THE UNIVERSITY OF CHICAGO

ESSAYS ON THE ECONOMICS OF EDUCATION

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE IRVING B. HARRIS
GRADUATE SCHOOL OF PUBLIC POLICY STUDIES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY

ROHEN SHAH

CHICAGO, ILLINOIS

JUNE 2025

For Ashvin Mama

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my dissertation committee for their constant support and invaluable feedback: Ariel Kalil, John List, Jens Ludwig, Lesley Turner. I am also very grateful for the team at the Behavioral Insights and Parenting Lab for the formative experience over the past seven years, especially my close collaborations with Susan Mayer and Michelle Michelini.

Equally helpful for my research has been my teaching experience at Chicago, and I am very grateful to have had a mentor and friend in Jim Leitzel. I also cannot thank Jen Lombardo enough for the many opportunities she gave me to teach, and learn about teaching, with several generations of Harris students. I could not have gotten through this without the countless conversations and camaraderie with fellow PhD students, including Emileigh, Kisoo, Lucas, Goya, Noah, Rubina, Mythili, Ari, and David McMillon.

I am so grateful for my family: Raj, Sima, Jessica, Rikita, Riyaan, Arya, Mark, Sandi, Jack, and Daniel. My mom's work ethic and selflessness have shaped me profoundly. Seeing my dad dream big helped me think outside of the box. Jessica was my first friend and role model; seeing her persevere through countless obstacles that life put in front of her has been a constant source of inspiration. Rikita was present. (Just kidding! You are a creative and generous person who has enriched my life in countless ways, and I am so much better for having had you as a sister).

Finally, this entire dissertation would not have been possible without my wife, Rebecca Shah. I am grateful for the many ways you have shaped me into the person I am today. I am so thankful for the countless sleepless nights where you helped recruit schools and proofread every paragraph I wrote, all while being an

incredible mother to our daughter, Reina. Words cannot capture your impact on every project I do, in research and in life.

ABSTRACT

Educational interventions are often shown to be effective in lab or pilot RCTs, but then subsequently fail to retain their treatment impact when applied at scale. This dissertation consists of three field experiments, each evaluating the impact of an educational intervention. The common thread in each of these is that the interventions I evaluate have at least one element that makes them more feasible to scale relative to similar interventions.

In the first chapter, I conduct a large-scale field experiment on learning by teaching. While previous interventions show evidence of "learning by teaching" in lab settings, this study tests the impact in a field setting over an 8-week period. Classrooms are randomly assigned to have students (1) create explanation videos, (2) complete passive practice problems, or (3) placed in a control condition. The explanation treatment improved short-run scores by 0.17 SD and long-run grades by 0.07 SD relative to the practice-problem group. Notably, while both treatment groups improved relative to control, only the explanation treatment improved performance on novel problems, suggesting that explaining concepts enhances one's ability to understand deeply and generalize concepts.

In the second chapter, I evaluate the impact of an in-school tutoring program. While schools aim to have both "high dosage" and "small groups", budget constraints make it infeasible to deliver small group tutoring frequently at scale. In this paper, I test the relative importance of group size (quality) versus tutoring frequency (quantity). Students at a middle school were randomly assigned to either 1) a control condition, or to receive in-school math tutoring 2) twice a week in 2-student groups, or 3) three times a week in 3-student groups. Importantly,

the total cost per student is the same in both treatment conditions. I find that the 2-student group tutoring led to a significant improvement in math skills (0.23 SD), whereas the equal-cost, more frequent tutoring in the 3-student groups did not lead to a significant improvement in math skills.

In the third chapter, co-authored with Ariel Kalil and Susan Mayer, we evaluate an intervention to increase attendance at preschool parent engagement events. We designed an intervention using a combination of financial incentives and two tools from behavioral economics: loss-framing and reminder messages. The treatment parents were given a loss-framed $25 per event incentive to attend 8 events sponsored by their preschools, as well as weekly text message reminders about the events. Relative to other similar RCTs, our smaller financial incentive is more feasible to implement at scale. We find no extensive margin treatment effect: the intervention did not increase the fraction of parents who attended at least one event. However, we find a 32% intensive margin treatment effect. This tells us that while behavioral tools can help already-involved parents engage more with preschools, they are not enough to reach disengaged parents. This study was recently accepted for publication in *Applied Economics*.

CHAPTER 1

WHEN THE STUDENT BECOMES THE MASTER

1.1 Introduction

Suppose that Jenny is a typical student who puts a modest amount of effort in school. Her friend, Jack, asks Jenny for help with the most recent homework assignment. As Jenny attempts to help, she realizes that she did not understand the material as well as she thought. Jenny engages more with the material, and through the process of generating explanations for Jack, Jenny comes to understand the concepts more deeply.

As modern economies increasingly rely on human capital for growth, economists have sought ways to develop human capital efficiently (Fryer, 2017). While economists have widely documented "Learning by Doing" (Arrow, 1962) in the context of firm production, it is not clear how this can be leveraged to improve human capital development for students. In this study, I propose that *teaching* is the "doing" that can lead to deep and efficient student learning. The problem is that the Jennys of the world typically do not have a friend asking them for help for every homework assignment. Is there a scalable way to provide students like Jenny more opportunities to engage in "doing" math?

In this paper, I test the impact of creating weekly math explanation videos on students' math skills. I conduct a field experiment in partnership with 20 public middle and high schools in the Midwestern United States during the spring 2024 semester. Math teachers at these schools who taught multiple periods (sections) of the same math class were invited to enroll their classrooms in the study. For

each teacher, their class periods were randomly assigned to 1) Control, 2) "Passive Task", and 3) "Video Creation" conditions. Upon enrollment, all three class periods were given a baseline math assessment and survey, and then a post-test after three months. The control classroom had business-as-usual homework assignments between the baseline and post-test. The "Passive Task" classroom was assigned one weekly PSAT[1] math question (hereafter, "Task") on Google Form, in addition to their regular homework assignments, for eight weeks. The "Video Creation" classroom was assigned the exact same weekly PSAT math question on the same schedule as the "Passive Task" classroom, with a sole difference: the Video Creation classroom's tasks asked them to create a video explaining their solution to the task to a hypothetical student. Students submitted their videos directly to their teachers via Flipgrid, Google Drive, Google Form, etc. based on the teacher's preference. This study enrolled 128 classrooms, for a total of 2,523 students.

The post-test at the end of the intervention contained 15 PSAT questions on topics that were relevant to the intervention "Tasks". A subset of these (6 questions) were *exact* Task questions that students in the Passive Task and Video Creation groups were assigned (hereafter, "Task Questions"). The rest (9 questions) were PSAT questions not previously assigned, but which still assessed the topics covered by the intervention Tasks (hereafter, "Novel Questions").

I find that the overall average completion rate across all students and tasks in the Passive Task classrooms was 84%, whereas that of the Video Creation classrooms was 62%. This difference might be expected given that the effort required to complete a task is higher for the Video Creation classrooms than the Passive

1. in some cases, questions were taken from the ACT, SAT, or middle school state tests depending on the grade level and topic covered in that teacher's class

Task classrooms. I estimate an Intent-to-Treat (ITT) effect of the treatment on overall math class grades and the post-test score, including a breakdown of performance on the "Task Questions" and "Novel Questions" subsets of the post-test. I find that relative to the Passive Task treatment, the Video Creation treatment led to a statistically significant improvement in both overall math class grade (0.07σ) and post-test score (0.22σ). There was no significant difference in either outcome between the Passive Task and the control groups. Further, when considering the "Task Questions" and "Novel Questions" subsets of the post-test, I find that the Video Creation group outperforms the control and Passive Task groups for both types of questions, whereas the Passive Task group only outperforms the control group for "Task Questions" but not for "Novel Questions." This implies that the Video Creation treatment was particularly effective at helping students generalize what they learned.

One potential mechanism driving the overall treatment effect might be the total amount of effort that a student would choose to spend on math as a result of being assigned the Video Creation task. On the one hand, the video creation might sufficiently motivate students to put in additional study effort beyond the time it takes to record themselves solving the problem. On the other hand, students might have an inelastic total amount of time that they would spend on math (Cotton et al., 2020). In this case, the time it would take to record themselves solving the math problem would take away from the time they spend preparing. To disentangle these potential mechanisms, I estimate a proxy for student effort for each task. Below the text of the Task problems, I provided a link to a "help" video on YouTube for each task that explains the solution to a similar problem. While the help video for each task was identical for the Passive Task and Video

Creation classrooms, I could track the total number of clicks on the help video separately for each classroom. I estimate the probability of clicking on a help video for any given task to be 11% for the Passive Task classrooms, and 25% for the Video Creation classrooms. This difference was statistically significant, and highlights that asking students to create videos could be way to induce student effort.

Unlike many cluster RCTs in education that randomly assign a treatment at the teacher or school level, the design of this experiment allows for the estimation of a "teacher-level" treatment effect. This is because the random assignment of classrooms was stratified by teacher, implying that every teacher has both a "Passive Task" and "Video Creation" classroom. This yields a distribution of treatment effects, and I find that the treatment effect of Video Creation relative to the Passive Task was positive for approximately 80% of teachers, and negative for 20%.

It is not obvious that the Video Creation task should lead to a positive impact on math skills relative to the Passive Task because of the lower likelihood of students completing the Video Creation task. If the completion rate for the Passive Task is sufficiently higher than the Video Creation task, we might expect a negative Intent-to-Treat effect on math skills. For instance, if most students in a teacher's Passive Task classroom complete their tasks, but virtually no students in that teacher's Video Creation classroom complete their tasks, then we would expect students in the Passive Task classroom to have a greater improvement in their math skills than those in the Video Creation classroom. I leverage the aforementioned distribution of "teacher-level" treatment effects to explore the relationship between relative task completion and treatment effect. I measure the "relative compliance rate" as the task completion rate in a teacher's Video Creation

classroom minus that of their Passive Task classroom. I find a statistically significant, positive correlation between the relative compliance rate and treatment effect. This suggests that the effectiveness of the Video Creation treatment depends on a teacher's ability to induce their students to create videos. It might also suggest that for the sake of maximizing student welfare, teachers who are unable to induce students to create videos might be better off switching to the passive task.

This work relates to recent laboratory experiments in behavioral economics that show that people are substantially more likely to learn things they discover on their own as opposed to hearing from others (Conlon et al., 2022), as well as work showing that students benefited from being asked to "give advice" via a short survey (Eskreis-Winkler et al., 2019). This also contributes to our understanding of how humans "transfer" knowledge to new situations. The cognitive science literature denotes different stages of "generalizing" knowledge, where the initial step is to recognize a "new problem" as one similar to a familiar scenario, followed by creating a mental map between the familiar and new scenario, and finally using that map to solve the new problem (Samat et al., 2019). A part of what might hinder this knowledge transfer is the ability to retrieve memory of the familiar scenario (Bordalo et al., 2020, 2021). Another hindrance might be not understanding the structure of the original problem well enough to be able to create the "map" between the old and new scenario (Ahn et al., 1992). In this experiment, I showed that students who did the passive task were able to retrieve the memory of those tasks because they outperformed the control group, but only the students who created videos were able to outperform the control group on "new scenario" questions. This might suggest that memory retrieval is not the issue, but rather the

video creation process helps students understand the structure of the problem well enough to be able to make the requisite "mental map" relating their knowledge of the topic to the "new problem," thereby generalizing what they had learned.

This also contributes to the literature on scaling in economics. Many interventions that have large positive impacts in the lab often fail to replicate when brought to the field (List, 2022). While lab studies indicate the potential for improving outcomes through "learning by teaching", this study presents a way to successfully implement this idea in a large-scale field setting to improve non-lab outcomes. Similarly, qualitative and lab studies in education research highlight the importance of engaging students in "active learning" to improve outcomes (Bonwell and Eison, 1991; Freeman et al., 2014; Markant et al., 2016). However, there has been mixed success with implementing active learning strategies in classroom settings, sometimes resulting in a negative effect relative to the status quo (Berlinski and Busso, 2017). Given the challenge in asking teachers to change the way they teach, this study shows that a lighter-touch intervention that simply adds a few assignments might be a more reliable way to engage students in active learning such that the effect holds at scale.

This also relates to the work by labor economists on motivating students. Provision of high-quality resources alone does not guarantee utilization by students who would benefit (Robinson et al., 2022). It is even difficult to motivate students when providing financial incentives (Brownback and Sadoff, 2020; Burgess et al., 2021; Fryer, 2017; Sadoff, 2014). One reason why financial incentives on "outputs" such as test scores might not be effective is because students might not know how to convert inputs (i.e. studying) into outputs (Cotton et al., 2020). Nevertheless, student motivation to put forth effort can also play a role in test performance

(Gneezy et al., 2019). In general, incentivizing inputs is more reliable than incentivizing outputs (Gneezy and List, 2013). This study contributes to this literature by highlighting a way to incentivize student effort without the use of financial incentives. Randomly assigning students to create videos doubled their likelihood of using the "help video" resource, which indicates a substantial difference in student effort. This treatment might therefore be a more scalable way to induce student effort as it does not require schools to budget for prizes or financial incentives for students.

Finally, in a world where Generative Artificial Intelligence (GenAI) such as Large Language Models (LLM) like ChatGPT are more commonplace, a common concern is that students can now easily put in less cognitive effort in tasks completed at home, where LLM usage cannot be monitored. This study shows that video creation at home is one type of homework assignment that could mitigate this concern, given the finding that students are more likely to put in effort in this task than the passive-task counterpart. Another contribution of this study is to highlight the potential downside in plans to scale up tutoring provision by replacing human peer tutors with LLMs. Doing so would lead to a loss in potential human capital gains by the would-be peer tutors, and quantifying this is necessary for a full cost-benefit accounting of such plans.

This paper is organized as follows. Section 2 discusses the prior work on learning by teaching and the rationale for this intervention. Section 3 presents a theoretical framework of student effort choice. Section 4 describes the experimental design, and Section 5 shows the results. I discuss mechanisms and the distribution of treatment effects in Sections 6 and 7 respectively, and conclude in section 8.

1.2 Background

1.2.1 Prior Evidence for Learning by Teaching

Conceptually, teaching may lead to more learning than traditional techniques because teaching may be more engaging in two different senses. First, teaching may lead people to spend more time and effort with academic content. This might happen if the teacher altruistically cares about the learner's outcome, or if they socially care about the learner's perception of the teacher's abilities. Second, teaching may force the teacher to engage more actively with the content, including thinking about the content in different ways as part of thinking through how to communicate that idea to someone else.

Evidence for "learning by teaching" primarily comes from lab settings, but the extent to which the treatment effect generalizes to the field is unclear. Lab studies have shown that students score better on a quiz when they are preparing to tutor someone, as opposed to preparing for a quiz (Fiorella and Mayer, 2013; Guerrero and Wiley, 2021). Students are randomly assigned to either a control condition or a "tutoring expectation" condition. Control participants are told that they have 10 minutes to study for a quiz on a physics topic (the Doppler Effect) and are provided with study materials. The treatment participants are told that they have 10 minutes to prepare to tutor someone on the Doppler Effect, and are given the same study materials that the control group is provided. However, the treatment group is then given a quiz rather than actually tutoring someone (they are debriefed about this deception afterward), and the authors find a significant, positive effect on the quiz score. This sheds light on a mechanism through which tutors might learn from tutoring: they prepare more deeply before their tutoring

sessions. While this evidence is suggestive that students might learn by tutoring, it is not clear whether the results from this lab setting would generalize to the field. In particular, the impact measured in the lab is from a single preparation session, and the effect might not be sustained over an entire semester. Additionally, if students in a field setting choose not to review content before tutoring, then the results of the study would not apply.

There have not been many field experimental studies that have identified the effect of tutoring on the tutor's knowledge. Those that do often suffer from substantial identification threats. For instance, some studies on peer tutoring measure what students learn after both receiving and providing tutoring to a peer (De Backer et al., 2012; King et al., 1998). These designs measure the improvement in skills over time (pre-post design), but do not have a "control group" of students that do not engage in the activity. 5th grade students in Greene et al. (2018) were assigned to tutor 3rd and 4th graders. The 5th grade students were randomly assigned to either receive training or not receive training prior to this tutoring. However, there was no random assignment to a control group (5th graders who did not tutor), and therefore the study is not designed to identify the effect of tutoring on the tutor's knowledge. AbdulRaheem et al. (2017) randomly assigns a single classroom from one school to control and a single classroom from another school to a peer-tutoring treatment. This design makes it impossible to disentangle the treatment effect of peer tutoring from the teacher-, classroom-, or school-fixed effects.

Mitchell et al. (2016) does randomly assign 4th grade students to either a control condition or a condition where they tutor a 2nd grade student (with or without training). The teacher decided the student-tutor pairs based on personality

matches, which in and of itself is not a threat to internal validity. The researchers found that while the 2nd grade students benefitted from receiving tutoring, the 4th grade students did not benefit from providing tutoring. The biggest issue with this study is the sample size. With less than 15 students per treatment arm (43 4th grade students for three experimental conditions), the study is underpowered to detect modest treatment effects.

Two studies attempt to identify the impact of tutoring on learning by randomizing which subject a student tutors. Romero et al. (2022) studied cross-age tutoring within a primary school in Kenya. The sample size is large, and the outcome is the tutor's own grades. They find that the tutors had little impact on their own test scores from tutoring math (as opposed to tutoring English). Noteworthy here is that there is a 5-year gap between the tutor and student. This might be too large of a gap to expect the tutor's own math performance (5th grade) to improve from tutoring the student (1st grade), especially because the tutors are self-selected high performers. Fuchs and Malone (2021) assigned master's students in education to tutor either Math ($n = 25$) or English ($n = 17$) to elementary school students. However, the assignment was based on scheduling constraints rather than random assignment. They find large effects (0.5 to 0.75 SD) for an assessment on the math topic that was tutored (fractions).

1.2.2 Rationale for Video Creation as an Intervention

Given the effectiveness of learning by teaching in lab settings, we might wonder why this is not already universally used as a tool for human capital development in schools. While there have been calls for universal school-wide peer tutoring

programs, few exist (Kraft and Falken, 2021).

One reason might be that only the type of person who chooses to teach would benefit, or what economists call "selection on gains." However, it might be the case that the benefits from teaching are universal, yet only a select few engage in the activity. For instance, it might be the case that while everyone would benefit from providing peer tutoring, only students near the top of the achievement distribution are asked to provide peer tutoring. If this is the case, then making the opportunity to provide peer tutoring universal would help mitigate the achievement gap.

There are at least three barriers that could explain why peer tutoring programs are not widespread. 1) It is logistically difficult to coordinate peer tutoring sessions. Scheduling synchronous sessions in a way where most students have an opportunity to provide tutoring takes considerable effort, and this high transaction cost may not be worth the perceived benefit. 2) Programs designed to give peer tutors the opportunity to learn might not be ideal for the students receiving tutoring. The tutoring programs with the highest impact on students are ones where screened professionals provide the tutoring (Nickow et al., 2020). This implies that the opportunity cost for students receiving tutoring from an untrained peer might be too high to justify their participation. 3) Students might be unwilling to engage in an activity that reveals their academic ability to their peers. Prior work has shown that making academic effort publicly visible to students' peers can have an adverse impact if students highly value their social status (Bursztyn et al., 2019).

An intervention that overcomes these concerns is one where students create video explanations to a hypothetical peer as homework assignments. 1) The asynchronous nature of this task mitigates the need to coordinate schedules, making participation more feasible. 2) The risk that the recipient of tutoring might re-

ceive a lower quality experience than alternative uses of their time is no longer a concern. 3) Adverse peer effects are avoided because the audience is the teacher rather than peers. This intervention might not capture all of the benefits that synchronous peer tutoring might entail. For instance, if the back-and-forth interaction in peer tutoring leads to a substantial impact on the tutor, then this video creation intervention misses out on this benefit by only focusing on the "initial explanation" phase (Kobayashi, 2022). Additionally, if the tutor's motivation to put effort into their tutoring stems from caring about the student's outcome, then this intervention would not capture this benefit because the audience of the video is a hypothetical rather than an actual peer. Despite this lower potential benefit, the video creation intervention might still have a high benefit-cost ratio because of its low cost and potential for scalability.

1.3 Theoretical Framework

This section describes a model of "learning by teaching" that highlights mechanisms as well as potentially adverse affects. I also describe a model of peer tutoring, highlighting the tradeoffs from introducing AI tutoring, in part because of the learning loss from the tutors themselves if they are replaced with AI tutors.

1.3.1 A Model of Student Effort Choice

I adapt the worker effort framework in DellaVigna et al. (2022) to describe student i's optimal effort as a function of grade incentives, the cost of effort, and

a social preference parameter A. The utility function to maximize is:

$$u(e_i) = p \cdot e_i - c_i(e_i) + A \cdot e_i$$

Where e_i is the amount of effort, the piece rate $p > 0$ is the incentive in the form of class grades, $c_i(\cdot)$ is the cost of effort for student i, and A is a social parameter that reflects how much you care about others (i.e. altruism) or their perception of you (i.e. social image). I assume that effort maps one-to-one to observable grade outcomes. A unique solution is guaranteed by assuming $c'() > 0$, $c''() > 0$, and $\lim_{e\to\infty} c'(e) = \infty$. The first order condition to this maximization problem is:

$$\begin{aligned} 0 &= p - c'(e_i) + A \\ c'(e_i) &= p + A \\ e_i &= (c^{-1})'(p + A) \end{aligned}$$

Now, suppose there are two types of effort that a student can choose to engage in. I denote e_1 as *receptive effort* and e_2 as *generative effort*, which signifies the type of learning the student is engaged in (passive versus active). Generative effort is effort that results in generating an output, such as an explanation, as opposed to receptive effort, which might involve reading an explanation. The piece-rates associated with receptive and generative effort are p_1 and p_2 respectively, where $p_2 > p_1$. The cost functions associated with receptive effort and generative effort are $c_1(\cdot)$ and $c_2(\cdot)$ respectively, where $c_2(e) > c_1(e),\ \forall e$. Students are aware that generative effort has a higher benefit, but also that generative effort has a higher cost. The cost functions for these efforts can vary for each student, which can

result in different levels of investment of each type of effort among students that face the same grade incentives. For now, I assume that the costs and benefits of each type of effort type are independent from each other.

Suppose that a teacher is considering two alternative assignments where the student's work will be visible to the teacher. One assignment ($T = 0$) is a direct function of receptive effort e_1, and the other assignment ($T = 1$) is a direct function of generative effort e_2. This results in the following utility maximization problem:

$$\max_{e_1,e_2\geq 0} u(e_1, e_2) = p_1e_1 + p_2e_2 - c_1(e_1) - c_2(e_2) + A(e_1 \cdot \mathbb{1}_{T=0} + e_2 \cdot \mathbb{1}_{T=1})$$

For a student who is assigned task $T = 0$, where receptive effort is socially incentivized, the optimal level of e_1 effort is $e_1^* = (c_1^{-1})'(p_1 + A)$. For a student who is assigned $T = 1$, where generative effort is socially incentivized, the optimal level of e_2 effort is $e_2^* = (c_2^{-1})'(p_2 + A)$. Note that if no socially visible task is assigned, then the optimal effort values are $e_1^* = (c_1^{-1})'(p_1)$ and $e_2^* = (c_2^{-1})'(p_2)$.

For students who are assigned $T = 0$, there is no change in the optimal value of e_2 relative to receiving no socially incentivized task. The change in optimal effort e_1 relative to no socially incentivized task is $\Delta e_1 = (c_1^{-1})'(p_1+A)-(c_1^{-1})'(p_1) > 0$.

For students who are assigned $T = 1$, there is no change in the optimal value of e_1 relative to receiving no socially incentivized task. The change in optimal effort e_2 relative to no socially incentivized task is $\Delta e_2 = (c_2^{-1})'(p_2+A)-(c_2^{-1})'(p_2) > 0$.

In other words, relative to being assigned $T = 0$, a student being assigned $T = 1$ results in them exerting a higher level of e_2 effort, but a lower level of e_1 effort. This implies that a treatment that socially incentivizes generative effort relative to a counterfactual of incentivizing receptive effort could result in a *negative treatment*

effect if Δe_2 is sufficiently smaller than Δe_1. That is, a student might be better off being assigned a receptive task rather than a generative one if the receptive task results in a substantial increase in receptive effort, whereas the generative task results in a small increase in generative effort. This would depend both on the student's relative costs of generative and receptive effort, and also on the mapping between each type of effort and math skills (or utility) outcomes.

1.3.2 Effort Complementarity

The previous section assumes additive separability in the cost and benefit of the two effort types. Dropping this assumption implies that the two components of effort could be either substitutes or complements in production. Students might view the total amount of time spent on math as relatively inelastic (Cotton et al., 2020). If that is the case, then the two types of effort might be treated as substitutes. This would imply that being assigned a generative task such as creating videos might make you invest less effort in receptive tasks such as watching videos. On the other hand, a student may view these two types of effort as complementary. In this case, being assigned a task requiring generative effort would make them more likely to also engage in receptive effort. That is, being assigned to create a video would make them *more* likely to watch a math video as well.

1.3.3 Skill Transfer

Suppose the skill variable y has two components, y_1 and y_2 corresponding to *local* and *generalizable* skills respectively. y_1 refers to knowing how to do something exactly a certain way, whereas y_2 refers to one's ability to generalize how to do

tasks similar to ones they know. The work in cognitive psychology implies that active learning effort (e_2) is a pre-requisite to develop generalizable (y_2) skills (Ahn et al., 1992; Boaler et al., 2022). This would imply that $\frac{\partial y_2}{\partial e_1} = 0$. This implies that the likelihood of the aforementioned negative treatment effect of being assigned to teach someone is much lower for generalizable skills, and also that the counterfactual passive task treatment would not have a positive treatment effect on the y_2 component of skill.

1.3.4 A Model of Peer Tutoring

Suppose a school has a population of N students, and of these τ have the ability to serve as peer tutors. Each tutor has the capacity to tutor g students. Suppose that $\tau + g\tau < N$, implying some students do not receive tutoring in equilibrium.

There is a distribution of "benefits from tutoring" amongst the $N - \tau$ students who are not peer tutors, and the way to select the $g\tau$ students who receive tutoring is simply by selecting those with the highest benefit.

Now, suppose that *AI Tutoring* is introduced, which allows for all remaining $(N-\tau-g\tau)$ students to receive tutoring. Assume this benefit is quantifiable. While this benefit may seem to come at no cost, there might be a substitution where some human tutoring amongst the $g\tau$ students gets switched to AI tutoring. Even if the quality of AI tutoring is the same from the students' perspective, the *tutors would lose out* on skill development. This negative effect should be factored in when estimating the total impact of AI tutoring.

Additionally, suppose that γ percent of the students who don't otherwise choose to get tutoring are "cheaters", where they would use the AI tutoring as a way to

get answers to homework questions and reduce their cognitive load of studying at home as a result of having access to this AI tutoring. This decrease in their own learning would impact performance on in-school administered tests, and thus be quantifiable.

This implies that the introduction of AI tutoring would need to weigh the benefit of the students who now get tutoring that previously could not with the costs of 1) decreased learning opportunities for the peer tutors themselves, and 2) decreased skills for those who would now lessen their own at-home effort as a result of having access to AI tutoring.

This project conducts a "learning by teaching" intervention in a way that provides an estimate for cost (1) describe above, but also explores whether cost (2) could be mitigated by assigning an at-home task that is more difficult to "cheat" on, thus inducing more student effort relative to a passive task where AI could do all of the work.

1.4 Experimental Design

1.4.1 Sample and Recruitment

District and School Recruitment

During December 2023 and January 2024, I recruited school districts via emails to the superintendents. Emails were sent to all districts with at least 5,000 students in Illinois, Wisconsin, Iowa, Michigan, and Ohio. I sent 371 emails and received 94 replies, of which 29 indicated interest in learning more. Of these, 13 agreed to participate and shared information about the study with middle and high school

math teachers. The most common reasons for districts not being able to participate were scheduling conflicts with other school improvement initiatives and extensive research review processes that would not fit in our timeline for the current school year.

The characteristics of participating school districts varied widely. The average percentage of students on Free or Reduced Lunch was 41%, and ranged from 8% to 93%. The high school graduation rates varied from 72% to 98%, with an average of 88%.

Teacher Recruitment

Next, math teachers at participating school districts were sent an interest form with information about the study. Teachers who taught at least two periods (sections) of the same math class were eligible. Teachers received $500 for their effort in helping implement the study. Recruitment took place on a rolling basis between February 5th and March 8th, 2024. In total, 47 teachers filled out the interest form, of which 41 teachers chose to proceed with the study after learning more. Of these 41, 28 were high school teachers and 13 were middle school teachers.

Finally, teachers distributed parent permission forms to all students in their classes. The permission form indicated whether the students were allowed to take a short survey with demographic information and their math background, as well as whether the teacher could share identifiable data with the researchers (including student-generated videos). Students whose parents did not give permission still participated in the tasks according to their class period, but they did not take the survey and only de-identified data was shared for these students.

1.4.2 Randomization

Random assignment was done at the class period level, stratified by teacher. Individual-level randomization is difficult to implement because the treatment involves assignments given at the classroom level. SUTVA violations (List, 2024a) might especially be a concern for individual-level randomization in this context.

Teachers were eligible to enroll their classes in the study if they taught at least two periods (sections) of the same math class. If they taught exactly two periods, then one was randomized to the "Video Creation" treatment and the other to the "Passive Task" treatment (see Section 1.4.3 for a description of these treatments). If they taught three or more periods, then they were randomized to 1) Video Creation, 2) Passive Task, and 3) Control as long as each period had at least 15 students. If there were fewer than 15 students in a period, then the smaller two periods were treated as a single unit for the purpose of randomization (See the Pre-Analysis Plan[2] for the full description of the relative class size cutoff algorithm used to determine random assignment). This was done to ensure that there was sufficient power to detect differences between the Passive Task and Video Creation treatments, with the pure control classrooms being an additional comparison arm in case enough teachers with more than two periods were recruited.

Randomization occurred after all students in a teacher's class periods took a pre-test consisting of consisting of 15 multiple-choice questions taken from either the PSAT, ACT, SAT, or grade-level state standardized math test. The questions were chosen based on the topics the teachers indicated they planned to cover between March and May 2024. This in-class pre-test was 25 minutes, and teachers

2. `https://www.socialscienceregistry.org/trials/11884`

administered these on a rolling basis between March 4th and March 29th. Randomization occurred on a rolling basis after all the pre-tests were completed for each teacher.

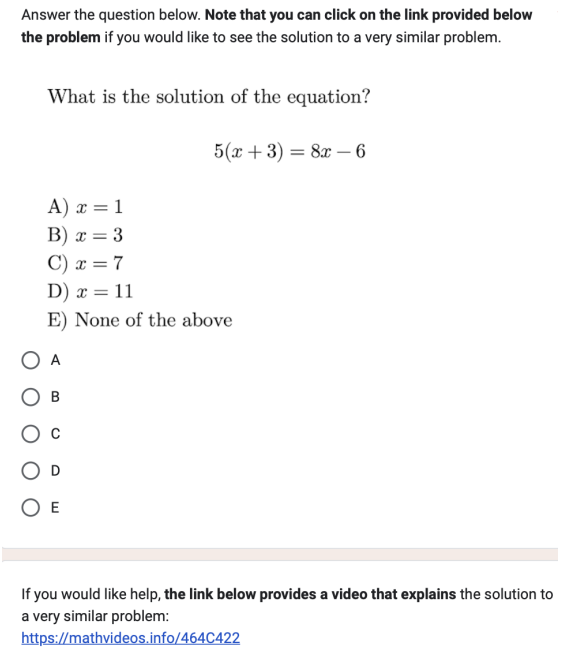
1.4.3 Treatment Description

Each week, teachers assigned students in their Passive Task classroom a PSAT (or in some cases ACT, SAT, or grade-level state test) question to complete for homework via Google Form. Teachers had some discretion on whether to skip weeks or whether to have two tasks in the same week depending on their preference and alignment with their curriculum. Teachers gave input for the topics that they preferred that the tasks covered, and the researcher selected questions from a pool of aforementioned standardized test questions. Teachers also had discretion on how they incentivized the task, with the recommendation being that it be for completion credit. While most teachers followed this, some were unable to do so because of school-wide policies that disallowed the use of homework for credit.

Below the Task question on each weekly Google Form, students were provided with a link to a "help" video on YouTube that explained the solution to a similar problem. These videos were selected by the researcher. In some cases where no videos were easily available, a video was created for the purpose of this study. The videos were shared on the Google Form in a bitly[3] link, which allowed the researcher to measure the total number of clicks on that link. While it was not possible to measure whether an individual student clicked on the link, the bitly link was unique for each classroom for each task. As pre-registered, I use this

3. `https://bitly.com/`

information to determine the probability of clicking on the link for each student. An example of this is shown in Figure 1.1.


Answer the question below. **Note that you can click on the link provided below the problem** if you would like to see the solution to a very similar problem.

What is the solution of the equation?

$$5(x+3) = 8x - 6$$

A) $x = 1$
B) $x = 3$
C) $x = 7$
D) $x = 11$
E) None of the above

○ A
○ B
○ C
○ D
○ E

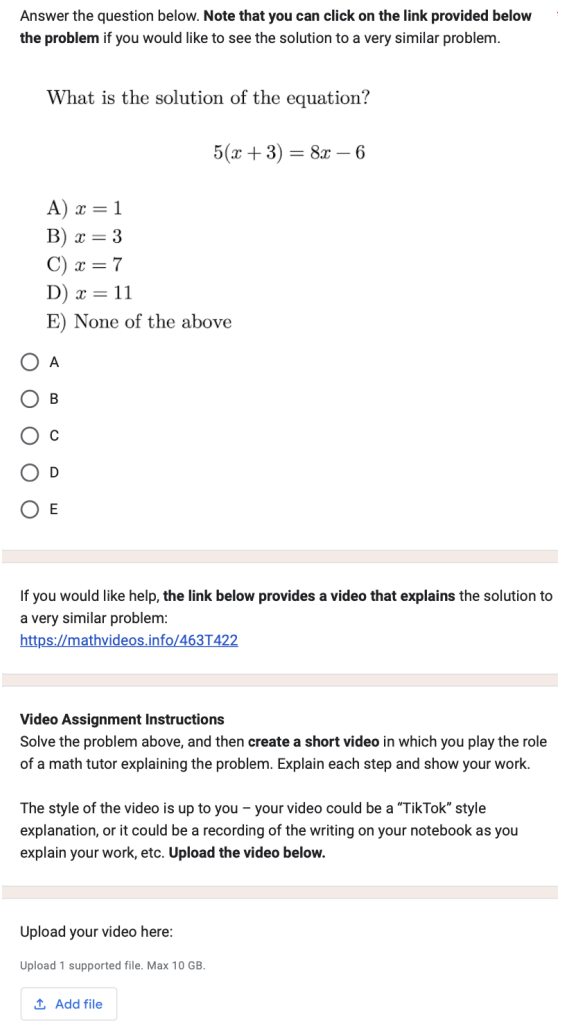
If you would like help, **the link below provides a video that explains** the solution to a very similar problem:
https://mathvideos.info/464C422

Figure 1.1: Example of Passive Task

In the Video Creation classrooms, the teachers assigned the same weekly task on the same schedule as their Passive Task classrooms. The only difference was that their Google Form asked them to submit a video explaining the solution. Teachers chose the exact mechanism by which students submitted their video depending on the teacher's preference. Most teachers used FlipGrid to collect student videos. Some teachers used Google Form directly, or asked students to send the video via email or uploading to a Google Drive folder. The Google Form for the video creation classrooms also had the same YouTube help video, but with a different bitly link that allowed for a separate identification of the number of total clicks in the Passive Task versus Video Creation classrooms. An example of this is shown

in Figure 1.2.

Three teachers assigned this as an in-class activity because their school did not allow for assigning homework. As pre-registered, these three teachers were dropped after it was determined that an exception could not be made for the purpose of this study.


Answer the question below. **Note that you can click on the link provided below the problem** if you would like to see the solution to a very similar problem.

What is the solution of the equation?

$$5(x+3) = 8x - 6$$

A) $x = 1$
B) $x = 3$
C) $x = 7$
D) $x = 11$
E) None of the above

○ A
○ B
○ C
○ D
○ E

If you would like help, **the link below provides a video that explains** the solution to a very similar problem:
https://mathvideos.info/463T422

Video Assignment Instructions
Solve the problem above, and then **create a short video** in which you play the role of a math tutor explaining the problem. Explain each step and show your work.

The style of the video is up to you – your video could be a "TikTok" style explanation, or it could be a recording of the writing on your notebook as you explain your work, etc. **Upload the video below.**

Upload your video here:

Upload 1 supported file. Max 10 GB.

Add file

Figure 1.2: Example of Video Creation Task

1.4.4 Outcomes

Two outcomes I use to estimate math skills are: 1) Semester 2 math class grade, and 2) A 15-question post-test, consisting primarily of PSAT and ACT questions, on the topics related to the Tasks for that teacher's classes. This 15-question test included 6 questions that were taken from the exact assigned Task questions, and 9 additional questions that were not exact task questions but covered the same topics.

The two secondary outcomes that shed light on mechanisms are 1) the students' self-reported math confidence, and 2) the students' likelihood of viewing the "help" video. The students' likelihood of viewing the "help" video is a proxy for the amount of effort they put into the task, and was measured as the number of total clicks on their class period's link divided by the total number of students in that period.

1.4.5 Covariates

The teachers administered the pre-test along with a survey that asked about students' age, gender, race, math confidence, growth mindset, as well as past experience with tutoring and video creation. Only students whose parents gave active consent and identifiable data-sharing permission took the survey.

1.5 Results

Sample Attrition

Overall, about half of the students in the sample were women, 10% were black, and 23% Hispanic. There are no significant differences in observable characteristics by treatment condition.

Some students did not take the post-test at the end of the year, primarily due to absences on the day the post-test was administered. While some teachers were able to administer make-up tests for students who were absent on the post-test day, others were constrained by final exam scheduling and chronic absenteeism from students who missed both days. Given that the post-test is the primary outcome variable, I consider a student to have dropped out of the study if they did not take the post-test.

Using this definition of "drop out", I measure the attrition rate as the fraction of students who dropped out. The attrition rate was 13.6%, 13.0%, and 13.9% for the control, Passive Task, and Video Creation groups respectively. There is no significant difference in these attrition rates across treatment conditions.

Compliance

Given the additional effort required to create a video (Video Creation treatment) relative to simply completing the google form (Passive Task treatment), we would expect the task completion rate for the Video Creation students to be lower than that of the Passive Task students. I define the task "completion rate" for each student as a continuous variable between 0 and 1 that indicates the proportion of

all 8 tasks that student completed[4]. I find that the overall task completion rate is 83.7% for the Passive Task students, and 62.4% for the Video Creation students. When subsetting on only the students who did not attrite from the study, these numbers are 87.8% for the Passive Task students and 66.5% for the Video Creation students.

I additionally pre-registerd a binary "compliance" variable for each student that has a value of 1 if the student completed at least half of the assigned tasks (i.e. if the student completed 4 or more tasks if their teacher assigned all 8 tasks), and 0 otherwise. With this definition, the compliance rate for the Passive Task students is 90.1% and that of the Video Creation students is 69.8%. When subsetting to the students who did not drop out of the study, the compliance rate is 94.7% for the Passive Task students, and 74.6% for the Video Creation students. While I use these compliance rates to estimate a Local Average Treatment Effect (LATE) in the appendix, the primary results I report in this study are Intent-to-Treat (ITT) estimates because they are more policy-relevant.

1.5.1 Impact on Math Skills

The following regression was used to estimate the treatment effect on math skills. This regression was pre-registered in the Pre-Analysis Plan[5]:

$$y_{i,t} = \beta_0 + \beta_1 T_i + \beta_2 C_i + \beta_3 y_{i,t-1} + \tau_j + \varepsilon_i$$

4. a few teachers had to drop tasks and had fewer than 8. In these instances, the completion rate is the total number of tasks that student completed divided by the total number assigned by the teacher.

5. `https://www.socialscienceregistry.org/trials/11884`

Here, $y_{i,t}$ is the value of outcome at the end of the experiment, $y_{i,t-1}$ is the baseline level of outcome, T_i is a binary indicator where 1 = either Video Creation or Passive Task classroom and 0 otherwise, C_i is a binary indicator where 1 = Video Creation task classroom and 0 otherwise, τ_j is the teacher fixed effect for teacher j, and ε_i is the error term. Standard errors of coefficients are clustered at the classroom level, because that was the level at which the treatment was assigned.

Math Grades

Table 1.1 shows the treatment effect on the overall class grade, normalized for each teacher and measured in standard deviation units. Class grades were measured either on a percentage scale (0-100) or on a 4-point scale. Class grades were unavailable for some teachers depending on the Data Use Agreement with that district. Additionally, a few middle school teachers from one district had no overall numerical course grades available due to district policy. With these restrictions, normalized course grades were available for 1,603 students. Overall, I find that the video creation treatment led to a statistically significant 0.068 Standard Deviation increase in math class grade relative to the Passive Task treatment.

Post-Test Scores

Table 1.2 shows the main treatment effect on the PSAT post-test. The first column shows the results for the full sample, indicating that the Passive Task did not lead to a significant improvement in the Post Test score overall, whereas the Video Creation treatment led to a significant impact compared to the Passive

Table 1.1: Treatment Effect on Math Grades

	(1) Math Class Grade
Assigned Either Task (Passive or Video Creation)	0.00288 (0.0553)
Video Creation	0.0680** (0.0315)
Baseline Grade Controlled	Yes
Teacher Fixed Effects	Yes
Observations	1,603

Note: The Assigned Either Task variable in the first row is binary with a value of 0 for students in the control group, and a value of 1 for students in either the Passive Task or Video Creation group. The Video Creation variable in the second row is binary with a value of 0 for students in either the control or Passive Task group, and a value of 1 for students in the Video Creation group. The outcome is the student's overall math course grade for semester 2, measured in standard deviation units. Clustered standard errors (at the classroom-level) are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Task.

For the first four teachers that participated, the post-test questions were chosen by the teachers. These four teachers graded the post-test themselves and sent final scores to the researcher, and a breakdown of how each student performed on each question is unavailable. The second column excludes these four teachers, showing the treatment effect for only the subset of students where the post-test grading was automated on Google Form (Quiz Mode) and student performance on each question is available.

The 15-question post-test included 6 questions that were *exact* intervention questions completed by the Passive Task and Video Creation classrooms (Task Qs), and 9 questions that covered the same topics, but had not been assigned as

tasks (Novel Qs). The treatment effect on the subset of Task and Novel post-test questions is shown in columns (3) and (4) respectively, using the same sample of students as in column (2). The third column shows that there was a significant impact of the Passive Task on the exact Task questions, and the fourth column shows that there was no significant impact of the Passive Task on the Novel questions. The Video Creation treatment has a significant impact relative to the Passive Task for both types of questions.

Table 1.2: Treatment Effect on Post-Test

	(1) Post-Test	(2) Post-Test
Assigned Either Task (Passive or Video Creation)	0.113*	0.124*
	(0.0581)	(0.0625)
Video Creation	0.216***	0.167***
	(0.0414)	(0.0429)
Baseline Score Controlled	Yes	Yes
Teacher Fixed Effects	Yes	Yes
Observations	2,130	1,869

Note: The Assigned Either Task variable in the first row is binary with a value of 0 for students in the control group, and a value of 1 for students in either the Passive Task or Video Creation group. The Video Creations variable in the second row is binary with a value of 0 for students in either the control or Passive Task group, and a value of 1 for students in the Video Creation group. The outcome in Columns 1 and 2 is the 15-question post-test consisting of standardized test questions that were relevant to the teacher's curriculum. Column 1 is the full sample, whereas Column 2 only contains teachers for whom a breakdown of the type of post-test question is available. All outcomes are in standard deviation units. Clustered standard errors (at the classroom-level) are in parentheses.
*** p<0.01, ** p<0.05, * p<0.1.

Knowledge Generalization

For instruction to be broadly useful, it is important for students to be able to generalize what they learn and apply it to other contexts, rather than learning in a way that is akin to rote memorization (Bonwell and Eison, 1991). Cognitive psychologists refer to this type of generalization as *knowledge transfer* (Samat et al., 2019). One aspect of this transfer process depends on rote memory of the initially learned concept, and another depends on understanding the concept well enough to be able to create a map between the previously learned concept and a new problem (Ahn et al., 1992).

Table 1.3 shows the impact of the Passive Task and Video Creation treatments. These regressions are the same as those in columns (2)-(4) of Table 1.2, except that Table 1.3 shows the treatment effect for each treatment relative to the control group. Note that the treatment effect coefficients in column (1) are a weighted average of those in columns (2) and (3). We see that the Passive Task leads to a significant impact on the exact Task Questions relative to the control group, but no significant impact on Novel Questions. On the other hand, the Video Creation treatment leads to a significant impact on both exact Task Questions and on Novel Questions. This highlights that while Passive Tasks might be effective at helping students learn in a way that is akin to rote memorization, the Video Creation treatment is effective at helping students learn in a way that allows for knowledge transfer.

Table 1.3: Treatment Effect on Post-Test relative to Control

	(1) Post-Test	(2) Task Qs	(3) Novel Qs
Passive Task	0.124**	0.183***	0.0563
	(0.0625)	(0.0664)	(0.0579)
Video Creation	0.291***	0.354***	0.187***
	(0.0625)	(0.0670)	(0.0574)
Baseline Score Controlled	Yes	Yes	Yes
Teacher Fixed Effects	Yes	Yes	Yes
Observations	1,869	1,869	1,869

Note: The Passive Tasks variable in the first row is binary with a value of 1 for students in Passive Tasks group and 0 otherwise. The Video Creations variable in the second row is binary with a value of 1 for students in the Video Creations group and 0 otherwise. The outcome in Column 1 is the 15-question post-test consisting of standardized test questions that were relevant to the teacher's curriculum. The outcome in Column 2 is the student's score on the subset of the post-test questions that were exact treatment tasks, while Column 3's outcome is the student's score on the remaining questions on the post-test. All outcomes are in standard deviation units. Clustered standard errors (at the classroom-level) are in parentheses.
*** p<0.01, ** p<0.05, * p<0.1.

1.6 Mechanisms

Two potential mechanisms that could explain this treatment effect are 1) student effort and 2) student confidence. I pre-register that student effort would be proxied by the likelihood that students click on the "Help Video" link provided to them for each task, and that student confidence is estimated by a self-reported survey question administered right before the pre-test and post-test.

1.6.1 Student Effort

In each task, students in both the Passive Task and Video Creation classrooms were provided a link to a "Help" video on YouTube. While the video was the

same for the Passive Task and Video Creation classrooms for any given teacher-task dyad, the links in the Google Forms were embedded in separate bitly links for each classroom. The bitly link allows for tracking of the total number of times that a link has been clicked. While it cannot be ascertained whether a given student clicked on a link, or even whether the same link was clicked multiple times by the same person, this total provides an estimate of the likelihood that a student in a given classroom clicked on the link.

I estimate V_k as the Viewing likelihood for any given classroom. This is computed by taking the total number of clicks for a given classroom across all tasks, and then dividing it by the total number of students in the classroom, and then again dividing by the number of tasks. I estimate the following regression, where the unit of observation is a classroom:

$$V_k = \alpha_0 + \alpha_1 C_k + u_k$$

Here, V_k is the Viewing likelihood, C_k is a binary indicator where 1 = Video Creation class period and 0 otherwise, and u_k = error term. Note that control classrooms are excluded from this by construction because not having any tasks means that a Viewing likelihood cannot be estimated. The results for this regression are shown in Table 1.4.

The average likelihood of clicking on a video is 11% for students in Passive Task classrooms, and is over twice that (25%) for studentss in Video Creation classrooms. This result contributes to disentangling the opposing forces on student effort described in section 1.3. Here, a student who is inelastic with regards to the total amount of time invested in math would be *less* likely to click on this help video

Table 1.4: Impact of Video Creation Treatment on Click Likelihood

	(1) Click Likelihood
Video Creation	0.143*** (0.0341)
Constant	0.110*** (0.0239)
Observations	100
R-squared	0.153

Note: Each unit of observation in this regression is a classroom. Video Creation is a binary variable with a value of 1 if that classroom was assigned to the Video Creation treatment, and 0 if assigned to the Passive Task treatment. The outcome variable is the probability of a student in that classroom clicking on the help video resource for any given task. The standard error of the coefficients are in parenthesis.
*** p<0.01, ** p<0.05, * p<0.1.

link when assigned to the Video Creation treatment relative to the Passive Task treatment. On the other hand, a student who puts a high enough value on the video viewer's perception of them and treats passive and active effort as complementary would be more likely to click on the link. These results highlight that the video creation task might be sufficient enough to induce student motivation to apply more effort, which is generally hard to move (Burgess et al., 2021; Cotton et al., 2020). Additionally, this shows that a mechanism through which teaching can help people learn is that teaching induces people to put in more time preparing, implying that lab studies that hold the total amount of preparation time constant are underestimating the true general equilibrium impact of "learning by teaching" interventions.

1.6.2 Student Confidence

One possible mechanism through which this intervention would improve student skills is through improving their confidence in their math skills, and this confidence would lead to a skill-improving change in "mindset" (Yeager and Dweck, 2012). If this were the case, we would expect the treatment students to end the semester with higher changes in confidence than the control group. Another possibility is that as the students in the Video Creation classrooms explained math in their videos, their lack of initial understanding became clearer to them. This is what psychologists call "Illusion of Explanatory Depth" (Rozenblit and Keil, 2002). If this is prevalent, then we would expect a zero or negative change in math confidence among the treatment group. Table 1.5 shows the regression results of the treatment on math confidence at the end of the semester, normalized for each teacher and measured in standard deviation units.

I find that the video creation treatment led to an insignificant change in math confidence. One possibility is that both mechanisms are at play, and canceling each other out, where math confidence is initially lowered when students explain math and get a realistic sense of their skills, but then sustained explanations over time makes them improve their overall math confidence as they gain experience with explaining math. Future iterations of this study could have a weekly confidence check-in question to explore the dynamics of confidence as students continue to produce videos.

Table 1.5: Impact of Video Creation Treatment on Math Confidence

	(1) Math Confidence
Assigned Either Task (Passive or Video Creation)	0.0152 (0.0430)
Video Creation	-0.0290 (0.0313)
Baseline Confidence	0.665*** (0.0182)
Teacher Fixed Effects	Yes
Observations	1,779

Note: The Assigned Either Task variable in the first row is binary with a value of 0 for students in the control group, and a value of 1 for students in either the Passive Task or Video Creation group. The Video Creation variable in the second row is binary with a value of 0 for students in either the control or Passive Task group, and a value of 1 for students in the Video Creation group. The outcome is the student's overall self-reported level of confidence in their math skills (on a scale from 1 to 10), measured in standard deviation units. Clustered standard errors (at the classroom-level) are in parentheses.
*** p<0.01, ** p<0.05, * p<0.1.

1.7 Distribution of Treatment Effects

One feature of this experimental design is that it allows for the estimation of an individual treatment effect for each teacher. Class sections (e.g. 3rd versus 4th period) are typically randomly assigned in schools, and this study ensures that each teacher has at least one class period that creates videos and one period that does the passive task. While there might still be classroom peer effects at play, and while the sample size for any given teacher is typically only between 30-90 students, we can think of this study as having conducted a set of 41 small RCTs, one for each teacher.

I estimate the treatment effect of the Video Creation treatment relative to the Passive Task treatment on the PSAT post-test for each teacher individually, controlling for the baseline PSAT pre-test. This allows for a distribution of treatment effects which is shown in Figure 1.3. From this, we can see that approximately 80% of teachers had a positive treatment effect, and 20% had a negative treatment effect.

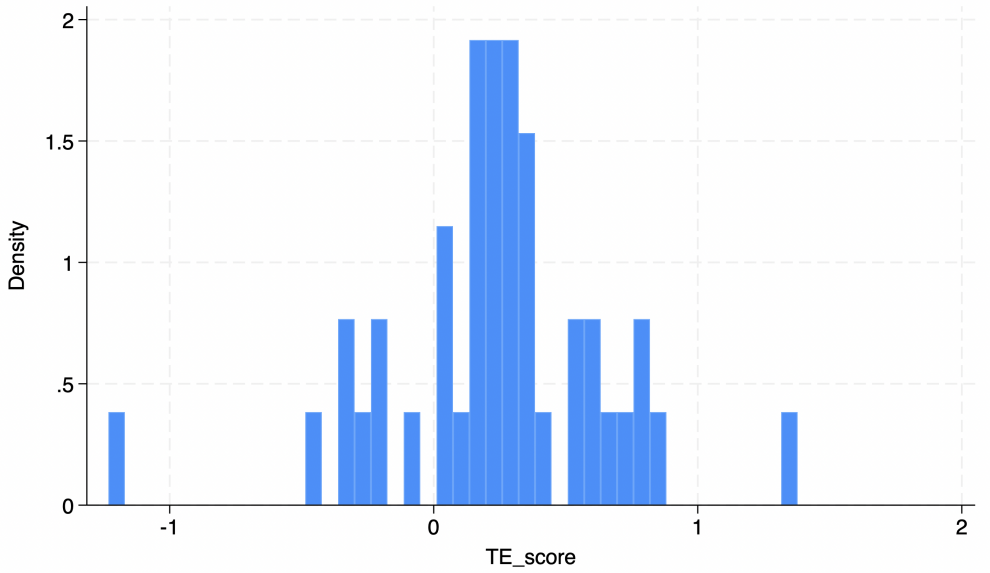


Figure 1.3: Distribution of Treatment Effects

As described in Section 1.3, a negative treatment effect is possible in this design for students who would choose to complete their assigned task if they are in the Passive Task classroom, but not if they are in the Video Creation classroom. The distribution in Figure 1.3 shows that some teachers had negative treatment effects, including one that had a large (greater than 1 SD) negative treatment effect. I assess whether the difference in task completion rate between a teacher's Passive Task and Video Creation classrooms predicts a teacher's treatment effect size.

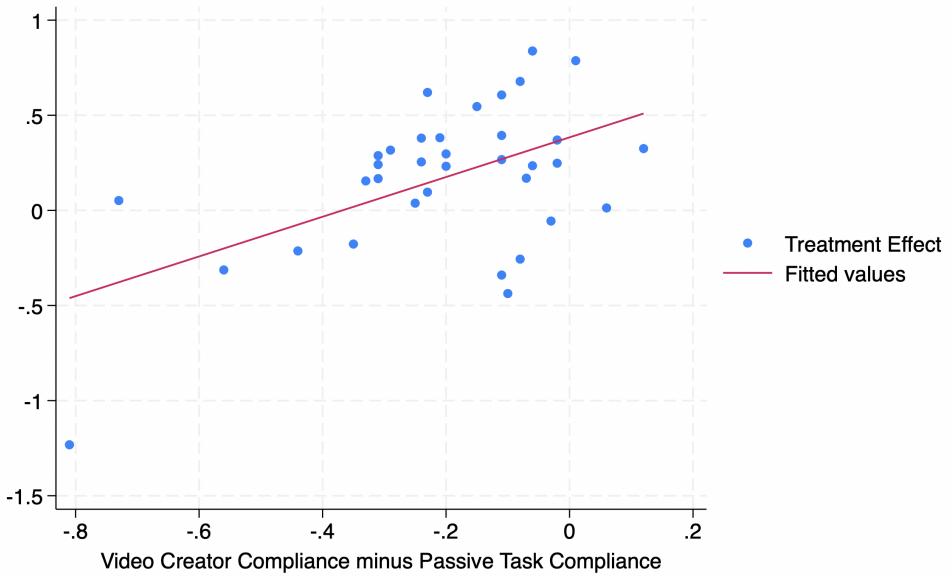


Figure 1.4: Relationship Between Compliance and Treatment Effect

Figure 1.4 shows the relationship between compliance and treatment effect. The x-axis shows a teacher's task completion rate for their Video Creation classroom minus that of their Passive Task classroom, and the y-axis shows the treatment effect size. Note that most teachers have a negative value for the x-axis variable because the task completion rate for their Passive Task classrooms was greater than that of their Video Creation classroom. We visually see a positive relationship, which is confirmed by a statistically significant positive slope as shown in Table 1.6.

This implies that teachers who were able to induce high compliance in their video creation classroom were able to achieve a high treatment effect. This relationship is not causal, so it is unclear whether this relationship is because of the high compliance rate or because of some other characteristic about the teachers who managed to induce a high compliance rate. Future work could see whether ex-

Table 1.6: OLS regression of Compliance on Treatment Effect

	(1) Treatment Effect
Difference in Compliance Rate (Video Creation minus Passive Task)	1.044*** (0.290)
Observations	35
R-squared	0.281

Note: The independent variable in this regression is the compliance rate of that teacher's Video Creation class section minus the compliance rate of that teacher's Passive Task class section. The outcome variable is the treatment effect, measured in standard deviation units, for each teacher. The standard error of the coefficient is in parenthesis.
*** p<0.01, ** p<0.05, * p<0.1.

perimentally inducing higher compliance, for example by having these assignments count for a high fraction of the overall course grade, leads to a higher treatment effect.

1.8 Conclusion

One goal of social science research is to estimate the impact of interventions that have the potential to be implemented widely. A pilot version of this intervention involved randomly assigning some students to a treatment where they tutored a student in a grade below them. Even when measures were taken to ensure that the tutor had higher baseline skills than the student, the compliance rate was exceedingly low. High financial incentives were not enough to induce a sufficiently large compliance rate to detect an Intent-to-Treat effect. There were additional costs involved in logistically ensuring that students had a space (even if it was virtual, a shared link) to meet. The nature of the intervention tested in this

project was a result of attempting to ensure that this paper tested a version of a program that we may expect when implemented on a larger scale, or "policy-based" evidence as per List (2024b). The direct implementation cost of this project was approximately $10 per student.[6] Given the treatment effects in this study, the benefit-cost ratio is estimated to be on the high end of other successfully implemented educational interventions (Guryan et al., 2023; Kline and Walters, 2016)

However, as this intervention and others like it are scaled, we should be mindful of the distribution of treatment effects. In particular, if some students have a high probability of a negative treatment effect, then scaling should be done with caution. The treatment effect in this study was consistently positive for teachers who had a high compliance rate of video creation relative to the passive task. If a teacher notices that their compliance rate is low and cannot do much to change it, then having the non-compliers engage in the passive task might be a way to maximize societal welfare. Of course, there is a moral hazard if students know that they will be assigned a less challenging task if they simply refuse to participate for the first few assignments.

Future versions of this project that have a larger number of teachers and covariates could utilize machine learning methods such as causal forest to help identify determinants of compliance to judge whether teachers should assign these tasks (Davis and Heller, 2017; Wager and Athey, 2018). Future work could also analyze the student generated videos to see if there are characteristics of created videos that are predicted to have higher treatment effects[7] Additionally, there might be

6. $25,000 cost, primarily on teacher incentives, for approximately 2,500 students

7. The data agreements made with districts in this study did not allow for matching videos

potential for a version of this intervention to impact learning in other contexts, such as training workers to improve productivity by asking them to teach or create video explanations for others.

with test scores and did not allow researcher access to videos from enough classrooms.

CHAPTER 2

THE TRADE-OFF BETWEEN QUALITY AND QUANTITY

2.1 Introduction

Recent work has spurred nationwide interest in the promise of high-dosage tutoring as a means to close the achievement gap for adolescents (Guryan et al., 2023; Nickow et al., 2020). This has led to several studies on how tutoring can be delivered at scale (Cortes et al., 2024; Kraft et al., 2022; Kraft and Lovison, 2024; Kraft et al., 2024; Robinson et al., 2022, 2024). Two features of tutoring to consider are 1) tutoring group size and 2) tutoring frequency. Some work suggests that it would be ideal to have a group size of no more than 2 students, and a frequency of 3-4 times per week (Guryan et al., 2023). However, the cost involved in doing this might not be feasible for many schools. If a school were to implement a lower-cost version of this ideal tutoring program, should they cut their spending by having larger groups, or by reducing the frequency?

In this paper, I evaluate the impact of an Indiana KIPP[1] charter middle school's math tutoring program. All 343 students in the middle school (grades 6-8) were randomly assigned to either a control condition or to receive in-school math tutoring during another elective class period. Of the 149 students who were randomly assigned to receive tutoring, 62 were randomly assigned to receive tutoring in 2-student groups twice per week, and 87 were randomly assigned to receive tutoring in 3-student groups thrice per week. The total cost per student is equal in these two

1. https://en.wikipedia.org/wiki/KIPP

treatments ($40 per week per student). So, this design allows us to test whether a budget-constrained school would get a higher return on investment from providing more frequent tutoring in larger groups, or less frequent tutoring in smaller groups.

Overall, I find that the 2-student group tutoring led to a significant improvement on math skills (0.23 SD), whereas the equal-cost, more frequent tutoring in the 3-student groups did not lead to a significant improvement in math skills. These results are robust to: alternative specifications of the outcome variable; "tutor" fixed-effectss; attrition tests; and multiple hypothesis testing concerns (List, 2024a).

This study contributes to the literature on human capital development. Many interventions aimed at improving child skills target either the home environment or the classroom environment (Alsan and Eichmeyer, 2021; Kalil et al., 2024, 2023a; Mayer et al., 2023; Shah et al., 2023; York et al., 2019). This study is an example of an intervention that takes place during the school day but not by teachers, thereby targeting student skills without monopolizing either teachers' or parents' time.

This study also contributes to our understanding of the nature of the educational production function (Lazear, 2001). The results in this study imply that the gain in instructional quality from a smaller group size outweighs the gains from a larger volume of exposure to tutoring. This is consistent with other work showing that reducing tutoring group size can improve outcomes (Kraft and Lovison, 2024; Robinson et al., 2024). However, these studies show that the more expensive intervention (smaller group size) is better, which implies that additional assumptions must be made about the nature of the cost and benefit functions to judge which version of the program has a higher ROI. The design in the current study frees us

from needing to make such assumptions in determining the ROI. This work also relates to the literature on most efficient use of funds in experimental interventions in education (Fryer et al., 2022; List and Shah, 2022).

The rest of this paper is organized as follows. Section 2 describes the experimental setting and design. Section 3 reports the results, which are discussed in Section 4.

2.2 Experimental Design

2.2.1 Institutional Setting

The Knowledge is Power Program (KIPP) is the largest network of charter schools in the United States. This paper reports on an RCT done with one KIPP school in Indiana. The population served by this school is low-income, with 97% receiving free or reduced-price lunch. While the school is a K-12 school that serves over 1,500 students, the tutoring intervention focuses on middle school (grades 6, 7, and 8), which had 343 students enrolled as of January 2024.

The tutoring program began during the 2021-22 school year, and was implemented again in the 2022-23 school year. The tutoring done in school, and students are pulled from non-core classes (electives) during the school day to receive tutoring, mitigating the need to stay before or after school.

The experimental evaluation of this tutoring took place between January 2024 and May 2024. All enrolled students in middle school were automatically enrolled to be a part of the study.

2.2.2 Randomization

A total of 12 classrooms – four in each grade – participated in the study. Stratified by classroom, students within each classroom were randomly assigned to one of three treatment conditions: 1) Control, 2) Two-student group tutoring twice per week, and 3) Three-student group tutoring thrice per week.

To ensure that students in groups were of similar mathematical ability, students within each classroom were first sorted by baseline performance on the MAP math standardized test, and then students were (on paper) paired into groups of 2's and 3's with the students closest to them on baseline performance. Then, these "groups" of 2's and 3's were assigned to either receive tutoring or to be in the control group. This process ensures that the tutoring groups consisted of students with relatively homogeneous skills, but also that students across the skills spectrum all had a chance of being assigned to receive tutoring. The full set of randomization protocols is described in the Pre-Analysis Plan[2].

2.2.3 Treatment

From January until May of 2024, 149 of the 343 students were randomly assigned to receive tutoring in either groups of 2 or 3 students. Of the 149 students who received tutoring, 62 were randomly assigned to receive tutoring in 2-student groups twice per week, and 87 were randomly assigned to receive tutoring in 3-student groups thrice per week.

2. `https://www.socialscienceregistry.org/trials/12858`

2.2.4 Outcome

The outcome variable of this study is the MAP Math assessment, which is created by the NWEA. NWEA assessments are used by over 50,000 schools and districts in 149 countries. There are over 16.2 million students using NWEA[3]. This assessment is aligned with widely used instructional standards. The MAP Growth uses the RIT (Rasch Unit) scale to help measure and compare academic achievement and growth. Specifically, the scale measures levels in academic difficulty. The RIT scale extends equally across all grades, making it possible to compare a student's score at various points throughout their education.[4]

2.2.5 Missing Data

Some students who moved out of the district during the intervention, and others were not present during the day of the state-administered MAP Math test. In these cases where an endline score is not available, I count the students as having attrited from the sample. I perform attrition tests in the appendix to show that there is no systematic difference in attrition by treatment condition.

2.3 Results

2.3.1 Descriptive Results

The available administrative data for each student included the student's gender, race, age, and baseline performance (measured in January 2024) on the MAP

3. For more details about the MAP assessment and RIT scores, see: `https://www.nwea.org/`

4. See `https://teach.mapnwea.org/impl/maphelp/Content/AboutMAP/WhatRITMeans.htm`

math assessment.

Overall, the baseline observable characteristics are balanced. The attrition rate is also not significantly different across treatment conditions.

Compliance

Over the course of the 12 weeks of the intervention, the 2-student group treatment had a median of **16 sessions** (1.3 per week), with an average group size of 1.88 students per session. The 3-student group treatment had a median of **21 sessions** (1.8 per week), with an average group size of 2.65 students per session.

Because of absences, those assigned 2-student groups sometimes had 1:1 sessions, and those assigned to 3-student groups occasionally had 2:1 or 1:1 sessions. This brought the session group sizes down to 1.88 (rather than 2) and 2.65 (rather than 3) respectively. Additionally, students in both treatment conditions only attended about two-thirds as many sessions as originally intended. While the dosage was not as high as initially planned, the ratio of the number of additional sessions received by the 3-student groups relative to the 2-student groups was consistent with the design to keep the cost the same per student.

The histograms above in Figures 2.1 and 2.2 show the distribution of attendance for students in the 2-student and 3-student groups respectively. The maximum sessions for the 2-student group was 20, whereas that of the 3-student groups was 28. The median number of tutoring sessions was 16 for the 2-student groups, and 21 for the 3-student groups. Both have a left-skew distribution, with a majority of students attending at least 75% of the required sessions, and a handful of students that were chronically absent.

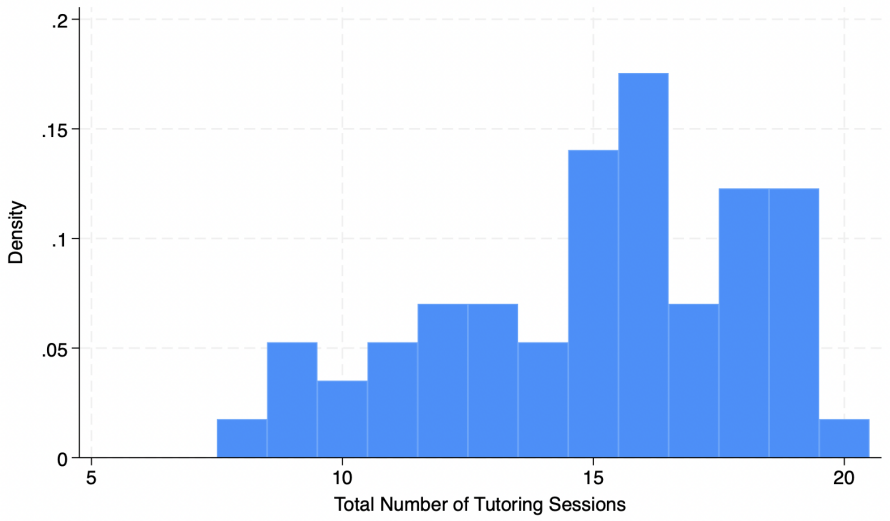


Figure 2.1: Histogram for Total number of sessions for 2-student groups

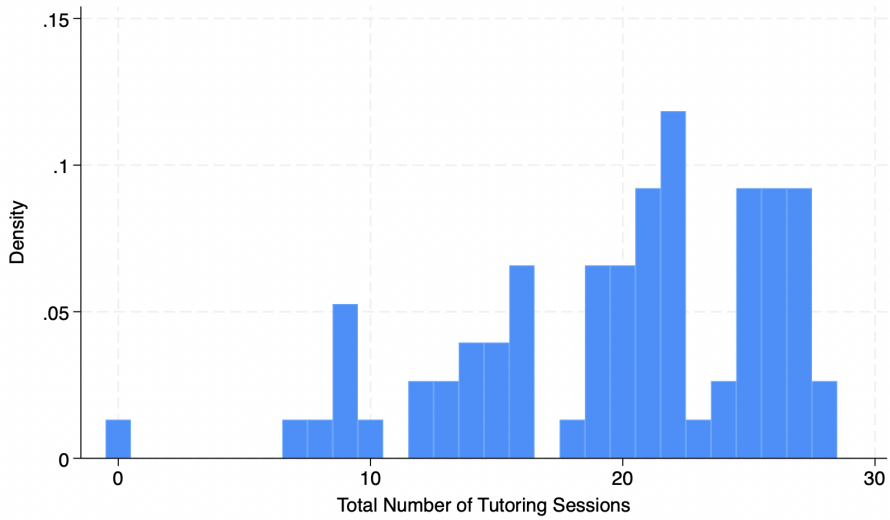


Figure 2.2: Histogram for Total number of sessions for 3-student groups

2.3.2 Main Regression Results

The following pre-registered regression was used to estimate the treatment effect on math skills:

$$y_{i,t} = \beta_0 + \beta_1 T_{2i} + \beta_2 T_{3i} + \beta_3 y_{i,t-1} + \gamma_j + \varepsilon_i$$

Here, $y_{i,t}$ is the value of outcome at the end of the experiment, $y_{i,t-1}$ is the baseline level of outcome, T_{2i} is a binary indicator where 1 = assigned to 2-group tutoring twice a week and 0 otherwise, T_{3i} is a binary indicator where 1 = assigned to 3-group tutoring thrice a week and 0 otherwise, γ_j is the teacher fixed effect for teacher j, and ε_i is the error term.

Here, we see in Table 2.1 that there was no significant impact of the 3-student group tutoring on math skills, but the 2-student group tutoring led to an approximately 4 point increase on the MAP math assessment. This is approximately 0.23 Standard Deviations, which is economically substantive. The second column shows that the results do not change when the available covariates are included as controls in our regression.

Table 2.2 shows that the results are practically similar when we choose alternative specifications of the outcome. In the first column, we see that the 2-student group tutoring led to a 6 percentile increase in MAP math performance, which is substantial. There was no significant impact for the 3-student group tutoring. The second column uses the growth in MAP math score as an outcome (instead of using the endline value as the outcome and the baseline value as a control). Here again we find that there is a similar, large and significant impact of the 2-student group tutoring on math skills, but no impact of the 3-student group tutoring.

Table 2.1: Impact of Group Tutoring on Math Scores

	(1) RIT Score	(2) RIT Score
2-Student Group Treatment	3.966***	3.866***
	(1.349)	(1.356)
3-Student Group Treatment	0.421	0.582
	(1.076)	(1.022)
Baseline RIT Score	0.821***	0.811***
	(0.0416)	(0.0421)
Female		-0.812
		(0.908)
Black		-3.346
		(2.060)
Hispanic		-2.097
		(2.264)
Age		-3.087***
		(0.905)
Constant	41.81***	87.84***
	(8.747)	(15.45)
Observations	302	302
Classroom Fixed Effects	Yes	Yes

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Note: Clustered standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1.

Table 2.2: Impact of Group Tutoring on Math Percentile

	(1) RIT Percentile	(2) RIT Score Growth
2-Student Group Treatment	6.136***	3.382**
	(1.980)	(1.375)
3-Student Group Treatment	1.648	-0.287
	(1.596)	(1.072)
Baseline Percentile	0.843***	
	(0.0391)	
Constant	7.014***	4.477***
	(1.188)	(0.687)
Observations	302	302
Classroom Fixed Effects	Yes	Yes

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Note: Clustered standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1.

2.3.3 Robustness Checks

Table 2.3 shows the main results when including a tutor fixed effect. Note that this is only possible to run for a regression that only involves students who received tutoring, so the comparison is between the 2-student and 3-student groups, rather than between the treatments and the control group. Further, tutors who only taught one type of student had to be dropped. This leaves us with an underpowered sample, but the magnitudes of the treatment effect are consistent in showing that the 2-student group tutoring leads to about a 0.2 SD point estimate improvement in math skills relative to the 3-student group tutoring.

Table 2.3: Treatment Effect with Tutor Fixed Effects

	(1) RIT Score	(2) RIT Percentile
3-Student Group Treatment	-3.695*	-5.164
	(2.017)	(3.100)
Baseline RIT Score	0.886***	
	(0.0875)	
Baseline Percentile		0.860***
		(0.0769)
Constant	31.68	12.51**
	(19.62)	(5.545)
Observations	89	89
Classroom Fixed Effects	Yes	Yes
Tutor Fixed Effects	Yes	Yes

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Note: Clustered standard errors in parentheses.
*** p<0.01, ** p<0.05, * p<0.1.

2.4 Discussion

Taken at face value, the results of this study imply that it would be advantageous to limit tutoring groups to a size of 2 students, even if it means tutoring has to be less frequent to make up for the additional cost. This might be because a small enough group size allows for closer relationships to develop between the tutor and students, and that this is necessary for tutoring to be effective.

One potential concern is that while the 3-student group sessions were assigned to receive 1.5 times as many sessions as those in the 2-student groups, they only received 1.33 times as many sessions in practice. However, this difference in im-

plementation does not seem substantial enough to explain the null effect of the 3-student group tutoring relative to control. Note that we can rule out that the tutors who performed the 2-student group tutoring were of higher quality than those who performed the 3-student group tutoring, because tutors were assigned both types of sessions, and the results are consistent when we include tutor fixed effects.

A caveat in generalizing this result is that the tutors were college students, and while they received some tutor training, they did not receive extensive training on classroom management. If one's ability to tutor 3 students well depends on classroom management skills, then it might be the case that tutors with classroom teaching experience would fare better in the 3-student group tutoring sessions than the tutors in this study did.

CHAPTER 3

ENGAGING PARENTS WITH PRESCHOOLS

3.1 Introduction

Many observational studies show that parents who attend school events, volunteer at school, are in communication with teachers, and engage in school activities have children who perform better in school (for example, Castro et al., 2015; Dizon-Ross, 2019; Domina, 2005; Hill and Tyson, 2009; Liu and White, 2017; McNeal Jr, 2012; Wang and Sheikh-Khalil, 2014). There is some causal evidence of the positive effects of various forms of parental engagement in schools (Avvisati et al., 2014). Consistent with the high perceived benefit of parent engagement, publicly supported preschools such as Head Start are required to spend substantial funds promoting it (Zigler and Muenchow, 1992).

Despite their prevalence, there is almost no experimental work evaluating the benefits of parent engagement programs. An RCT to evaluate existing parent engagement programs would need to exogenously induce parents to attend the engagement programs to which they currently have access. Such an evaluation would need to have a strong "first stage," whereby a substantially larger fraction of treatment parents attends engagement programs relative to control parents. Existing research, which shows that parental attendance tends to be low, suggests that a novel approach is needed (Avvisati et al., 2010; Marti et al., 2018; Mendez, 2010). In this paper, we test whether a behaviorally informed intervention that combines financial incentives and behavioral tools could lead to a substantial increase in parent attendance at engagement events. If successful, this approach could not only

be used by programs themselves but could be incorporated into future research to test the causal impact of parental attendance on child outcomes.

We use an RCT to test the combined impact of loss-framed financial incentives and text-message reminders on parental attendance for 319 parents at preschool-sponsored family engagement events at six subsidized preschools in Chicago, IL from November 2018 to March 2019. Our experiment used an opt-out design. Only 20 parents opted out. We include the opt-outs in the main analysis to identify the intent-to-treat effect because it is more policy-relevant.

Treatment parents were offered $25 per event for eight events. All parents in our sample earned below the federal poverty line ($25,100 per year for a family of four at the time of the study), and had an estimated median hourly wage of $10. The average length of an event was 90 minutes. If we assume 30 minutes of commute time each way to and from the events, the $25 compensation for 2.5 hours is approximately equal to the parents' median hourly wage. In addition to weekly text message reminders with details about the event(s) that week, parents in the treatment group received their financial incentive using a loss-frame. That is, parents were initially given $200 in a virtual account (redeemable at the end of the experiment), but $25 was deducted from their account for each missed event. Each week, treatment parents received a second text message that indicated how much money they had remaining in their account.

We find that the treatment led to no significant difference in the fraction of parents that attended at least one event; 57% of control parents and 56% of treatment parents attended at least one event. However, we find a statistically significant 32% (7 percentage-point) increase in the attendance rate among parents who already attended at least one event: among parents who attended at least one event, the

overall attendance rate across all eight events was 22% for control parents and 29% for treatment parents. There was no significant heterogeneity in treatment effect by time of day or event length. We develop a theoretical framework to discuss our findings. One reason for the lack of treatment effect and low overall attendance rates might be that parents perceive such school-sponsored events as having a low expected return on investment for their time.

Our treatment design was based on existing studies designed to increase parental engagement in children's learning, broadly defined, especially in low-income families. Several recent experimental studies show that tools drawn from behavioral economics can boost parental engagement and improve child outcomes, such as reading, math, and preschool attendance among low-income families (Kalil et al., 2021; List et al., 2018; Mayer et al., 2023, 2019).

Three experimental studies relate closely to the present work. Gennetian et al. (2019) and Hill et al. (2021) use a bundle of behavioral tools including reminders, commitment devices, and personalized invitations to increase parental attendance at preschool events. While Gennetian et al. (2019) find that behavioral tools increase attendance, Hill et al. (2021) find no significant effect. One behavioral tool not included in these studies is loss framing, which the present study tests. Fryer et al. (2015) test whether financial incentives increase parental attendance at parenting workshops at a preschool run by that research team. They found that a $100 incentive for a 90-minute workshop led to a substantial effect on attendance. The present study differs from Fryer et al. in two ways. First, the incentive we offer is only a quarter as large, and second, our study is situated in the ongoing parental engagement events and efforts offered by existing preschools in the community. This natural setting can help mitigate external validity concerns.

We contribute to the human capital development literature. Parent engagement is a crucial aspect of child development; we test whether engagement in school events can be increased in a natural setting. Given our opt-out design, the results are more likely to hold at scale as compared to studies requiring opt-in participation (Mayer et al., 2021). An experimental test for the effectiveness of parental engagement on student outcomes will require a treatment that could reliably provide a substantial exogenous increase in parental engagement. Our study provides evidence that modest financial incentives, even when combined with reminders, do not induce a substantial increase in engagement in preschool events.

We also contribute to the economics of education literature on parent communication. York et al. (2019) found that reminder messages to parents led to improved child literacy. Kraft and Rogers (2015) found that communication with parents of high school students increased the likelihood of passing a summer course. Castleman and Page (2017) found that a text message intervention helped increase parental engagement in the college enrollment process. Our intervention found a treatment effect only for parents who were already engaging with their school. This study provides evidence of the limitations of parent communication interventions in engaging otherwise disengaged parents.

This study also adds to the recent economics literature on combining financial incentives with other tools to change behavior (Arad et al., 2023; List and Shah, 2022). For example, Barrera-Osorio et al. (2020) provided financial assistance to parent associations in Mexico, as well as information to individual parents. They find that while information led to a change in parent behavior at home, financial assistance did not lead to behavior change for students or parents. Economists have used loss aversion to increase the effectiveness of incentives in education (Fryer

et al., 2022; Imas et al., 2017; Levitt et al., 2016). In some instances, loss-framed incentives may have negative consequences, such as neglecting unincentivized aspects of the task (Pierce et al., 2020). However, a recent field experiment in Uganda showed that loss-framed incentives can increase labor productivity (Bulte et al., 2020) and found no negative incentive spillovers (Bulte et al., 2021).

The remainder of this paper is organized as follows. In Section 2, we briefly describe the context of preschool parental engagement programs. In section 3, we describe our sample and intervention. In Section 4, we present our main results. In Section 5, we describe a theoretical model we use to interpret our results. Section 6 concludes with policy implications.

3.2 Institutional Background

Although research has demonstrated that increasing parents' direct engagement with their children increases children's academic success (Cunha et al., 2006; Fiorini and Keane, 2014; Heckman and Masterov, 2007; Villena-Roldan and Ríos-Aguilar, 2012), there is little evidence on the efficacy of the kinds of programs that preschools provide under the umbrella of family engagement. From its inception, Head Start has emphasized parent involvement. A founding principle of the Head Start program was the "maximum feasible participation" of the parents (Harmon, 2004; Zigler and Muenchow, 1992; Zigler and Styfco, 2010). The Head Start Code of Federal Regulations[1] states:

> A program must, at a minimum, offer opportunities for parents to

1. Chapter XIII part 1302.51b; available online at `https://eclkc.ohs.acf.hhs.gov/policy/45-cfr-chap-xiii/1302-51-parent-activities-promote-child-learning-development`

> participate in a research-based parenting curriculum that builds on parents' knowledge and offers parents the opportunity to practice parenting skills to promote children's learning and development.

Parent engagement in preschool is also required in the "Every Student Succeeds Act," and many state requirements for preschools also follow Head Start's code for parent engagement.[2] A few observational studies have considered parent engagement in these types of programs, mainly finding positive correlations between parent engagement and child academic and behavioral outcomes (Ansari and Gershoff, 2016).

Regulations are clear that preschool programs should promote parent engagement. Yet, there are few guidelines about what preschools should do to accomplish this. The Head Start Performance Standards hold that teachers must regularly communicate with parents about their child's schooling, hold at least two parent conferences a year, have at least two home visits, provide parents the chance to volunteer at the school, implement intentional strategies to engage parents in their children's learning and development, offer activities that support parent-child relationships and child development including language, dual language, literacy, and bi-literacy development as appropriate, and provide family engagement services in the language and cultural context of the family.[3]

In addition to the federal regulations, The National Head Start Association promotes the *Two Generations Together* initiative, which is focused on increasing awareness of the "two-generation" adult education and job training models that

2. See `https://www.everystudentsucceedsact.org/title-1--1-1-3-1-1-1-1-1`

3. See the list of requirements here: `https://eclkc.ohs.acf.hhs.gov/policy/45-cfr-chap-xiii/1302-34-parent-family-engagement-education-child-development-services`

are part of the comprehensive child and family services delivered by Head Start programs across the country (Dropkin and Jauregui, 2015). Two-generation approaches focus on creating opportunities for and addressing the needs of children and their caregivers together to create economic stability for the family. Providing parent engagement programs is not costless: not only must preschools dedicate personnel and space to host and promote the events, but regulations also require preschools to develop and submit plans for engaging parents. Those plans sometimes must be reviewed by multiple individuals.[4]

Family engagement activities offered by Head Start programs include school information sessions, conferences with teachers, social events, parent-child activities, parent education programs, social service programs, and volunteer opportunities at the school. In general, research shows that disadvantaged parents communicate less with their children's teachers and are less likely to attend parent-teacher meetings and other school events (McQuiggan and Megra, 2017; Turney and Kao, 2009). So, it is perhaps not surprising that parental attendance at preschool-sponsored parent engagement events tends to be low (Avvisati et al., 2010). For example, Mendez (2010) reported that parents attended fewer than two parent-engagement workshops out of nine offered. In another descriptive study, average attendance at Head Start parenting events was reported to be about 21% (Marti et al., 2018).

At least three reasons might explain the low attendance. One reason is structural barriers to attendance, such as work conflicts, lack of transportation, or other inflexible obligations. A second reason might be that parents don't believe the events are worth attending. This may be because the expected benefit from

4. See `https://www.everystudentsucceedsact.org/title-1--1-1-3-1-1-1-1-1`, Section 1010, parts B and C

attending is low or because the opportunity cost of attending is high. A third potential reason for low attendance is that cognitive biases, such as inattention or present bias, might influence parenting decisions (Mayer et al., 2019). Parents may want to attend an event and believe it is worthwhile to attend, but simply forget to attend or forget to arrange transportation to attend. The intervention we describe here addresses the second and third of these reasons simultaneously: financial incentives are intended to offset parents' opportunity cost, and reminder messages are intended to offset the influence of present bias and inattention.

3.3 Experimental Design

3.3.1 Sample and data collection

This experiment was conducted between November 2018 and March 2019 at six preschools serving low-income children in Chicago. Five of these preschools were Head Start centers. From administrative records, we were able to access three covariates: the child's age, the child's gender, and the family's primary language. While parent characteristics were not available at the individual level, the following averages were provided by the preschools: 9% of parents were unemployed (twice the national average), the hourly wage was $10 (40% below the national average at the time), 8% have a bachelor's degree or higher, and the racial breakdown consists of approximately 39% black, 38% Hispanic, 12% white, 4% Asian, and 6% Multi-racial families.

There were 319 parents in total at the six centers, of whom 159 were assigned to the control group and 160 to the treatment group. The randomization was stratified by center to ensure that treatment status was balanced within each

center. Schools informed parents that they might receive text messages as a part of a project designed to increase parent engagement in preschools. While schools did not mention randomization or financial incentives, it is possible that some parents realized that they were randomly assigned to receive (or not receive) messages and financial incentives by communicating with parents in the other group. All parents at a center were automatically enrolled in the experiment unless they opted out via text, email, or phone call. Twenty parents opted out (this is discussed in more detail in Section 3.4.3).

During the 4-month intervention period, different centers had varying numbers of parent events but no mechanism for collecting attendance at these events, so research assistants attended the events in person and collected attendance. We counted the participant as being "present" for a given event as long as any adult member of their child's family attended that event.

To maintain consistency across centers, we chose eight events per center for which we tracked attendance.[5] Any event open to all parents was eligible for inclusion, and if a center had more than eight such events, we chose a random subset of eight to include in our intervention. For centers that had fewer than eight events, we organized additional events with that center so they could offer eight events to parents.

Our primary outcome is a binary variable indicating whether or not the parent attended at least one event. Using a linear probability model, the treatment effect will indicate the fraction of parents who were induced to attend at least one event. Our secondary outcome measure is the attendance rate, which is the proportion of

5. at one center, the total ended up being seven events due to unanticipated weather-related cancellation

the eight events that each parent attended. For example, if a parent attended two out of eight events, we calculated their attendance rate as 25%. In addition, four of the six centers offered at least one event beyond the eight used in our study. We also measured attendance at these additional events, which allows us to assess whether there were any positive or negative incentive spillovers (see Section 3.4.5).

Events lasted from 60 to 120 minutes, with a median of 90 minutes. No event was offered on a weekend, and 60% were offered on a Thursday. The start times of events ranged from 7:30 AM to 6:00 PM, with a median start time of 4:00 PM and a mode of 3:30 PM. Most events focused on parent-child interactions, such as "Holiday Crafts Night", "Healthy Smiles Parent-Child Workshop", and "Chili and Chill Family Fun Night."

A few focused exclusively on parents, such as "Digital Literacy" and "Employment Workshop." For these events where the children are not expected to join, all schools provided on-site childcare for the duration of the events.

3.3.2 Treatment description

We did not make any contact with the control group beyond taking attendance at events. In contrast, for the treatment group, we sent two text messages per week and offered them $25 for attending each of the eight events. One of the text messages (sent on Sundays at 6:00 PM) reminded parents of events taking place that week as follows:

> *Plan to go to and sign in at [Preschool Name]'s event this week. You or another adult who cares for [Child Name] may go.*
>
> *[Event Name] is on [Day of Week] [Date] at [Time]. Hold onto the $*

in your Bank. Mark your calendar!

Parents received a second text message on Fridays at 6:00 PM, reminding them of the financial incentive. Treatment parents were told that they could redeem $200 at the end of the study (March 2019), but would lose $25 for each event missed. The weekly Friday text was as follows:

> *Your balance is [Amount] as of [Date]. Remember you started with 200 and lose 25 for every event you miss.*

Head Start is targeted to parents whose household income is at or below the federal poverty line. At the time of the study, the federal poverty line for a family of three was $20,780, and for a family of four was $25,100. The median hourly earnings of workers in the bottom quintile of the earnings distribution is $10.22 (Ross and Bateman, 2019). Assuming transportation to and from the event takes about an hour, the median total time spent to attend each event would be 2.5 hours. This makes our incentive equivalent to median parents' financial opportunity cost.

The cash incentive was provided in a behaviorally-informed design to capitalize on loss aversion. Loss aversion refers to the idea that people put a greater weight on losses than on equivalent gains. The theory of loss aversion implies that people will be more responsive when money is taken from them versus when an equal amount of money is given to them (Kahneman et al., 1991).

3.3.3 Limitations

One limitation of this design is that the treatment bundles financial incentives, loss framing, and text message reminders so we cannot identify their unique contri-

bution to any treatment impact. We bundled the treatments because our goal was to do a practical efficacy test to boost parental engagement as much as possible.

Another limitation is that parents could not collect their financial incentive until the end of the intervention. Adherence to this design may be challenging for parents who are present-biased; that is, disproportionately valuing the present. Present bias is a relevant cognitive bias that might affect parenting decisions (Mayer et al., 2019). As such, this works against our finding a treatment impact. The treatment impact could also be diminished if parents did not trust that they would receive the financial incentive. Theoretically, a clawback design can give the money upfront and take it back for failure to comply. We did not think it was ethical to adopt this approach with a sample of low-income parents.

Another compliance concern is that we cannot be sure that participants received the text messages or that their phone numbers did not change during the course of the intervention. We updated phone numbers if and when the schools did so. However, in our prior work with low-income parents we found that 97% of the phone numbers we had on file were still active one year after we had collected them (Kalil et al., 2020, 2023b,c; Mayer et al., 2023).

3.3.4 Power Analysis

To estimate the required sample size for this RCT, we assume the standard significance level of 0.05 and power of 0.80. Further, we assume that the available covariates will explain 20% of variation in the outcome, based on pilot testing. With these assumptions, a target sample size of 300 will allow us to detect a minimum effect size of 0.29 SD. While this is on the high end of expected effect

sizes for educational interventions, it is likely achievable for this project based on pilot testing because of the very low level of baseline engagement among parents. That is, given that most parents do not attend any events to begin with, a 0.29 SD treatment effect can be achieved with a modest increase in parental attendance.

3.4 Results

3.4.1 Descriptive statistics

Table 3.1 shows the means and standard deviations for the treatment and control groups for the three covariates available from administrative data. There was no significant difference across treatment and control for any of the variables.

Table 3.1: Descriptives and Balance Test

	Control			**Treatment**				
	N	Mean	SD	N	Mean	SD	Diff	SE
Female	159	0.51	0.50	160	0.50	0.50	−0.009	0.056
Spanish	159	0.22	0.42	160	0.24	0.43	0.024	0.047
Child's Age	159	4.18	0.63	160	4.08	0.61	−0.104	0.070

Note: The Diff column is the coefficient of a regression of treatment status on the variable, and SE is the robust standard error of that coefficient. Female is the proportion of children who are female. Spanish is the proportion of whose primary language is Spanish rather than English. Child's Age is the child's age (in years) as of November 2018.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure 3.1 shows the distribution of total number of events attended by parents. The mode is zero and the median is one. Our primary outcome variable is a binary variable with a value of 1 for parents who attended at least one event, and 0 otherwise. Our secondary outcome is attendance rate, which we compute by dividing the total number of events a parent attended by the total number of events offered (eight for 89% of parents, and seven for the remaining 11% whose

school had an unexpected weather-related cancellation for the eighth event).

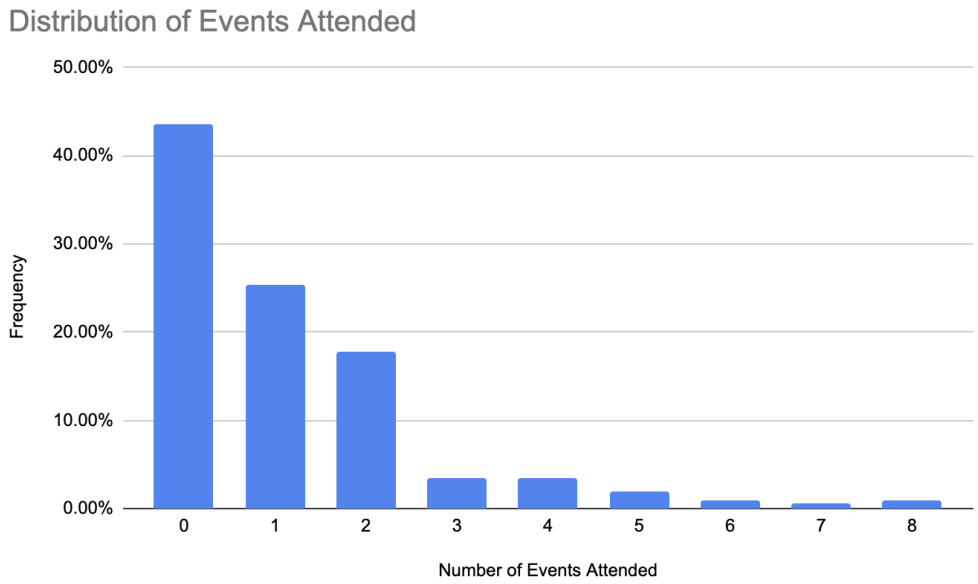


Figure 3.1: Distribution of Total Events Attended

3.4.2 Extensive Margin Treatment Effect

We use our experimental data to estimate the following regression:

$$A_i = \alpha + \beta T_i + \gamma X_i + \delta_s + \varepsilon_i$$

A_i is an indicator for whether parent i attended at least one event, T_i is the treatment status indicator, X_i is a vector of observable child demographics (age, gender, and Spanish as primary language), δ_s represents school fixed effects, and ε_i is the error term. Table 3.2 provides estimates of β, with and without school fixed effects and child covariates.

We find that 57% of parents in the control group attended at least one event, and that fraction in the treatment group (56%) was not statistically significantly

Table 3.2: Treatment Effect (Extensive Margin) on Attendance

	(1) Attended	(2) Attended	(3) Attended
Treatment	-0.0161	-0.0145	-0.0135
	(0.0557)	(0.0491)	(0.0493)
Female			0.0394
			(0.0498)
Spanish			0.0652
			(0.0648)
Child Age (Years)			0.0240
			(0.0409)
Constant	0.572***	0.680***	0.562***
	(0.0394)	(0.0658)	(0.181)
Observations	319	319	319
School FE	No	Yes	Yes
Covariates	No	No	Yes

Note: Robust Standard errors are in parentheses. The outcome is a binary variable indicating whether or not the parent attended at least one event. Female is an indicator for whether the child is female, Spanish is an indicator for whether the household's primary language is Spanish, and Child Age is the child's age in years as of November 2018.
*** p<0.01, ** p<0.05, * p<0.1.

different ($p = 0.77$). The result does not change when the regression includes school fixed effects or covariates.

3.4.3 Attrition and Robustness Checks

Because the study had an opt-out design, all 319 parents were automatically enrolled in the study unless they opted-out via text, email, or phone call. 20 participants dropped out by week two of the study, but none did so after that.

However, the dropout rate was not balanced between treatment and control: 4 control parents (2.5% of the control group) dropped out, and 16 treatment parents (10% of the treatment group) dropped out. This difference is statistically significant ($p = 0.006$).

It was easier for treatment parents to opt out because they were already receiving text messages and could opt out simply by replying to a text. Control parents, on the other hand, would have had to initiate a text message, email, or phone call to opt out. It is also likely that the study was more salient to treatment parents, given the text messages they were receiving. This may have reminded them that they were in a study of which they no longer wished to be a part.

We stopped tracking a parent's attendance after they dropped out. If there is a relationship between a parent's decision to drop out and their attendance rate, then removing these 20 observations may lead to an imbalance in expected unobservable characteristics between treatment and control. Therefore, we kept these observations and imputed zero for their attendance for two reasons. First, the modal parent attended zero events and second, there was zero attendance among these 20 parents for the data we do have for them prior to their dropping out.

The binary outcome in the regressions in Table 3.2 has zero imputed as the outcome for the 20 missing observations. This might depress the true treatment effect in any case, because most of these observations were in the treatment group. In Table 3.3, we show the results of running the same regression as in Table 3.2, with some alternative imputations for the missing outcomes.

In column 1 of Table 3.3, the missing values of the outcome have 1 imputed, rather than 0. The treatment effect is still statistically insignificant, showing that

Table 3.3: Treatment Effect with Alternative Outcome Specifications

	(1) Attended (Alt)	(2) Attended (Low)	(3) Attended (High)
Treatment	0.0634	-0.0385	0.0883*
	(0.0471)	(0.0485)	(0.0480)
Female	0.0141	0.0358	0.0177
	(0.0482)	(0.0491)	(0.0489)
Spanish	0.0530	0.0663	0.0520
	(0.0619)	(0.0624)	(0.0641)
Child Age (Years)	0.0433	0.0248	0.0425
	(0.0382)	(0.0401)	(0.0391)
Constant	0.540***	0.589***	0.513***
	(0.169)	(0.177)	(0.174)
Observations	319	319	319
School FE	Yes	Yes	Yes
Covariates	Yes	Yes	Yes

Note: Robust Standard errors are in parentheses. All outcomes are a binary variable indicating whether or not the parent attended at least one event. The outcome in column 1 imputes "1" for all missing outcome data, as opposed to the 0 that was imputed in the regression in Table 3.2. Columns 2 and 3 impute missing data such that column 2 produces a lower bound on the treatment effect, and column 3 an upper bound. Specifically, column 2 imputes a 0 for all missing treatment observations and a 1 for all missing control observations, and column 3 does the reverse. Female is an indicator for whether the child is female, Spanish is an indicator for whether the household's primary language is Spanish, and Child Age is the child's age in years as of November 2018.
*** p<0.01, ** p<0.05, * p<0.1.

our main findings do not depend on whether we impute 0 or 1 for the missing values of the binary outcome. As a bounding exercise, columns 2 and 3 of Table 3.3 show the lower and upper bound of the treatment effect. In column 2, we impute 0 for missing treatment values, and 1 for missing control values, which gives the lower bound of the treatment effect. In column 3, we impute 1 for all

missing treatment values and 0 for all missing control values, which gives an upper bound of the treatment effect. None of these estimates of β are significant at the 5% level, showing that the imputation technique is not driving the results.

3.4.4 Intensive Margin Treatment Effect

Table 3.4 shows the main regression from earlier in column 1, followed by alternative outcome specifications in columns 2 and 3. The outcomes are binary variables where 1 indicates whether that parent has attended at least 2 and 3 sessions, respectively, and zero otherwise. Here, we see that while the treatment may not induce a parent to attend at least 1 session, there is a statistically significant (α=.05 level) impact of the treatment on the likelihood that a parent attends at least 3 events. This shows that while there is not an "extensive" margin treatment effect, there might be an "intensive" margin treatment effect, such that parents who already attend sessions are more likely to attend additional sessions as a result of the treatment.

To explore this further, Table 3.5 shows the treatment effect when the outcome is a continuous variable representing the attendance rate, which is computed as the total number of events a parent attended divided by the total number of events.

Column 1 shows the regression results from the full sample. The attendance rate for the control group is 12.9%, and that of the treatment group is 16.5%, leading to a 3.6 percentage point treatment effect. This is the Intent To Treat (ITT) estimate of being assigned to the treatment group on attendance. This difference is statistically significant at the 10% level, but not the 5% level. Column 2 shows the results of the regression when the sample is limited to parents who

Table 3.4: Treatment Effect on Attending At Least 1, 2, 3 Events

	(1) At Least 1 Event	(2) At Least 2 Events	(3) At Least 3 Events
Treatment	-0.0135	0.0541	0.0741**
	(0.0493)	(0.0479)	(0.0343)
Female	0.0394	0.0451	0.0292
	(0.0498)	(0.0486)	(0.0368)
Spanish	0.0652	0.136**	0.0758
	(0.0648)	(0.0651)	(0.0524)
Child Age (Years)	0.0240	0.0256	0.0120
	(0.0409)	(0.0390)	(0.0252)
Constant	0.562***	0.335*	0.0941
	(0.181)	(0.176)	(0.115)
Observations	319	319	319
School FE	Yes	Yes	Yes
Covariates	Yes	Yes	Yes

Note: Robust Standard Errors are in parenthesis. The outcome in column 1 is a binary variable where 1 indicates that the parent attended at least 1 event, and 0 otherwise. Similarly, columns 2 and 3 are binary variables where 1 indicates that the parent attended at least 2 and 3 events respectively, and 0 otherwise. Female is an indicator for whether the child is female, Spanish is an indicator for whether the household's primary language is Spanish, & Child Age is the child's age in years as of November 2018.
*** p<0.01, ** p<0.05, * p<0.1.

attended at least one event. Because the extensive margin effect is zero in this sample, the group of parents who attend at least one event in the treatment and control are arguably comparable on unobservable characteristics. Any difference in their outcomes could therefore be interpreted as the intensive margin treatment effect under this assumption.

Of the 91 control parents who attend at least one event, the attendance rate

Table 3.5: Treatment Effect (Intensive Margin) on Attendance

	(1) Attendance Rate	(2) Attendance Rate
Treatment	0.0356*	0.0700**
	(0.0195)	(0.0270)
Female	0.0249	0.0326
	(0.0203)	(0.0271)
Spanish	0.0548**	0.0469
	(0.0264)	(0.0345)
Child Age (Years)	0.00862	0.00496
	(0.0142)	(0.0190)
Constant	0.144**	0.241***
	(0.0629)	(0.0817)
Observations	319	180
School FE	Yes	Yes
Covariates	Yes	Yes

Note: Robust Standard Errors are in parenthesis. The outcome is the attendance rate for each parent, computed as the percentage of all 8 (or 7) events attended by that parent. The regression in column 1 includes all observations, whereas the regression in column 2 conditions the regression on the parent attending at least one event. Female is an indicator for whether the child is female, Spanish is an indicator for whether the household's primary language is Spanish, & Child Age is the child's age in years as of November 2018.
*** p<0.01, ** p<0.05, * p<0.1.

across all events is 22%: parents attend an average of 1.78 events. Of the 89 treatment parents who attend at least one event, the attendance rate is 29% across all events: parents attend an average of 2.33 events. Column 2 of Table 5 shows that this 7-percentage-point difference is statistically significant at the 5% level. In relative terms, this is a 32% intensive margin treatment effect on attendance rate. In practical terms, treatment parents attended 0.55 additional events (out of

eight), which corresponds to a 0.37 standard deviation increase. This magnitude of treatment effect is considered large among RCTs in education (Kraft, 2020). Despite this high relative treatment effect, the level of attendance for the treatment group is still low in absolute terms, which we interpret using a theoretical framework described in Section 3.5.

We explore heterogeneity in the treatment effect of the continuous-specified attendance rate. We looked at heterogeneity based on two quantitative features of events that we could observe: event length and event timing. The events lasted between 60 and 120 minutes, with a median length of 90 minutes. The treatment effect on attendance rate was larger for shorter events, but this difference was not statistically significant ($p = 0.334$). For event timing, the median start-time for events was 4:00 PM (Note: school pick-up time for most centers was between 3:30 and 4:30 PM). There was no significant difference in the treatment effect for events that took place in the evening versus during the daytime, nor was the interaction term significant when event start time is treated as a continuous variable.

3.4.5 Incentive Spillover

One potential concern with our treatment is that rather than increasing the total number of events a parent attends, our treatment might simply make parents reprioritize which events to attend. For instance, if a parent in the treatment group was planning to attend exactly three events in total for the year, then our treatment might simply induce them to attend our events, for which they would be paid, instead of other events the school may offer. In that case, our treatment would have a negative *incentive* spillover.

In contrast, attending more events as a result of the treatment might make a parent develop a habit of attending events. As a parent attends more events, they might view the barriers to attending as less costly or the benefits to be gained as greater than previously perceived. Or, perhaps they form friendships with other parents or teachers at the additional events they attend, which increases their expected return on attending events in general. This increased likelihood of habit formation to attend events would be a potentially positive incentive spillover.

We can assess incentive spillovers for 211 parents because four of the six centers offered at least one event beyond the eight used in our study.[6] There were seven such events in total across the four centers. These events took place within the four-month window of our study, after the eight incentivized events. Specifically, for 134 parents (42% of our sample) we have attendance data for at least two events beyond the eight in the experiment. The treatment effect on the attendance rates at these unincentivized events is shown in Table 3.6.

The results show that there was a positive incentive spillover: treatment parents are nearly twice as likely as control parents to attend one of the unincentivized events. Specifically, 14.8% of the treatment parents attended at least one event after the incentives ended, compared to 8.3% of the control parents. This difference is statistically significant at the 10% level but not the 5% level.

One concern might be that the treatment parents might not be aware that the incentives have ended. However, this is unlikely to be the case because treatment parents received weekly text messages with the specific name of each upcoming incentivized event, and a reminder of the incentive for attending that event. Such

6. The treatment effect for these four centers is similar to the treatment effect for the overall sample.

Table 3.6: Treatment Effect on Attendance at Unincentivized Events

	(1) Unincentivized Attendance Rate	(2) Unincentivized Attendance Rate
Treatment	0.0653*	0.0694*
	(0.0359)	(0.0362)
Female		-0.00164
		(0.0369)
Spanish		-0.00124
		(0.0360)
Child Age (Years)		0.0328
		(0.0313)
Constant	0.0825***	-0.0176
	(0.0231)	(0.138)
Observations	211	211
School FE	No	Yes
Covariates	No	Yes

Note: Robust Standard Errors are in parenthesis. The outcome is the attendance rate for each parent at events that were not financially incentivized by the program, after the treatment ended. These regression only include the 211 parents at the 4 centers that had these additional events. Female is an indicator for whether the child is female, Spanish is an indicator for whether the household's primary language is Spanish, & Child Age is the child's age in years as of November 2018.
*** p<0.01, ** p<0.05, * p<0.1.

texts were not sent out for the unincentivized events.

While this sample of seven total events across four centers might not be convincing enough to confidently claim that the treatment causes parents to begin a regular habit of attending events after incentives are removed, the evidence is suggestive. This also shows negative incentive spillovers are likely not a concern: our treatment does not seem to divert attendance away from unincentivized events.

3.5 Theoretical Framework

To better interpret our findings, we next present a model of parental decisions to engage with their child's school.

3.5.1 A Simple Model of Parental Decision Making

Suppose $e_i \in \{0, 1\}$ represents parent i's decision on whether to engage with their child's school (such as attending school-sponsored events), where $e_i = 1$ indicates engaging and $e_i = 0$ is not engaging. Suppose B_k is the parent's perceived benefit from attending event k, and $c_i \in \mathbb{R}^+$ is the parent's opportunity cost of attending event B_k. If event k is an hour long, then c_i would represent that parent's hourly wage. Let $F(c)$ represent the cumulative distribution function for hourly wage (opportunity cost) c. The parent's utility function $u(e_i)$ is given by:

$$u(e_i) = (e_i)(B_k) + (1 - e_i)(c_i)$$

A parent's optimal decision would therefore be to attend ($e_i^* = 1$) the event when $u(1) \geq u(0)$, which happens when $B_k \geq c_i$. This happens with probability $F(B_k)$.

3.5.2 Reminders

One way in which a reminder can influence behavior in this context is to simply inform the parent of something that they were not aware of (such as the event's time and location, or that it is happening at all). Another channel is that even if the parent was aware of the event details, a reminder can bring the event "top

of mind" for the parent. Parents who would optimally choose to attend an event (ones with $c_i < B_k$) might not attend if the event is not on the top of their mind. This might occur if the parent has multiple competing demands on their attention. For this type of parent, a text message reminder can prompt action.

Suppose there are two types of parents: attentive and inattentive. Let this be denoted by $\alpha_i \in \{0, 1\}$, where $\alpha = 1$ denote the 'attentive' type of parent, who is both informed of an event and it is on their top of mind. For this type of parent, a text message reminder would not change anything. Let $\alpha = 0$ denote an inattentive parent, who is either uninformed of the event, or is informed but the event is not at the top of their mind. We will not distinguish among the reasons a parent might be inattentive, and will simply note that a reminder can influence this type of parent if their optimal decision is to attend ($c_i < B_k$). Note that we are ignoring the possibility that a parent might be so inattentive that they miss a reminder message; we are assuming that even the $\alpha = 0$ parent would notice a reminder message in time for the event. Overlaying attention with opportunity cost, our type space for parents is $\tau = (\alpha_i, c_i) \in \{0, 1\} \times \mathbb{R}^+$. We can amend the simple utility function above to incorporate α as follows:

$$u(e_i) = (e_i)(\alpha_i B_k) + (1 - e_i)(c_i)$$

The attentive types have the same utility function as before, and their decision to attend will be dependent on whether $c_i < B_k$. However, an inattentive type would never choose to attend the event, assuming $c_i > 0$.

Let us now define a reminder intervention as one that sets a parent's $\alpha_i = 1$. This means that for already attentive types, nothing changes. Additionally, those

with $c_i > B_k$ will choose not to attend regardless of their attentiveness type, and will therefore also not be affected by this intervention. The only parent that would be prompted to go because of a reminder is one with both $\alpha_i = 0$ and $c_i < B_k$.

Denoting the proportion of attentive parents $\mathbb{P}[\alpha_i = 1] = \theta$, the expected proportion of parents who attend event k without the reminder intervention is $\theta F(B_k)$. The expected proportion with the reminder intervention would be $F(B_k)$, making the expected treatment effect $(1 - \theta)F(B_k)$.

3.5.3 Financial Incentives with Loss Aversion

Now to incorporate financial incentives, suppose that a parent is given m_1 dollars for attending the event, and m_0 for not attending the event. This makes the utility:

$$u(e_i) = (e_i)[\alpha_i(B_k + v(m_1))] + (1 - e_i)[c_i + v(m_0)]$$

Where $v(m)$ is given by the standard reference-dependent *loss aversion* value function:

$$v(m) = \begin{cases} m - r, & m \geq r \\ \gamma(m - r), & m < r \end{cases}$$

Loss aversion is when $\gamma > 1$. Empirical estimates of γ in the literature are around $\gamma \approx 2$, implying that a losses have twice the impact of an equivalent gain (Benartzi and Thaler, 1995).

Consider an intervention that gives m dollars to parents for attending event k. If the money is to be given after the event, we can think of the reference value r as being 0. With $m_1 = m, m_0 = 0, r = 0$, this gives $v(m_1) = m, v(m_0) = 0$,

implying that for an attentive type of parent, the optimal decision is to attend if $B_k + m \geq c_i$.

If instead the parent is given m dollars up front, and then are asked to return the money in case they do not attend the event, this now changes our reference to $r = m$. With $m_1 = m, m_0 = 0, r = m$, this gives $v(m_1) = 0, v(m_0) = -\gamma m$, implying that for an attentive type of parent, the optimal decision is to attend if $B_k + \gamma m \geq c_i$. This implies that a loss averse framing would induce a higher share of parents to attend the event, since it increases the opportunity cost threshold above which a parent would choose to not attend.

The new expected proportion of parents who would attend the event after an intervention that combines reminders with a loss-aversion-framed incentive of m dollars would be $F(B_k + \gamma m)$. This would make the expected treatment effect equal to $F(B_k + \gamma m) - \theta F(B_k)$. Note that the treatment effect is larger when cognitive biases are stronger (the larger the γ and lower the θ).

3.5.4 Interpreting our findings

If all parents behave like rational agents ($\gamma = 1$ and $\theta = 1$), the expected treatment effect is $F(B_k + m) - F(B_k)$. Given the low baseline attendance rate in the control group, $F(B_k)$ is a fairly low number, meaning most parents have an opportunity cost greater than perceived benefit B_k. Because our monetary incentive is approximately equal to the median financial opportunity cost, we would expect that receiving m in addition to the perceived event benefit B_k would induce a majority of parents to attend in a frictionless world. That is, we would expect $F(B_k + m)$ to be greater than 0.50 if there were no structural barriers preventing

parents from attending beyond the opportunity cost of attending. However, we estimate that $F(B_k+m)$ is far below this, because the treatment group's attendance rate was less than 20%, let alone 50%. Note that lack of substantive treatment effect is even more puzzling if parents have cognitive biases ($\gamma > 1$ and $\theta < 1$).

One potential interpretation here is that structural barriers to attendance, such as work conflicts or inflexible obligations, are far too prevalent for interventions such as ours to be able to reasonably change parental engagement. Another interpretation is that parents' perceived benefit of attending an event, B_k, is in fact negative. That is, if parents view attending such events as having a psychic cost and no associated benefit ($B_k < 0$), then this is consistent with observing attendance rates of less than 50% despite parents being offered the median opportunity cost as financial compensation.

Whether it is structural barriers or no perceived value added for events, the implication here may be that financial compensation to increase parental engagement will likely not increase parental attendance to substantial levels unless the compensation is much higher than the median parent's hourly wage (for example, Fryer et al., 2015). Depending on cost constraints, this implies that providing parents with financial incentives is likely not a feasible strategy to increase parental engagement for most preschools.

3.6 Conclusion

We designed an experiment to test whether financial incentives combined with reminders could increase low-income parents' attendance at parent engagement events at their children's preschools. We focus on these events because schools

are mandated to offer them, because prior research shows that lack of attendance is a persistent problem, and because parent-school connections are theoretically relevant for children's skill development. We used financial incentives to test a theory about parents' assessments of the value of their attendance. We chose an amount for a financial incentive that approximated parents' opportunity costs in the labor market. To maximize the impact of the financial incentives we offered them to parents in a loss aversion framework. We used reminders – a common behavioral tool – to mitigate information frictions and parental inattention, that is, to make the information about the school events salient or "top of mind" to parents.

Our results show that the treatment had no impact on parents who did not attend any events, but did increase the attendance rate among parents who attended at least one event. The treatment also increased parents' likelihood of attending unincentivized events, which provides suggestive evidence that our treatment might help already-engaged parents strengthen a habit of engaging. Yet, even among parents who attend at least one event, the attendance rate for treated parents is still far below what many schools aspire to. It would be hard not to conclude from this study and related recent ones that preschools serving disadvantaged children should abandon or wholly reimagine their efforts to induce parents to attend school events. Structural barriers to attendance such as work conflicts may put a ceiling on expected parental engagement even in an ideal scenario.

Parents may also view such events as having a low expected return for their time, and may need to be compensated substantially more than their lost earnings to attend. Schools may have more success in parental attendance by offering events that parents perceive as being worthwhile. As such, future work can randomize

the type of events schools offer to better understand parents' preferences.

REFERENCES

AbdulRaheem, Y., Yusuf, H. T., and Odutayo, A. O. (2017). Effect of peer tutoring on students' academic performance in economics in ilorin south, nigeria. *Journal of Peer Learning*, 10(1):95–102.

Ahn, W.-k., Brewer, W. F., and Mooney, R. J. (1992). Schema acquisition from a single example. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(2):391.

Alsan, M. and Eichmeyer, S. (2021). *Experimental Evidence on the Effectiveness of Non-Experts for Improving Vaccine Demand*. National Bureau of Economic Research.

Ansari, A. and Gershoff, E. (2016). Parent involvement in head start and children's development: Indirect effects through parenting. *Journal of Marriage and Family*, 78(2):562–579.

Arad, A., Gneezy, U., and Mograbi, E. (2023). Intermittent incentives to encourage exercising in the long run. *Journal of Economic Behavior & Organization*, 205:560–573.

Arrow, K. J. (1962). The economic implications of learning by doing. *The review of economic studies*, 29(3):155–173.

Avvisati, F., Besbas, B., and Guyon, N. (2010). Parental involvement in school: A literature review. *Revue d'économie politique*, 120(5):759–778.

Avvisati, F., Gurgand, M., Guyon, N., and Maurin, E. (2014). Getting parents involved: A field experiment in deprived schools. *Review of Economic Studies*, 81(1):57–83.

Barrera-Osorio, F., Gertler, P., Nakajima, N., and Patrinos, H. A. (2020). *Promoting Parental Involvement in Schools: Evidence From Two Randomized Experiments*. The World Bank.

Benartzi, S. and Thaler, R. H. (1995). Myopic loss aversion and the equity premium puzzle. *The quarterly journal of Economics*, 110(1):73–92.

Berlinski, S. and Busso, M. (2017). Challenges in educational reform: An experiment on active learning in mathematics. *Economics Letters*, 156:172–175.

Boaler, J., Brown, K., LaMar, T., Leshin, M., and Selbach-Allen, M. (2022). Infusing mindset through mathematical problem solving and collaboration: Studying the impact of a short college intervention. *Education Sciences*, 12(10):694.

Bonwell, C. C. and Eison, J. A. (1991). *Active learning: Creating excitement in the classroom. 1991 ASHE-ERIC higher education reports.* ERIC.

Bordalo, P., Coffman, K., Gennaioli, N., Schwerter, F., and Shleifer, A. (2021). Memory and representativeness. *Psychological Review*, 128(1):71.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2020). Memory, attention, and choice. *The Quarterly journal of economics*, 135(3):1399–1442.

Brownback, A. and Sadoff, S. (2020). Improving college instruction through incentives. *Journal of Political Economy*, 128(8):2925–2972.

Bulte, E., List, J. A., and Van Soest, D. (2020). Toward an understanding of the welfare effects of nudges: Evidence from a field experiment in the workplace. *The Economic Journal*, 130(632):2329–2353.

Bulte, E., List, J. A., and van Soest, D. (2021). Incentive spillovers in the workplace: Evidence from two field experiments. *Journal of Economic Behavior & Organization*, 184:137–149.

Burgess, S., Metcalfe, R., and Sadoff, S. (2021). Understanding the response to financial and non-financial incentives in education: Field experimental evidence using high-stakes assessments. *Economics of Education Review*, 85:102195.

Bursztyn, L., Egorov, G., and Jensen, R. (2019). Cool to be smart or smart to be cool? understanding peer pressure in education. *The Review of Economic Studies*, 86(4):1487–1526.

Castleman, B. L. and Page, L. C. (2017). Parental influences on postsecondary decision making: Evidence from a text messaging experiment. *Educational Evaluation and Policy Analysis*, 39(2):361–377.

Castro, M., Expósito-Casas, E., López-Martín, E., Lizasoain, L., Navarro-Asencio, E., and Gaviria, J. L. (2015). Parental involvement on student academic achievement: A meta-analysis. *Educational research review*, 14:33–46.

Conlon, J. J., Mani, M., Rao, G., Ridley, M. W., and Schilbach, F. (2022). Not learning from others. Technical report, National Bureau of Economic Research.

Cortes, K., Kortecamp, K., Loeb, S., and Robinson, C. (2024). A scalable approach to high-impact tutoring for young readers: Results of a randomized controlled trial. Technical report, National Bureau of Economic Research.

Cotton, C., Hickman, B. R., List, J. A., Price, J., and Roy, S. (2020). Disentangling motivation and study productivity as drivers of adolescent human capital formation: Evidence from a field experiment and structural analysis. Technical report, National Bureau of Economic Research.

Cunha, F., Heckman, J. J., Lochner, L., and Masterov, D. V. (2006). Interpreting the evidence on life cycle skill formation. *Handbook of the Economics of Education*, 1:697–812.

Davis, J. M. and Heller, S. B. (2017). Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review*, 107(5):546–550.

De Backer, L., Van Keer, H., and Valcke, M. (2012). Exploring the potential impact of reciprocal peer tutoring on higher education students' metacognitive knowledge and regulation. *Instructional science*, 40(3):559–588.

DellaVigna, S., List, J. A., Malmendier, U., and Rao, G. (2022). Estimating social preferences and gift exchange at work. *American Economic Review*, 112(3):1038–1074.

Dizon-Ross, R. (2019). Parents' beliefs about their children's academic ability: Implications for educational investments. *American Economic Review*, 109(8):2728–2765.

Domina, T. (2005). Leveling the home advantage: Assessing the effectiveness of parental involvement in elementary school. *Sociology of education*, 78(3):233–249.

Dropkin, E. and Jauregui, S. (2015). Two generations together: Case studies from head start. *Washington, DC: National Head Start Association.*

Eskreis-Winkler, L., Milkman, K. L., Gromet, D. M., and Duckworth, A. L. (2019). A large-scale field experiment shows giving advice improves academic outcomes for the advisor. *Proceedings of the national academy of sciences*, 116(30):14808–14810.

Fiorella, L. and Mayer, R. E. (2013). The relative benefits of learning by teaching and teaching expectancy. *Contemporary Educational Psychology*, 38(4):281–288.

Fiorini, M. and Keane, M. P. (2014). How the allocation of children's time affects cognitive and noncognitive development. *Journal of Labor Economics*, 32(4):787–836.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., and Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the national academy of sciences*, 111(23):8410–8415.

Fryer, R. G. (2017). The production of human capital in developed countries: Evidence from 196 randomized field experiments. In *Handbook of economic field experiments*, volume 2, pages 95–322. Elsevier.

Fryer, R. G., Levitt, S. D., List, J., and Sadoff, S. (2022). Enhancing the efficacy of teacher incentives through framing: A field experiment. *American Economic Journal: Economic Policy*, 14(4):269–299.

Fryer, R. G., Levitt, S. D., List, J. A., et al. (2015). Parental incentives and early childhood achievement: A field experiment in chicago heights. Technical report, National Bureau of Economic Research.

Fuchs, L. S. and Malone, A. S. (2021). Can teaching fractions improve teachers' fraction understanding? insights from a causal-comparative study. *the elementary school journal*, 121(4):656–673.

Gennetian, L. A., Marti, M., Kennedy, J. L., Kim, J. H., and Duch, H. (2019). Supporting parent engagement in a school readiness program: Experimental evidence applying insights from behavioral economics. *Journal of Applied Developmental Psychology*, 62:1–10.

Gneezy, U. and List, J. A. (2013). *The why axis: Hidden motives and the undiscovered economics of everyday life*. Random House.

Gneezy, U., List, J. A., Livingston, J. A., Qin, X., Sadoff, S., and Xu, Y. (2019). Measuring success in education: The role of effort on the test itself. *American Economic Review: Insights*, 1(3):291–308.

Greene, I., Tiernan, A. M., and Holloway, J. (2018). Cross-age peer tutoring and fluency-based instruction to achieve fluency with mathematics computation skills: A randomized controlled trial. *Journal of Behavioral Education*, 27(2):145–171.

Guerrero, T. A. and Wiley, J. (2021). Expecting to teach affects learning during study of expository texts. *Journal of Educational Psychology*.

Guryan, J., Ludwig, J., Bhatt, M. P., Cook, P. J., Davis, J. M., Dodge, K., Farkas, G., Fryer Jr, R. G., Mayer, S., Pollack, H., et al. (2023). Not too late: Improving academic outcomes among adolescents. *American Economic Review*, 113(3):738–765.

Harmon, C. (2004). Was head start a community action program? another look at an old debate. *The Head Start debates*, pages 85–101.

Heckman, J. J. and Masterov, D. V. (2007). The productivity argument for investing in young children.

Hill, N. E. and Tyson, D. F. (2009). Parental involvement in middle school: a meta-analytic assessment of the strategies that promote achievement. *Developmental psychology*, 45(3):740.

Hill, Z., Spiegel, M., Gennetian, L., Hamer, K.-A., Brotman, L., and Dawson-McClure, S. (2021). Behavioral economics and parent participation in an evidence-based parenting program at scale. *Prevention Science*, pages 1–12.

Imas, A., Sadoff, S., and Samek, A. (2017). Do people anticipate loss aversion? *Management Science*, 63(5):1271–1284.

Kahneman, D., Knetsch, J. L., and Thaler, R. H. (1991). Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic perspectives*, 5(1):193–206.

Kalil, A., Liu, H., Mayer, S., Rury, D., and Shah, R. (2023a). Nudging or nagging? conflicting effects of behavioral tools. *Conflicting Effects of Behavioral Tools (January 4, 2023). University of Chicago, Becker Friedman Institute for Economics Working Paper*, (2023-02).

Kalil, A., Liu, H., Mayer, S., Rury, D., and Shah, R. (2023b). Nudging or nagging? conflicting effects of behavioral tools. *University of Chicago, Becker Friedman Institute for Economics Working Paper 2023-02*.

Kalil, A., Mayer, S., Oreopoulos, P., and Shah, R. (2024). Making a song and dance about it: The effectiveness of teaching children vocabulary with animated music videos. Technical report, National Bureau of Economic Research.

Kalil, A., Mayer, S., and Shah, R. (2020). Impact of the covid-19 crisis on family dynamics in economically vulnerable households. *University of Chicago, Becker Friedman Institute for Economics Working Paper 2020-143*.

Kalil, A., Mayer, S., and Shah, R. (2023c). Scarcity and inattention. *Journal of Behavioral Economics for Policy*, 7(1):35–42.

Kalil, A., Mayer, S. E., and Gallegos, S. (2021). Using behavioral insights to increase attendance at subsidized preschool programs: The show up to grow up intervention. *Organizational Behavior and Human Decision Processes*, 163:65–79.

King, A., Staffieri, A., and Adelgais, A. (1998). Mutual peer tutoring: Effects of structuring tutorial interaction to scaffold peer learning. *Journal of Educational Psychology*, 90(1):134.

Kline, P. and Walters, C. R. (2016). Evaluating public programs with close substitutes: The case of head start. *The Quarterly Journal of Economics*, 131(4):1795–1848.

Kobayashi, K. (2022). Learning by teaching face-to-face: the contributions of preparing-to-teach, initial-explanation, and interaction phases. *European Journal of Psychology of Education*, 37(2):551–566.

Kraft, M., List, J., Livingston, J., and Sadoff, S. (2022). Online tutoring by college volunteers: Experimental evidence from a pilot program. In *AEA Papers and Proceedings*.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational researcher*, 49(4):241–253.

Kraft, M. A. and Falken, G. T. (2021). A blueprint for scaling tutoring and mentoring across public schools. *AERA Open*, 7(1):1–21.

Kraft, M. A. and Lovison, V. S. (2024). The effect of student-tutor ratios: Experimental evidence from a pilot online math tutoring program. edworkingpaper no. 24-976. *Annenberg Institute for School Reform at Brown University*.

Kraft, M. A. and Rogers, T. (2015). The underutilized potential of teacher-to-parent communication: Evidence from a field experiment. *Economics of Education Review*, 47:49–63.

Kraft, M. A., Schueler, B. E., and Falken, G. (2024). What impacts should we expect from tutoring at scale? exploring meta-analytic generalizability.

Lazear, E. P. (2001). Educational production. *The Quarterly Journal of Economics*, 116(3):777–803.

Levitt, S. D., List, J. A., Neckermann, S., and Sadoff, S. (2016). The behavioralist goes to school: Leveraging behavioral economics to improve educational performance. *American Economic Journal: Economic Policy*, 8(4):183–219.

List, J. A. (2022). *The voltage effect: How to make good ideas great and great ideas scale*. Crown Currency.

List, J. A. (2024a). *Experimental Economics: Theory and Practice*. University of Chicago Press.

List, J. A. (2024b). Optimally generate policy-based evidence before scaling. *Nature*, 626(7999):491–499.

List, J. A., Samek, A., and Suskind, D. L. (2018). Combining behavioral economics and field experiments to reimagine early childhood education. *Behavioural Public Policy*, 2(1):1–21.

List, J. A. and Shah, R. (2022). The impact of team incentives on performance in graduate school: Evidence from two pilot rcts. *Economics Letters*, 221:110894.

Liu, Z. and White, M. J. (2017). Education outcomes of immigrant youth: The role of parental engagement. *The Annals of the American Academy of Political and Social Science*, 674(1):27–58.

Markant, D., Ruggeri, A., Gureckis, T., and Xu, F. (2016). Enhanced memory as a common effect of active learning. mind, brain, and education, 10 (3), 142-152.

Marti, M., Merz, E. C., Repka, K. R., Landers, C., Noble, K. G., and Duch, H. (2018). Parent involvement in the getting ready for school intervention is associated with changes in school readiness skills. *Frontiers in psychology*, 9:759.

Mayer, S., Shah, R., and Kalil, A. (2021). How cognitive biases can undermine program scale-up decisions. In *The Scale-Up Effect in Early Childhood and Public Policy*, pages 41–57. Routledge.

Mayer, S. E., Kalil, A., Delgado, W., Liu, H., Rury, D., and Shah, R. (2023). Boosting parent-child math engagement and preschool children's math skills: Evidence from an rct with low-income families. *Economics of Education Review*, 95:102436.

Mayer, S. E., Kalil, A., Oreopoulos, P., and Gallegos, S. (2019). Using behavioral insights to increase parental engagement the parents and children together intervention. *Journal of Human Resources*, 54(4):900–925.

McNeal Jr, R. B. (2012). Checking in or checking out? investigating the parent involvement reactive hypothesis. *The Journal of Educational Research*, 105(2):79–89.

McQuiggan, M. and Megra, M. (2017). Parent and family involvement in education: Results from the national household education surveys program of 2016. first look. nces 2017-102. *National Center for Education Statistics*.

Mendez, J. L. (2010). How can parents get involved in preschool? barriers and engagement in education by ethnic minority parents of children attending head start. *Cultural Diversity and Ethnic Minority Psychology*, 16(1):26.

Mitchell, R. J., Morrison, T. G., Feinauer, E., Wilcox, B., and Black, S. (2016). Effects of fourth and second graders' cross-age tutoring on students' spelling. *Reading Psychology*, 37(1):147–166.

Nickow, A., Oreopoulos, P., and Quan, V. (2020). The impressive effects of tutoring on prek-12 learning: A systematic review and meta-analysis of the experimental evidence. working paper 27476. *National Bureau of Economic Research*.

Pierce, L., Rees-Jones, A., and Blank, C. (2020). The negative consequences of loss-framed performance incentives. Technical report, National Bureau of Economic Research.

Robinson, C. D., Bisht, B., and Loeb, S. (2022). The inequity of opt-in educational resources and an intervention to increase equitable access. *Annenberg Institute at Brown University EdWorkingPaper*, (22-654).

Robinson, C. D., Pollard, C., Novicoff, S., White, S., and Loeb, S. (2024). The effects of virtual tutoring on young readers: Results from a randomized controlled trial. *EdWorkingPapers. com*.

Romero, M., Chen, L., and Magari, N. (2022). Cross-age tutoring: Experimental evidence from kenya. *Economic Development and Cultural Change*, 70(3):000–000.

Ross, M. and Bateman, N. (2019). Meet the low-wage workforce. *The Brookings Institute*.

Rozenblit, L. and Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive science*, 26(5):521–562.

Sadoff, S. (2014). The role of experimentation in education policy. *Oxford Review of Economic Policy*, 30(4):597–620.

Samat, C., Chaijaroen, S., and Wattanachai, S. (2019). The designing of constructivist web-based learning environment to enhance problem solving process and transfer of learning for computer education student. In *Innovative Technologies and Learning: Second International Conference, ICITL 2019, Tromsø, Norway, December 2–5, 2019, Proceedings 2*, pages 117–126. Springer.

Shah, R., Kalil, A., and Mayer, S. (2023). Engaging parents with preschools: Evidence from a field experiment. *University of Chicago, Becker Friedman Institute for Economics Working Paper*, (2022-97).

Turney, K. and Kao, G. (2009). Barriers to school involvement: Are immigrant parents disadvantaged? *The Journal of Educational Research*, 102(4):257–271.

Villena-Roldan, B. and Ríos-Aguilar, C. (2012). Causal effects of maternal time-investment on children's cognitive outcomes. Working Paper 285, Center for Applied Economics.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Wang, M.-T. and Sheikh-Khalil, S. (2014). Does parental involvement matter for student achievement and mental health in high school? *Child development*, 85(2):610–625.

Yeager, D. S. and Dweck, C. S. (2012). Mindsets that promote resilience: When students believe that personal characteristics can be developed. *Educational psychologist*, 47(4):302–314.

York, B. N., Loeb, S., and Doss, C. (2019). One step at a time: The effects of an early literacy text-messaging program for parents of preschoolers. *Journal of Human Resources*, 54(3):537–566.

Zigler, E. and Styfco, S. J. (2010). *The hidden history of Head Start.* Oxford University Press.

Zigler, E. F. and Muenchow, S. (1992). *Head Start: The inside story of America's most successful educational experiment.* Basic Books.