or instruction from teacher, parent, peer, tutor, counselor, etc...)

What are skills good for? Who cares?

- We can foster skills to increase productivity to increase pay
- We can also use skills to contribute directly to adult well-being (e.g. skills to benefit health, social interactions, become 'street smart')
- But who cares?
  - children?
  - parents?

Behavioral biases Affecting Practice

- Decisions involving immediate costs and long-term, uncertain, benefits, challenging
- Children are especially adverse to effort, failure
- Prefer not to practice or do homework
- Present biased
- Projection biased
- Salience biased

Behavioral Policies to Help Children Practice

- Offset immediate costs with immediate benefits
  - offer immediate rewards (e.g. praise, money)
  - make desirable action easier (e.g. easier applications)
- Reduce immediate opportunity costs
  - remove distractions (phones)
  - restrict TV, meeting friends until homework finished
  - create structure where only practice can occur, like classroom setting, compulosry schooling
- Emphasize benefits (e.g. reminders, motivation)

Limitations with Nudging Practice without Support

- Nudging practice without support is cheap: e.g. reminders, notifications, information notices, online encouragement, simplification
- But not very effective (DellaVigna and Linos, 2022), especially for continuous activities and habits
- Imagine classroom only for practice, but without support
- Better for one-time actions (e.g. application assistance) where child already wants action to get done

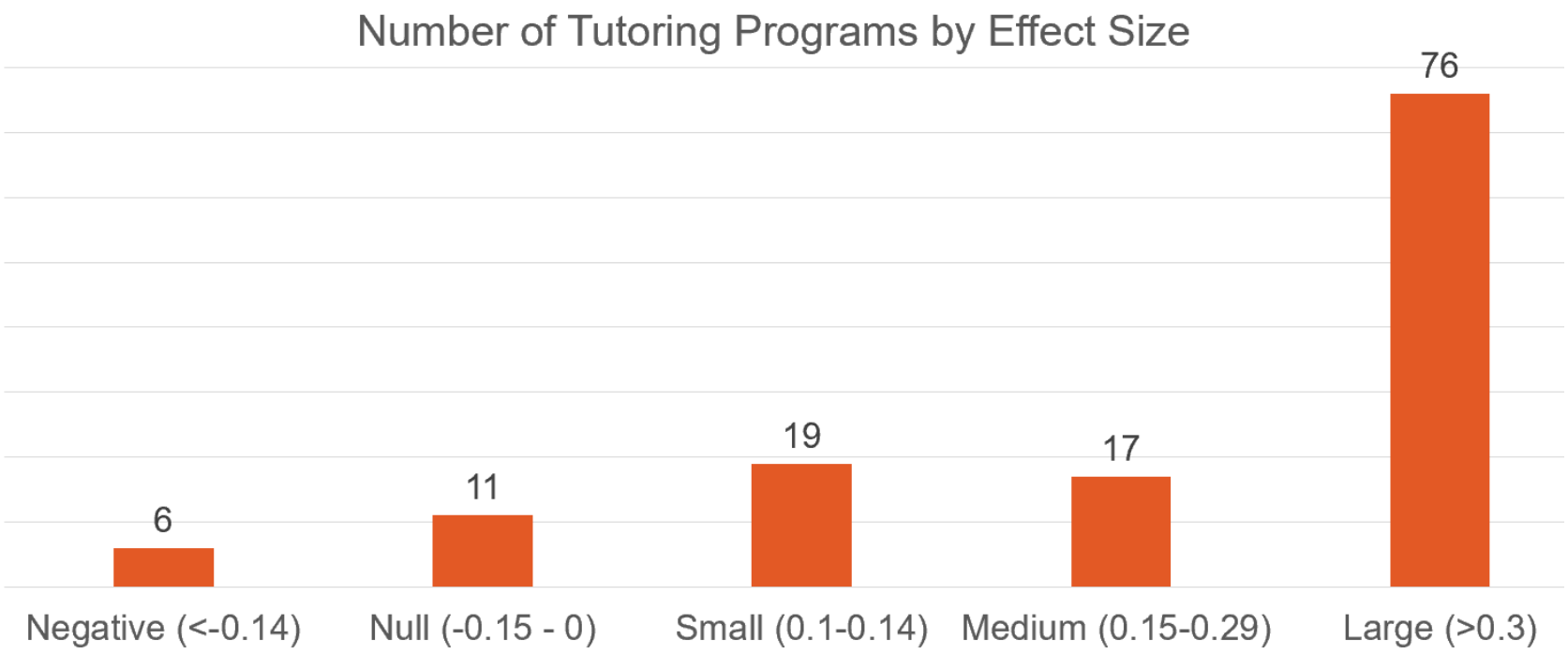
More effective education interventions involve practice AND support

- Ideal setup:
  - **Intensive support**: 1 teacher for 1 student
  - **Continuous support**: in the room, daily, sufficient time, offering feedback, available to answer questions
  - **High quality support**: e.g. clear instruction, mastery approach, encouragement, feedback
  - **Isolated support**: away from distraction

Tutoring

- One-on-one instructional programming by teachers, paraprofessionals, volunteers, or parents, allows students to progress at their own pace and receive immediate feedback and attention
- Tutoring overall benefits are impressive in a variety of settings
- A review of 96 randomized trials found consistent and substantial positive impacts on learning outcomes, with an average effects size of 0.29 standard deviation increase in test scores (usually english or math)

Nickow, Oreopoulos, and Quan, 2023



Avg. impact larger for programs 1) during school; 2) with teacher or paraprofessional as tutor; 3) 1:1; 4) 3-5 days a week

But even with 1:3 tutors, 1-2 days a week, after school and volunteer tutors, avg. effect size >0.2SD

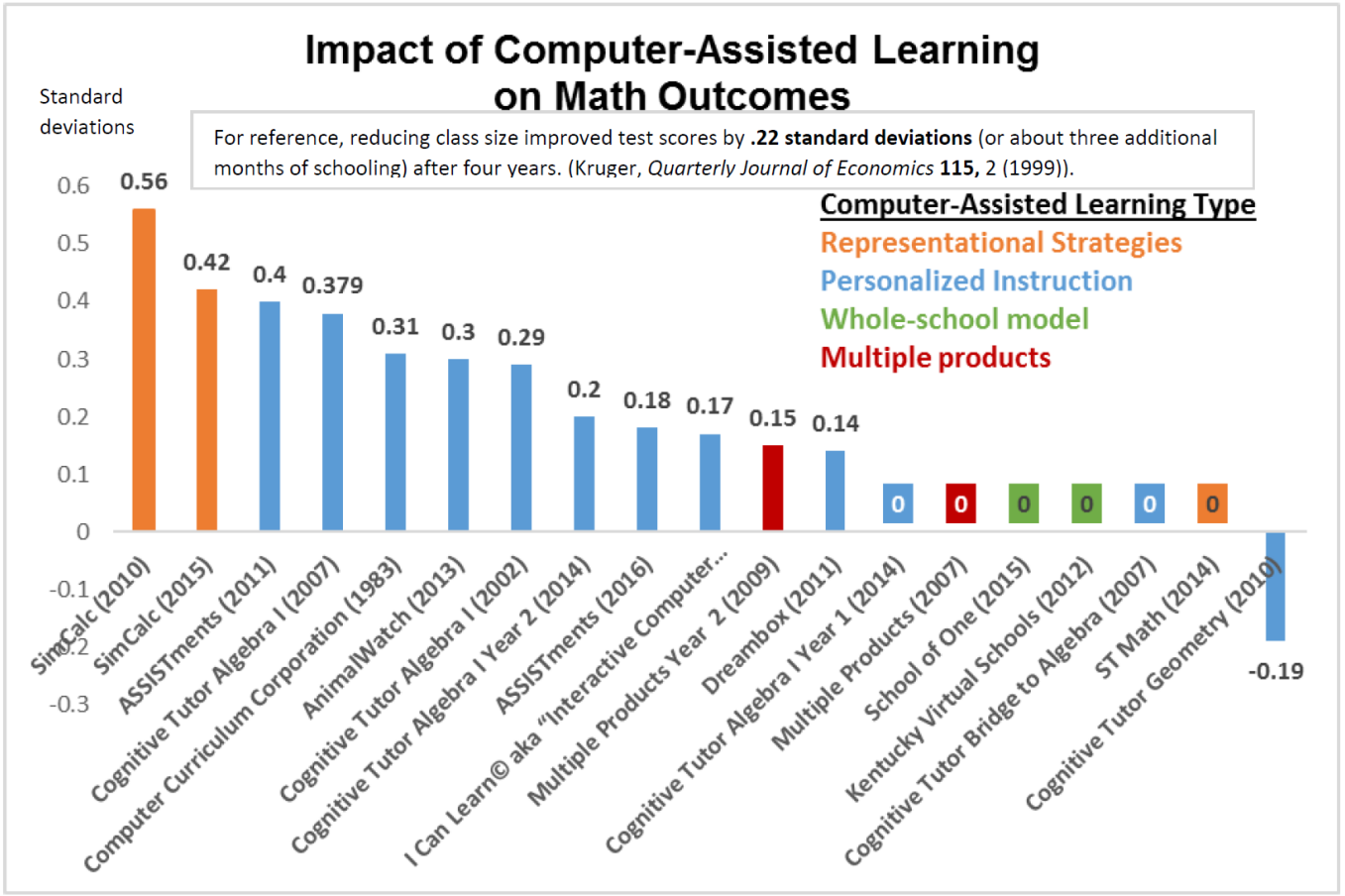
Challenges with Scaling Tutoirng

- Still expensive (thousands of dollars per student)
- Tutor supply limited and of varying quality
- Difficult to determine when and how to provide
- Low demand and low structure for tutoring outside of school
- Too many subjects to target (e.g. math, reading, science)

Alternative Solution for Promosting Support and Practice: School-based Computer Assisted Learning (CAL)

- Computer Assisted Technology CAL offers potential for simulating tutoring experience
- Example: Khan Academy
  - Roadmap of incremental short math videos and exercises to follow for Grades 3 to 12
  - Can progress at own pace
  - Receive immediate feedback and help
  - Mastery approach - keep trying until successfully completing questions
  - Roadmap can be customized for each student
  - Teacher, parents, tutors can observe student progress and respond if student is stuck

Escuta, Nickow, Oreopoulos, and Quan, 2020



12 of 19 RCTs of math CAL find impacts 0.14-0.56 SD, avg 0.18 SD

Challenges with scaling CAL

- Students not motivated to use it - low take-up outside of school (e.g. Little support - Beg et al. 2022))
- Incorporating technology requires a change to the education production function by those delivering education
- Teachers and parents not familiar with it, concerned about additional effort and time costs to learn it, skeptical to adopt compared to using previous year's curriculum, too busy
- Sometimes limited computer access, in school and at home

Overall Goal

- Provide intensive personalized support to facilitate effective practice that gets as close to 1:1 student:teacher ratio as possible
- Utilize technology AND support to facilitate high dosage instruction, time, effort, and quality feedback

Proposed Solution: Teach Teachers How to Use CAL Effectively

- Motivate, educate, and scaffold teachers for using CAL effectively as part of their curriculum.
- Key ingredient is assigning them a coach ('Khoach') to help train and support them throughout the year
- "Khoaches" meet with teachers weekly to set goals, troubleshoot issues, and discuss best practices
- They also help monitor and interpret data, identify to teachers students that need attention
- Khollaborators help construct a roadmap of KA activities and videos for students to follow incrementally. Students work at their own pace
- Program is flexible: khollaborator works with teacher to modify roadmap, use assignments, and/or customize roadmap for each student

Teacher instructions of KWiK

- Create a mastery goal (a roadmap for students to follow) - a set of exercises for your students to follow at their own pace throughout the quadmester
- Create an environment that encourages students to work on their mastery goal at least an hour a week
- Demonstrate at least one 'live exercise' to the class each week to remind students the importance of making mistakes and trying again
- Monitor progress, help those struggling and praid improvement
- Provide in-class time for KA activities to allow for an opportunity to observe students struggling the most
- Communicate with parents

Road Map for Rest of this Presentation

- **Experiment 1: one week student-level 'proof of concept' experiment in Nashville TN**
  - test scores improve from KA practice, but only in classrooms with sufficienty high average practice times
- **Experiment 2: one school year teacher-level field experiment in Arlington TX**
  - test scores improve for elementary school students but no or neg. impact for middle school students - main difference appears to be less time to practice in middle school
- **Quasi-experimental evidence from 'as good as random' assignment to Arlington teachers**
  - test scores improve for students in classrooms where teacher facilitates sufficiently long average practice times
- **Next Steps: LLM virtual tutors...**

Nashville Experiment Setting

- In March 2022, a trial experiment to measure Khan Academy's ability to teach was run in 7 Metro Nashville Public Middle Schools (Schools of Innovation) with about 50 math teachers.
- 'Proof of concept' goal of demonstrating 'week in the life of KWiK' could improve math performance
- 3183 potential participants in grades 6-8 were randomly assigned 1 of 2 grade-specific topics to learn over the course of a week.
- During class, students took a pre-test on both topics, and they were then given 1 hour of Khan Academy activities and videos on their topic to complete that week.
- 1 week later, teachers gave class time to wrap up KA practice and students took a similar post test.

Nashville Experiment Setting

Hello!
This activity is designed to help reinforce and improve your ability to understand some key math concepts, and to help get you ready for your upcoming TCAP Achievement Tests.

Enter your MNPS student number below to log in, or click here to watch a short instuctional video.

Student ID:

Re-Enter Student ID:

Next

Nashville Experiment Setting

This activity is designed to help reinforce and improve your ability to understand some key math concepts, and to help get you ready for your upcoming TCAP Achievement Tests.

1) Take a 6-question quiz to test your initial knowledge of the material.

- If you don't know a question, type **"Skip"** and move on.
- You won't need a calculator, but you may want scratch paper to work on.

2) Use the given username and password to log into Khan Academy and complete a series of videos and exercises.

Happy learning!

Next

Nashville Experiment Setting

Solve for h.

17 + 4h + 2 = 1 - 5h

Enter your answer as an integer.

Next

Grade 8, Topic A: Solving Equations
(Topics selected by *MNPS SOI's* Lead Math Instructional Specialist)

Nashville Experiment Setting

METRO NASHVILLE PUBLIC SCHOOLS

leo WILSON SHEEHAN LAB FOR ECONOMIC OPPORTUNITIES

Now let's practice on Khan Academy!

1) Put on your headphones

2) Click HERE to sign into Khan Academy. Below are a specific username and password that you should use to sign in (you may need to log out of another account first)
Username:
Password:

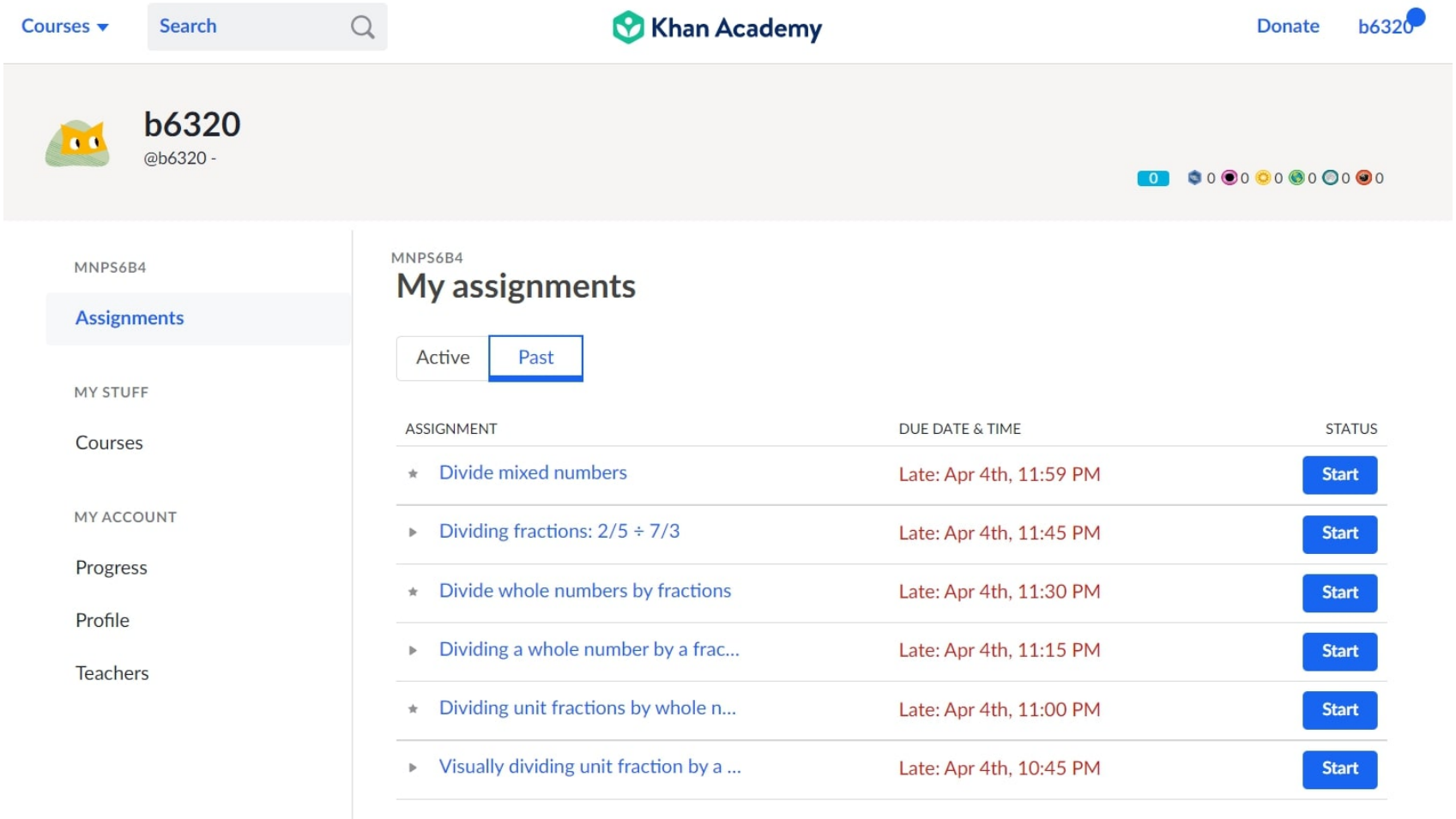
3) Start at the top and proceed sequentially until you complete the last exercise. Use paper and pen or pencil if you need to.

4) Don't feel frustrated if you have to try an exercise more than once before levelling up - learning from mistakes is how we improve.

5) Ask your teacher for help if you get stuck on an exercise or need a reminder of the instructions.

If you need more time, you can use free time this week to finish your practice.

Nashville Experiment Setting



Grade 6, Topic B: Dividing Fractions

MNPS: Progress through Practice

	N	Exercise 1 Watched	Exercise 1 Attempted	Exercise 1 Familiar	Exercise 2 Watched	Exercise 2 Attempted	Exercise 2 Familiar	Exercise 3 Watched	Exercise 3 Attempted	Exercise 3 Familiar
Grade 6										
Topic 1	191	0.75	0.72	0.66	0.68	0.65	0.40	0.54	0.40	0.22
Topic 0	184	0.59	0.56	0.41	0.49	0.43	0.27	0.69	0.49	0.22
Grade 7										
Topic 1	222	0.66	0.67	0.45	0.52	0.37	0.14	0.36	0.30	0.22
Topic 0	204	0.72	0.68	0.50	0.54	0.50	0.19	0.41	0.33	0.15
Grade 8										
Topic 1	172	0.79	0.64	0.39	0.58	0.53	0.32	0.51	0.53	0.38
Topic 0	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
Average Classmate # Familiar										
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: Using only students with both pre and posttests. Familiar status signifies score above 70%. The activities were presented in order, alternating between videos and exercises to progress through practice. Videos were between 2-7 minutes and exercises consisted of 4 or 7 exercises.

MNPS: Model

$$Y_{it} = \beta_0 + \beta_1 T_{it} + \beta_2 X_{it} + \epsilon_{it}$$

- 2 observations for each student, one for their how they performed on treated topic and one for control.
- Y_{it} is the standardized posttest score for student i on topic t
- X_{it} is the standardized pretest score for student i on topic t
- T_{it} is the treatment status of topic t for student i
- ϵ_{it} is the error term, clustered at the student level

Note: We produce an alternative model that includes a student fixed effect. This does not significantly change coefficients or significance.

MNPS: Results

	(1)	(2)	(3)	(4)
		By Grade		
	Full Sample	Grade 6	Grade 7	Grade 8
Treatment	0.22***	0.29***	0.09*	0.30***
	(0.04)	(0.06)	(0.06)	(0.07)
N	2260	750	852	658

	(5)	(6)	(7)	(8)
	By Average Class Practice			
	<5 Mins.	5-25 mins.	25-50 mins.	50-100 mins.
Treatment	0.06	0.12	0.25***	0.28***
	(0.23)	(0.07)	(0.05)	(0.07)
N	50	480	1156	566

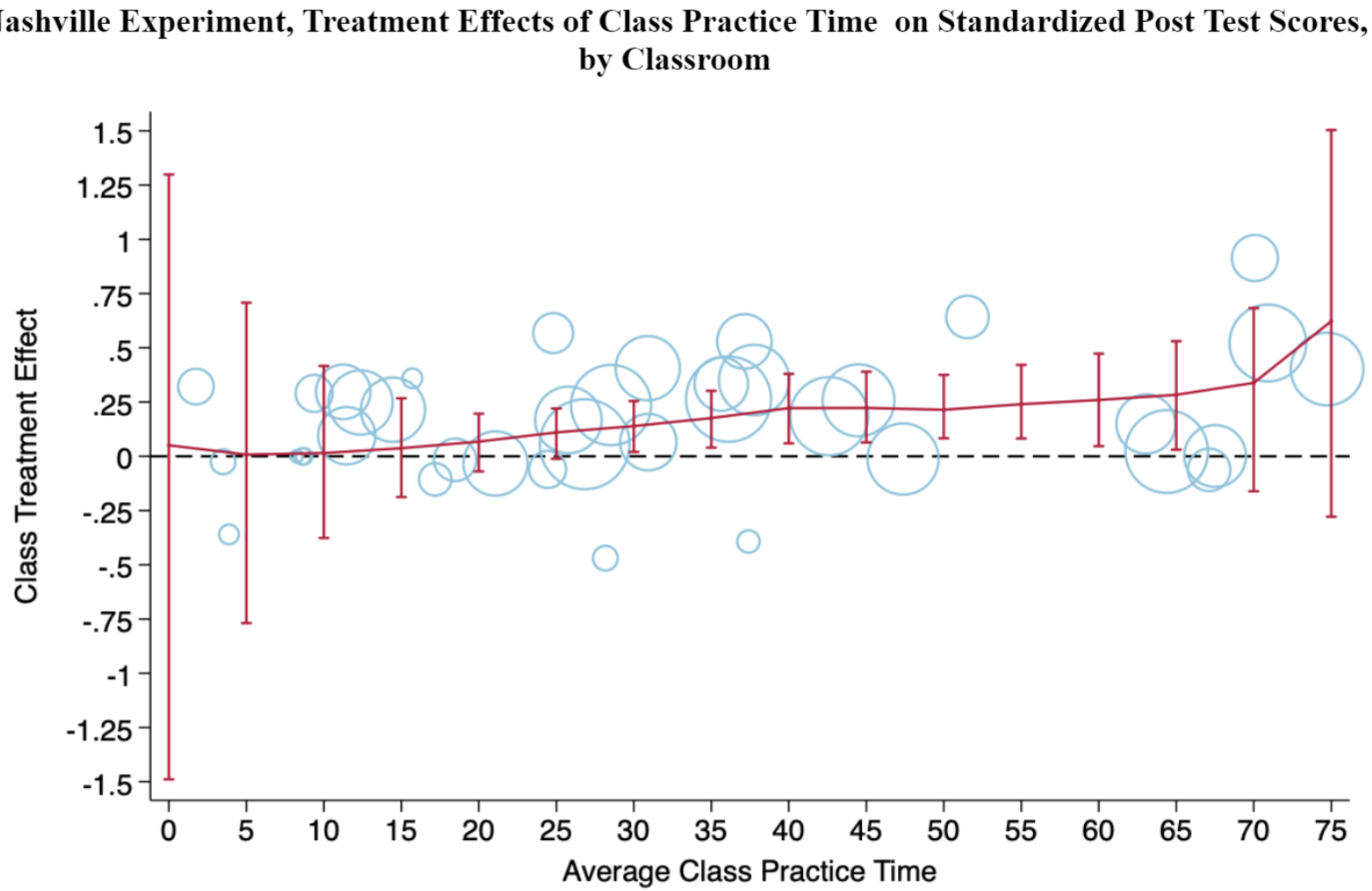
	(9)	(10)	(11)	(12)
	By Average Class Level-ups			
	<0.4 skills	0.4-0.8 skills	0.8-1.2 skills	>1.2 skills
Treatment	0.07	0.09	0.31***	0.36***
	(0.08)	(0.06)	(0.07)	(0.07)
N	386	622	586	666

	(13)	(14)	(15)
	By # of Activities Mastered		
	Exercise 1	Exercises 1-2	Exercises 1-3
Treatment	0.38***	0.43***	0.52***
	(0.05)	(0.07)	(0.08)
N	1110	640	436

Notes: Each student represents 2 observations. Scores on practiced and unpracticed topics are used for treatment and control, respectively. OLS estimations on the standardized posttest scores, with SEs clustered at the student level. All regressions control for standardized pretest scores. Class averages calculated leaving out the individual student.

With Student FEs | Lower bound | Upper bound

Avg. Effects by Classroom Practice Time



Takeaways from Nashville Experiment

- Using KA for reinforcing increasingly challenging math concepts significantly improves performance
- KWiK fidelity is critical: teachers need to facilitate sufficient practice time for students to master material (try again until levelling up)
- Effect sizes are impressively large when this happens (similar to high dosage tutoring)
- Caveat 1: short time horizon between practice and test
- Caveat 2: Counterfactual is not practicing, either because a) student was assigned to a different topic; b) time intended for practice used inefficiently (either due to teacher or student); c) teacher chose to use time differently

Arlington Experiment Setting

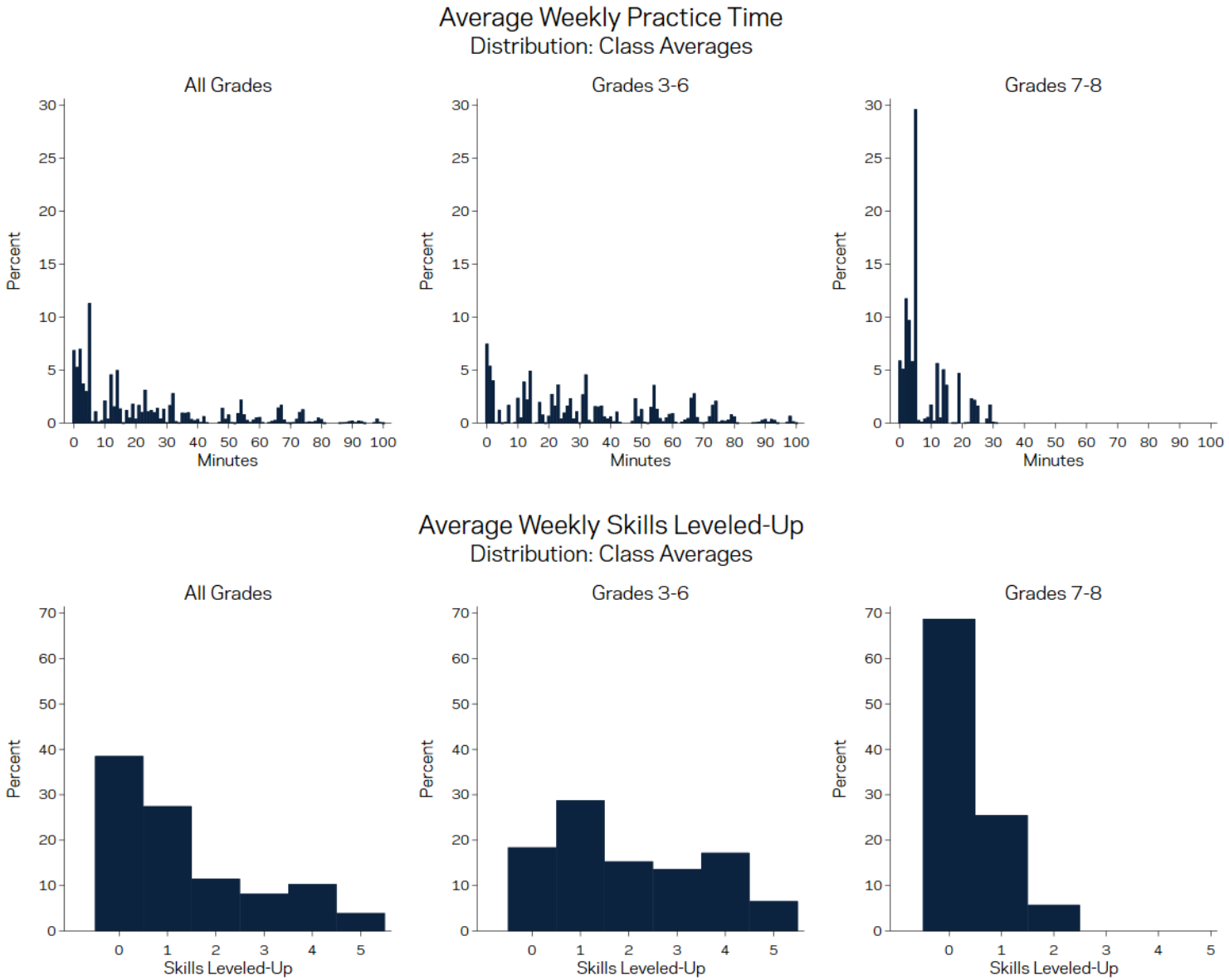
- At the end of the 2020-21 school year, we contacted all grade 3-8 math teachers in Arlington Independent School District about interest in using a CAL program in their class the following year
- 312 teachers expressed interest. They were grouped into 180 school/grade units to limit spillover between co-teachers
- After randomization (stratified by grade), 160 control teachers were selected to receive the program during the 2022-23 school year and contact was ceased

AISD: Grade Divisions

From the start, there was evidence supporting treating grades 3-6 (elementary) and 7-8 (middle) differently:

- Elementary classrooms blocked for 70 minutes per day for math, but teachers had flexibility to adjust since students remained in the same classroom for multiple subjects.
- Elementary classrooms had additional enrichment time throughout the week that teachers could use for any subject
- Middle school classrooms had 50 minutes per day with their math students
- Grades 3-6 average students per teacher: 43.7; Grades 7-8: 88.3
- Many teachers expressed disinterest in assigning Khan Academy as additional homework, especially given lack of tech access in some homes

AISD: Participation Distribution by Class



Control Group

AISD: Model

$$Y_{igs} = \beta_0 + \beta_1 T_{igs} + \beta_2 X_{igs} + \gamma_g + \epsilon_{igs}$$

- Y_{igs} is the standardized Math STAAR 2022 score for student i in grade g and school s
- T_{igs} is the treatment status of student i in grade g and school s
- X_{igs} is a matrix of person level controls for student i in grade g and school s
- γ_g is a grade level fixed effect, and ϵ_{igs} is the error term

AISD: Main Results

	I No Controls	II Grade FEs	III Grade FEs w/ Controls
All Grades	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 clusted at the grade/school level. Controls for III include: age, sex, race, ethnicity, days missed, english learner status, special ed status, free lunch eligibility.

By individual grade

TOT Effects

Arlington Experiment, Treatment on Treated Table

	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 Student Practice Time ≥ 20	0.321**	0.231**	0.388***	0.213*
	(0.129)	(0.111)	(0.143)	(0.115)
N	7,234	7,234	10,979	10,979
Average Student Practice Time ≥ 25	0.360**	0.259**	0.445***	0.248*
	(0.144)	(0.123)	(0.159)	(0.127)
N	7,234	7,234	10,979	10,979
Average Student Practice Time ≥ 30	0.402**	0.289**	0.504***	0.285**
	(0.160)	(0.136)	(0.178)	(0.141)
N	7,234	7,234	10,979	10,979
Average Student Practice Time ≥ 35	0.456**	0.329**	0.581***	0.333**
	(0.180)	(0.154)	(0.201)	(0.159)
N	7,234	7,234	10,979	10,979

Takeaways from Arlington Experiment

- Full sample program effect on student math scores insignificant
- This overall result masks distinct differences in estimates between elementary and middle schools, corresponding to distinct differences in practice time
- In elementary schools where average practice time was more than 30 minutes each week, state math scores were about 20% standard deviations higher for treated classrooms, but in middle schools where average practice time was less than 10 minutes each week, tests scores among treated students were significantly lower
- We posit that, given the little time middle school teachers have to teach math in Arlington, and their avoidance of homework, there is insufficient time for students to adopt mastery approach when practicing on KA, and this results in a worse substitute than the current curriculum
- Same conclusion as Nashville: KWiK fidelity is critical: teachers need to facilitate sufficient practice time for students to master material (try again until levelling up)

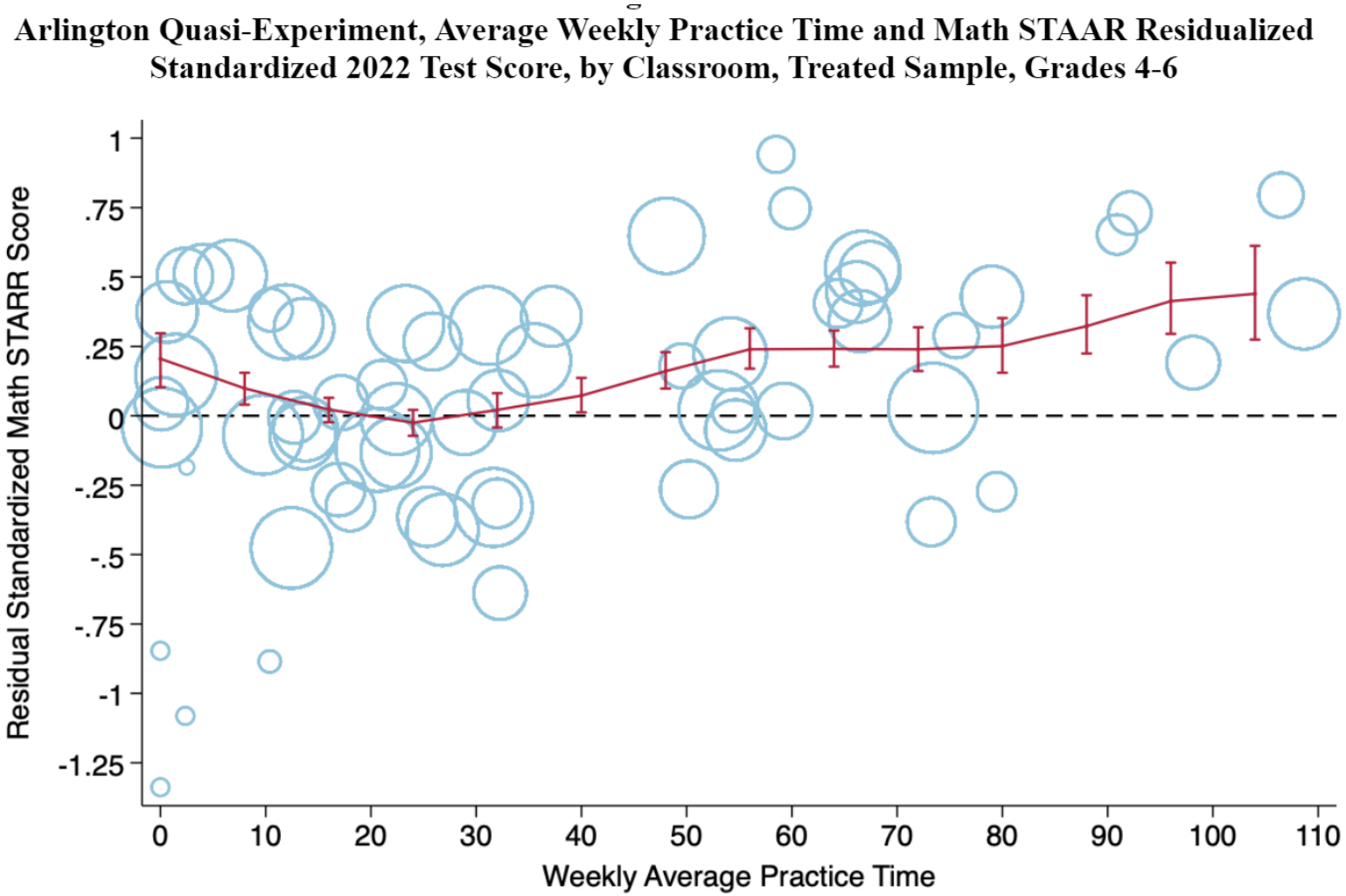
Caveats from Arlington Experiment

- Caveat 1: Breakdown of results by elementary vs middle school not in original pre-analysis
- Caveat 2: Estimates are imprecise, with p-values around 0.05
- Caveat 3: Even among elementary schools, high variance in fidelity

Quasi-Experimental Analysis

- Perhaps students 'as good as randomly assigned' to teachers within schools and grades, at least conditional on past test scores
- Similar to teacher value added models, except we examine the relationship between students assigned to teachers that facilitated more practice time on KA than others
- Past research suggests teacher value added credibly identified after conditioning on past scores (estimates similar to cases with student RA) (Chetty et al. 2014, 2017), Koedel et al. 2015, Kane et al. 2013)
- Caveat: are good teachers same ones who facilitate more KA practice?

Class Effects by Avg. Class Practice TIme



AISD: OLS relationship, both level-ups & practice minutes

	I	II	III
Full Sample			
1-2 Level-ups	0.356***	0.413***	0.426***
	(0.076)	(0.077)	(0.083)
2-5 Level-ups	0.656***	0.681***	0.746***
	(0.086)	(0.072)	(0.082)
5+ Level-ups	1.296***	1.316***	1.392***
	(0.139)	(0.125)	(0.139)
5-25 minutes	-0.239***	-0.076	0.022
	(0.051)	(0.064)	(0.090)
25-50 minutes	-0.477***	-0.368***	-0.101
	(0.075)	(0.075)	(0.081)
50+ minutes	-0.628***	-0.534***	-0.195*
	(0.110)	(0.099)	(0.111)
N	7,915	7,915	7,915

Notes: OLS regression of individual weekly practice on standardized 2022 Math STAAR scores. Does not include grade 3, as students do not test in grade 2. I: Controls for student demographics and 2021 STAAR Math Scores, Grade FEs. SEs clustered at the grade-school level. II: Adds School FEs. III: Adds Grade-School FEs.

Using Class Averages

Takeaways from Arlington Quasi-Experiment

- There exists a non-linear relationship in KA practice time among students with similar math backgrounds: only those averaging at least 30 minutes of practice each week or levelling up at least twice a week perform better by the end of the school year
- A similar relationship arises when looking at differences in performance by average classroom practice time and level ups
- Same conclusion as Nashville and Arlington Experiments: KWiK fidelity is critical: teachers need to facilitate sufficient practice time for students to master material (try again until levelling up)
- Caveat 1: assignment to classrooms may not be random
- Caveat 2: teachers who facilitate practice time could be better at in other ways

Qualitative Takeaways from Teacher Interviews

- High-practice teachers mentioned a strong sense of buy-in at the beginning of the program
- Deliberate plan for when to practice every week
- High-practice elementary school teachers often used candy, stickers, free time, leaderboards and sometimes grades as incentives
- Closely monitored student activity

Conclusions so far

- CAL offers potential to improve personalized support
- High variance around teacher facilitation of high dosage practice
- Potential for greater practice with dedicated practice time, close monitoring, high buy-in, and quality supervision and encouragement

Potential for LLM AI Tutoring

- LLM like CHatGPT can be prompted to provide high quality support
- Newest version allows voice to voice and video observation
- Potential to simulate ideal teacher to student setup
- Imagine mastery approach starting in early grades and progressing at own pace, with teacher supervision
- Remains to be seen how critical human touch is vs. high quality interaction

Next Steps, New Studies

- Chattnooga experiment with dedicated daily practice time and 'khanmigo'
- TDSB experiment with additional virtual 1:1 tutor
- A/B testing for increasing LLM tutoring engagement
- How to facilitate 1:1 model with technology and test its effectiveness exciting research agenda