THE TRADE-OFF BETWEEN QUALITY AND QUANTITY: EVIDENCE FROM A FIELD EXPERIMENT ON TUTORING

By

Rohen Shah

January 2026

COWLES FOUNDATION DISCUSSION PAPER NO. 2493

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS

YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

The Trade-off between Quality and Quantity: Evidence from a Field Experiment on Tutoring*

Rohen Shah†

January 14, 2026

Abstract

High-dosage tutoring has the potential to substantially raise adolescent academic achievement. However, at scale, schools may not have the financial ability to deliver small-group tutoring frequently. In this paper, I test the relative importance of group size (quality) versus tutoring frequency (quantity). I evaluate the impact of an in-school math tutoring program in a middle school in the Midwestern United States. Students are randomized to either 1) control, 2) receive tutoring twice a week in 2-student groups, or 3) receive tutoring 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.

JEL Classifications: H75, I20, J24

Keywords: Field Experiments, Program Evaluation, Tutoring

*This study was pre-registered on the AEA RCT Registry (AEARCTR-0012858) and overseen by the Social and Behavioral Sciences IRB from the University of Chicago (IRB24-0128). I am grateful for support and partnership from KIPP Indy and Accelerate, and would like to thank Jessica Schmaltz, Casey McLeod, Matthew Steinberg, and Jennifer Bronson. I would also like to thank Jens Ludwig, Ariel Kalil, John List, and participants at the 2026 AEA Annual Meeting for their helpful feedback.

Link to Replication Files: https://sites.google.com/view/rohen-shah/projects/kipp

†Yale University, Department of Economics. Email: rohen.shah@yale.edu.

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?

When scaling up programs, such as tutoring, it is important to not just "A/B" test, but to also choose treatment arms (Option "C" thinking) that are more likely to be implemented in practice (List, 2024). Given that a budget-constrained context is likely to be encountered at scale, it is important to generate "policy-based" evidence (List, 2022). In this paper, I use an RCT to test which budget-constrained version of a tutoring program, larger groups or reduced frequency, has a higher ROI.

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 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

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

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-effects, as well as concerns relating to attrition and multiple hypothesis testing (List, 2025).

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 (Kalil et al., 2024, 2023; Mayer et al., 2023; Shah et al., 2025; 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. The findings in this paper are consistent with other work showing that reducing tutoring group size can improve outcomes (Carlana and La Ferrara, 2025; Clarke et al., 2023; Kraft and Lovison, 2024; Robinson et al., 2024). However, these studies show that the more expensive intervention (smaller group size) is better. It may be the case that the smaller group size is no longer more effective when delivered less frequently so as to equalize costs, 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).

2 Experimental Design

2.1 Institutional Setting and Sample

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 was done in school, and students were pulled from elective classes during the school day to receive tutoring, mitigating the need to stay before or after school. The tutors were primarily college students, and received training from the school before beginning tutoring. 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. In the semester prior to the implementation of this study, all students received one 30-minute math tutoring session per week, in a 2-student group.

2.2 Randomization

A total of 12 classrooms – four in each of 6th, 7th, and 8th 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.

Of the twelve classrooms, six classrooms (two in each grade level) were considered by the school as being "Near Proficient" based on proximity to crossing the proficiency threshold on baseline MAP math scores. To accommodate the school's intention to target the tutoring to students in the Near Proficient classrooms, students in these classrooms had a higher probability of receiving treatment. In particular, approximately 60% of students in a Near Proficient classroom received tutoring, while approximately 25% of students in a "Not Near Proficient" classroom received tutoring. Of the 149 students who received tutoring, 104 (70%) were from a Near Proficient classroom, and 45 (30%) were from a Not Near Proficient classroom. Importantly, every classroom had at least some students in each treatment arm, and I include classroom fixed effects in the analysis. The full set of randomization and analysis protocols is described in the Pre-Analysis Plan[2].

2.3 Treatment

The treatment took took place from January to May of 2024. Of the 343 students in the school, 149 were randomly assigned to receive math tutoring. Of these, 62 were randomly assigned to receive tutoring in 2-student groups twice per week, and 87 received tutoring in 3-student groups thrice per week. Tutors received $40 for each session, regardless of the number of students in that session. This makes the total cost per student $40 per week, regardless of whether they received tutoring twice per week in a 2-student group, or thrice per week in a 3-student group.

Students were pulled from an elective class during the school day to receive the math tutoring. The program administrators varied the classes that students were pulled from so that students did not miss too much of any one class. A total of 7 tutors were hired through a local tutoring company. The tutors were college students, and the school provided training before the tutoring began, in addition to the training the tutors received from the tutoring company. Tutors took attendance at the tutoring sessions each week.

[2] https://www.socialscienceregistry.org/trials/12858

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]

The school tests the students 3 times per academic year (September, January, and May). For this study, I use the January 2024 test results as the baseline score, and the May 2024 test results as the endline variable.

3 Descriptive Results

3.1 Summary Statistics

The available administrative data for each student included the student's gender, race, age, and baseline performance (measured in January 2024) on the MAP math assessment. Table 1 shows the descriptive statistics, showcasing that there is no significant difference overall between the control and treatment groups across any of the relevant variables. Approximately 13% of students dropped out of the study by either transferring schools during the semester or by not being present during the assessment in May. This attrition rate is not significantly different across treatment conditions. There is also no significant difference between the 2:1 and 3:1 tutoring conditions for any of these available baseline variables.

[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

Table 1: Descriptive Statistics and Balance Tests

	Control (N=194)	**2:1 Tutoring** (N=62)	**3:1 Tutoring** (N=87)
Attrition	0.129 (SD: 0.336)	−0.024 (0.044) $p = 0.581$	0.001 (0.044) $p = 0.976$
Female	0.469 (SD: 0.500)	−0.084 (0.075) $p = 0.266$	−0.27 (0.066) $p = 0.678$
Age	13.2 (SD: 1.03)	0.001 (0.071) $p = 0.986$	0.039 (0.068) $p = 0.569$
Black	0.784 (SD: 0.413)	−0.029 (0.068) $p = 0.673$	0.006 (0.057) $p = 0.912$
Hispanic	0.170 (SD: 0.377)	0.036 (0.060 $p = 0.549$	0.004 (0.050) $p = 0.942$
Baseline Score	206.3 (SD: 16.5)	2.53 (1.82) $p = 0.164$	3.34** (1.70) $p = 0.050$

Note: The first column displays the mean for the control group for each variable. The second and third columns display the regression coefficient showing the difference between the control group and each treatment group with classroom fixed effects included. Robust standard errors are in parentheses below each coefficient. Attrition is a binary variable with a value of 1 if the student dropped out of the study by either transferring out of the school or being absent during the endline assessment.
*** p<0.01, ** p<0.05, * p<0.1.

3.2 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.

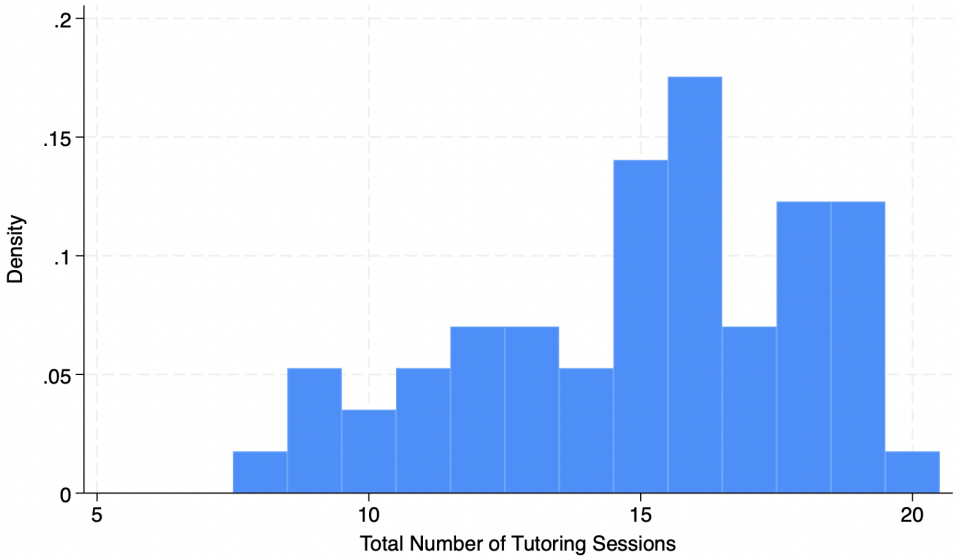

Figure 1: Histogram for Total number of sessions for 2-student groups

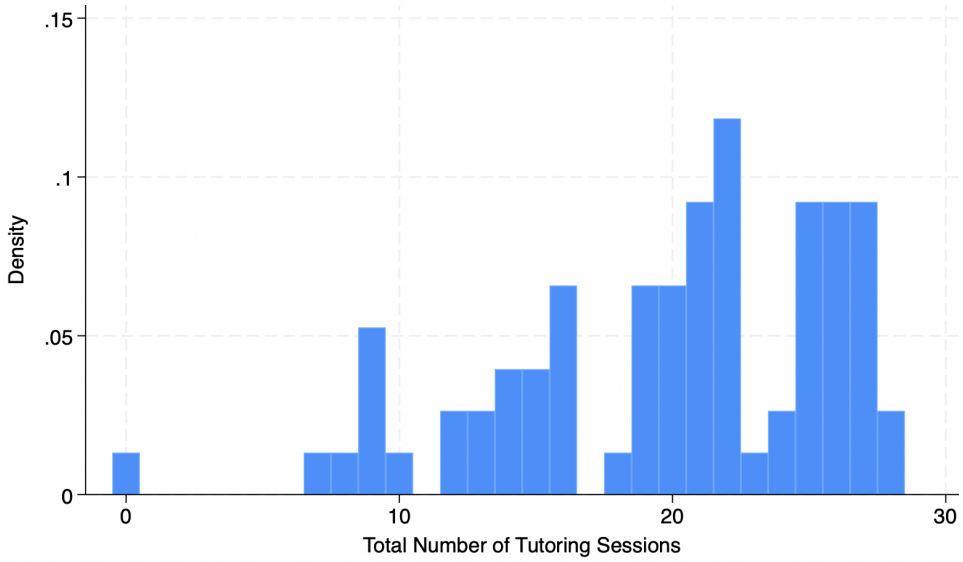

Figure 2: Histogram for Total number of sessions for 3-student groups

The histograms above in Figures 1 and 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.

4 Experimental Findings

4.1 Main Treatment Effects

The following pre-registered regression was used to estimate the ITT (Intent-to-Treat) 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 classroom fixed effect for classroom j, and ε_i is the error term.

Here, we see in Table 2 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. An exploratory analysis of heterogeneity shows that there is no significant difference in treatment effect by race, gender, or baseline math skills.

Additionally, the difference in treatment effect between the 2-student and 3-student group treatments is statistically significant ($p = 0.011$). The difference retains its statistical significance even after a conservative (Bonferroni) correction for multiple hypothesis testing, where the p-value is multiplied by the number of hypotheses (List, 2025). In this case, multiplying the p-values by 3 (for the hypotheses that the two treatments are each different from control, and different from each other) does not alter the overall finding that the 2-student group tutoring leads to a statistically significant improved math performance relative to both the control group and 3-student group tutoring.

Table 2: 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)
	$p = 0.004$	$p = 0.005$
3-Student Group Treatment	0.421	0.582
	(1.076)	(1.022)
	$p = 0.696$	$p = 0.569$
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)
Observations	302	302
Classroom Fixed Effects	Yes	Yes

Note: Robust standard errors in parentheses. 2-Student and 3-Student Group Treatments are indicators with a value of 1 for students assigned to that treatment condition. The Baseline RIT Score is the student's January 2024 MAP Math score. Female is a binary variable with a value of 1 for students who are listed as female. Black and Hispanic are binary variables indicating the student's race, with White being the omitted reference group. Age is the student's age in years as of January 2024. *** p<0.01, ** p<0.05, * p<0.1.

4.2 Alternative Outcome Specification

Table 3 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

Table 3: 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)	
Observations	302	302
Classroom Fixed Effects	Yes	Yes

Note: Robust standard errors in parentheses. The first column shows the treatment effect with the May MAP percentile as the outcome, controlling for the January MAP percentile. The second column shows the treatment effect where the outcome is the growth for each student, computed as the student's May MAP score minus their January MAP score.
*** p<0.01, ** p<0.05, * p<0.1.

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.

4.3 Tutor Fixed-Effect

One possible explanation for these results is that the 2-student group tutoring treatment had better tutors than the 3-student group tutoring treatment arm. We can explore whether this might be the case because a majority of tutors tutored students in both treatment arms, allowing us to include tutor fixed effects.

Table 4 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. The remaining sample was underpowered, but the magnitudes of the treatment effect are consistent in showing that the 2-student group tutoring leads to a 0.2

SD point estimate improvement in math skills relative to the 3-student group tutoring.

Table 4: 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)
Observations	89	89
Classroom Fixed Effects	Yes	Yes
Tutor Fixed Effects	Yes	Yes

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

4.4 Attrition

There were 41 students who dropped out of the study, either because they switched schools or were absent during the May MAP assessment. These students have an average baseline math score that is approximately 3 points, or 0.2 standard deviations, below that of the 301 students for whom the outcome is available (the p value of this difference is 0.116). When looking at the attritors and non-attritors separately, there is no significant difference in the baseline math performance by treatment status (Ghanem et al., 2023).

Following the Pre-Analysis Plan, Table 5 shows the main results, with the same regression specification as that of Table 2, but with the multiple imputation for the missing values, based on the available covariates and the baseline math scores. The overall results are unchanged, namely that the 2-student group tutoring treatment improves math skills by approximately 0.2 standard deviations, with no significant impact for the 3-student group tutoring.

Table 5: Treatment Effect after Imputing Missing Values

	(1) RIT Score	(2) RIT Score
2-Student Group Treatment	3.526***	3.462***
	(1.207)	(1.205)
	$p = 0.004$	$p = 0.004$
3-Student Group Treatment	0.317	0.470
	(0.943)	(0.891)
	$p = 0.737$	$p = 0.598$
Baseline RIT Score	0.821***	0.814***
	(0.0367)	(0.0374)
Female		-0.801
		(0.822)
Black		-3.387*
		(1.936)
Hispanic		-2.167
		(2.123)
Age		-3.124***
		(0.770)
Observations	343	343
Classroom Fixed Effects	Yes	Yes

Note: Robust standard errors in parentheses. 2-Student and 3-Student Group Treatments are indicators with a value of 1 for students assigned to that treatment condition. The Baseline RIT Score is the student's January 2024 MAP Math score. Female is a binary variable with a value of 1 for students who are listed as female. Black and Hispanic are binary variables indicating the student's race, with White being the omitted reference group. Age is the student's age in years as of January 2024. *** p<0.01, ** p<0.05, * p<0.1.

5 Conclusion

This study demonstrates an example of *Option C* thinking (List, 2024) in practice. The findings 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.

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. Given that less-experienced tutors are more likely to be used as tutoring programs scale up, the results from this experiment are more likely to be policy-relevant than one with only highly-skilled tutors.

References

Carlana, M. and La Ferrara, E. (2025). Apart but connected: Online tutoring, cognitive outcomes, and soft skills. *American Economic Review*, 115(10):3487–3513.

Clarke, B., Doabler, C. T., Sutherland, M., Kosty, D., Turtura, J., and Smolkowski, K. (2023). Examining the impact of a first grade whole number intervention by group size. *Journal of Research on Educational Effectiveness*, 16(2):326–349.

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.

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.

Ghanem, D., Hirshleifer, S., and Ortiz-Becerra, K. (2023). Testing attrition bias in field experiments. *Journal of Human resources*.

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.

Kalil, A., Liu, H., Mayer, S., Rury, D., and Shah, R. (2023). 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., 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.

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. 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., 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.

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

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

List, J. A. (2025). *Experimental Economics: Theory and Practice*. University of Chicago Press.

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.

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.

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*.

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. *Educational Evaluation and Policy Analysis*, page 01623737241288845.

Shah, R., Kalil, A., and Mayer, S. (2025). Engaging parents with preschools: Evidence from a field experiment. *Applied Economics*.

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.