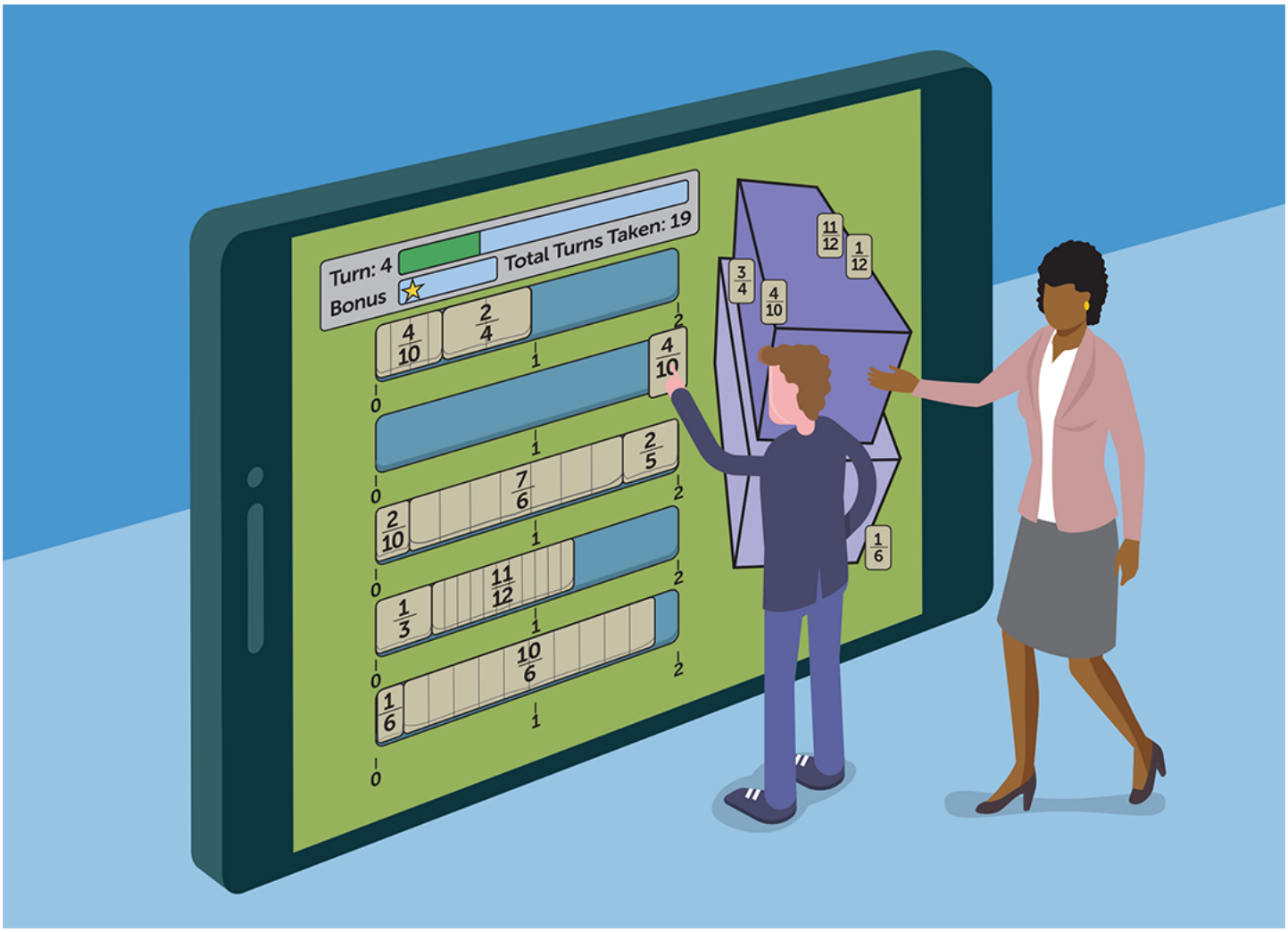
Evaluation of an Online Tutoring Program in Elementary Mathematics

Jeremy Roschelle, Britte Haugan Cheng, Nicola Hodkowski, Julie Neisler and Lina Haldar[1]
April 29, 2020



Digital Promise
Accelerating Innovation in Education

menloEDU

[1] Authors are Digital Promise staff except Britte Haugan Cheng (MenloEDU) & Lina Haldar (LCHaldar Consulting)

Suggested Citation

Roschelle, J., Cheng, B. H., Hodkowski, N., Neisler, J. & Haldar, L. (2020). Evaluation of an online tutoring program in elementary mathematics [Project Report]. San Mateo, CA: Digital Promise. Retrieved from: http://hdl.handle.net/20.500.12265/95

Acknowledgements

This research was supported by grants from the Bill & Melinda Gates Foundation, the Chan-Zuckerberg Initiative, and Schmidt Futures, under a subcontract from Cignition. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funders. We thank Elizabeth Tipton, Associate Professor in Statistics at Northwestern University, for graciously providing meta-analytic advice. The Digital Promise and MenloEDU team thanks the Cignition team for their collaboration in conducting this research, expert panelists, and all project contributors.

Contact Information

Email: jroschelle@digitalpromise.org

Digital Promise:

Washington, DC:
1001 Connecticut Avenue NW, Suite 935
Washington, DC 20036

San Mateo, CA:
2955 Campus Dr. Suite 110
San Mateo, CA 94403

Website: https://digitalpromise.org/

Executive Summary

Many students struggle with mathematics in late elementary school, particularly on the topic of fractions. In best evidence syntheses of research on increasing achievement in elementary school mathematics, Pelligrini et al. (2018) highlighted tutoring as a way to help students. Online tutoring is attractive because costs may be lower and logistics easier than with face-to-face tutoring. Cignition developed an approach that combines online 1:1 tutoring with a fractions game, called "FogStone Isle." The game provides students with additional learning opportunities and provides tutors with information they can use to plan tutoring sessions.

A randomized controlled trial investigated the research question: Do students who participate in online tutoring and a related mathematical game learn more about fractions than students who only have access to the game? Participants were 144 students from four schools, all serving low-income students with low prior mathematics achievement. In the Treatment condition, students received 20- to 25-minute tutoring sessions twice per week for an average of 18 sessions and also played the game. In the Control condition, students had access to the game, but did not play it often. Control students did not receive tutoring. Students were randomly assigned to a condition after being matched on pretest scores. The same diagnostic assessment was used as a pretest and as a posttest. The planned analysis looked for differences in gain scores (posttest minus pretest scores) between conditions. We conducted a t-test on the aggregate gain scores, comparing conditions; the results were statistically significant ($t = 4.0545$, $df = 132.66$, p-value $< .001$). To determine an effect size, we treated each site as a study in a meta-analysis. Using gain scores, the effect size was $g = +.66$. A more sophisticated treatment of the pooled standard deviation resulted in a corrected effect size of $g = .46$ with a 95% confidence interval of [+.23, +.70].

Students who received online tutoring and played the related game learned more; our research found the approach to be efficacious. The Pelligrini et al. (2018) meta-analysis of elementary math tutoring programs found $g = .26$ and was based largely on face-to-face tutoring studies. Thus, this study compares favorably to prior research on face-to-face mathematics tutoring with elementary students. Limitations are discussed; in particular, this is an initial study of an intervention under development. Effects could increase or decrease as development continues and the program scales. Although this study was planned long before the current pandemic, results are particularly timely now that many students are at home under shelter-in-place orders due to COVID-19. The approach taken here is feasible for students at home, with tutors supporting them from a distance. It also could work in many other situations where equity could be addressed directly by supporting students via online tutors.

Introduction

Many students struggle with mathematics in late elementary school, particularly on the topic of fractions (Behr et al., 1984; Streefland, 1991). The topic of fractions is important mathematically; it's part of an important strand of reasoning about rational numbers that develops over several years (Moss, 2005; Saxe et al, 2013; Siegler, Thompson, & Schneider, 2011). Further, weak performance in fractions predicts weak performance in Algebra and other more advanced mathematics (Booth, Newton, & Twiss-Garrity, 2014; Empson et al., 2011; Hackenberg, 2013; Thompson & Saldanha, 2003). The topic of fractions is one where concepts and procedures are both essential; without both, students' ability to solve problems with fractions is weak (National Research Council, 2001; Rittle-Johnson & Koedinger, 2009).

What might help struggling students, particularly those who are most vulnerable? In best evidence syntheses of research on increasing achievement in elementary school mathematics, Pelligrini et al. (2018) highlighted tutoring as a way to help students. Specifically, "particularly positive outcomes were found for tutoring programs" (p. 1). With one exception, Pelligrini et al. (2018) summarized studies in the literature were about face-to-face tutoring; only one online tutoring study was found. The synthesis reported "overall, the weighted mean effect size for one-to-one face-to-face tutoring was +0.26 ($k = 6$, $p < .001$), while the one-to-one online tutoring program had an effect size of -0.03." (p. 17) The online tutoring approach used tutors in India and Sri Lanka for students in England; it is possible that cultural, communication or curricular differences between countries made tutoring less effective.

Such results lend to further exploring online tutoring for elementary math students as worthwhile for two reasons. First, the costs may be lower and the logistics simpler for online tutoring compared to face-to-face tutoring because travel is not required. Second, the supply of highly qualified tutors may be in one geographic location, while demand for tutoring may be in another region. Online tutoring could bring talented tutors into settings where qualified tutors are not otherwise readily available. By matching tutors to students thoughtfully, it would be possible to allocate online tutoring in ways that address equity. One example would be to match students with greater need first, potentially on the basis of a diagnostic pretest.

We report on a rigorous evaluation of an online tutoring program that supported students of color and low-income students to learn challenging concepts related to fractions. Online tutoring sessions occurred twice a week for about 10 weeks and were about 25 minutes long. Students also used a related online fractions learning game. Unlike the one rigorous study in the literature of online tutoring, the tutors and students in this study all resided in the United States or Canada.

This is a preliminary report; we expect to prepare a thorough research report and would therein include a more detailed literature review as well as more complete analysis.

Research Design

The main research question was:

Do students who participate in online tutoring and a related mathematical game learn more about fractions than students who only have access to the game?

Population. Fifth-grade students were recruited from four school sites, with two groups in Central and one group in each of the other sites; hereafter we describe this as five sites. School names have been changed. The schools served populations where a majority of the students were Latinx. A majority of students were receiving free and reduced-price lunch. Over 40% were classified as English Language Learners. Students were nominated by the teachers on the basis of needing additional support. We recruited 148 students.

Intervention. In a program developed by Cignition, students were offered 10 weeks of tutoring, twice a week (in practice, there were fewer sessions due to absenteeism). Each session was approximately 25 minutes long. Tutors were experienced mathematics teachers who were carefully selected by Cignition for their experience in teaching mathematics and also based on an interview. The tutors received approximately 6 hours of training from Cignition, which covered Cignition's tutoring platform, efficacy program objectives and logistics, professional development on number talks, and best practices for teaching fraction content. Tutors met with students in an online environment in which they could talk and also each draw on a mutually-visible surface. Each tutor met consistently with the same students (3.5% of sessions had substitute tutors).

Students were able to play a game, "FogStone Isle," both before and after tutoring sessions. When students played the game, reports were generated for tutors on what concepts students might be struggling to understand. Also, tutors could assign follow-up work in the game after a tutoring session and receive a report on the student's work. Thus, game play and tutoring were interwoven to target areas of fraction understanding in most need.

Mathematical topics. Three topics in mathematics were covered: equivalence of fractions, comparing fractions, and adding fractions both with like and unlike denominators.

Experimental design. After taking the same diagnostic pretest, students in each research site were paired based on pretest scores. Each pair was randomized: one student was assigned to the Treatment condition and the other to the Control condition. If a student subsequently dropped out, the paired student was also dropped from the study. The number of students who completed the study was 144, evenly divided between the two conditions.

In the Treatment condition, students were assigned to a tutor and also assigned to play the game, as previously described. In the Control condition, students were able to play the game only and playing was optional. In both conditions, students continued to attend their existing classrooms and received ongoing instruction on fractions. All students in both conditions took the same posttest at the end of the study.

Instruments. The same diagnostic test was used both as a pretest and a posttest. It was developed to focus on the three mathematical topics listed above. To create the diagnostic

instrument, sources were consulted from researchers who had developed and validated diagnostics including from Saxe, Diakow, & Gearhart (2013); Wilkins, Norton, & Boyce (2013); and Izsak, Jacobsen and Bradshaw (2019). With these examples, a candidate test was refined and reviewed by assessment experts. It was then piloted in a cognitive lab (think-aloud) process to improve the clarity of items and find ones of appropriate difficulty and which elicited conceptual reasoning. Additionally, items were evaluated to identify floor and ceiling effects. The final test had a total of 18 items of which a total of 46 points was the highest possible score. When the test was scored, the scorers were blind to student identity, student location, and condition.

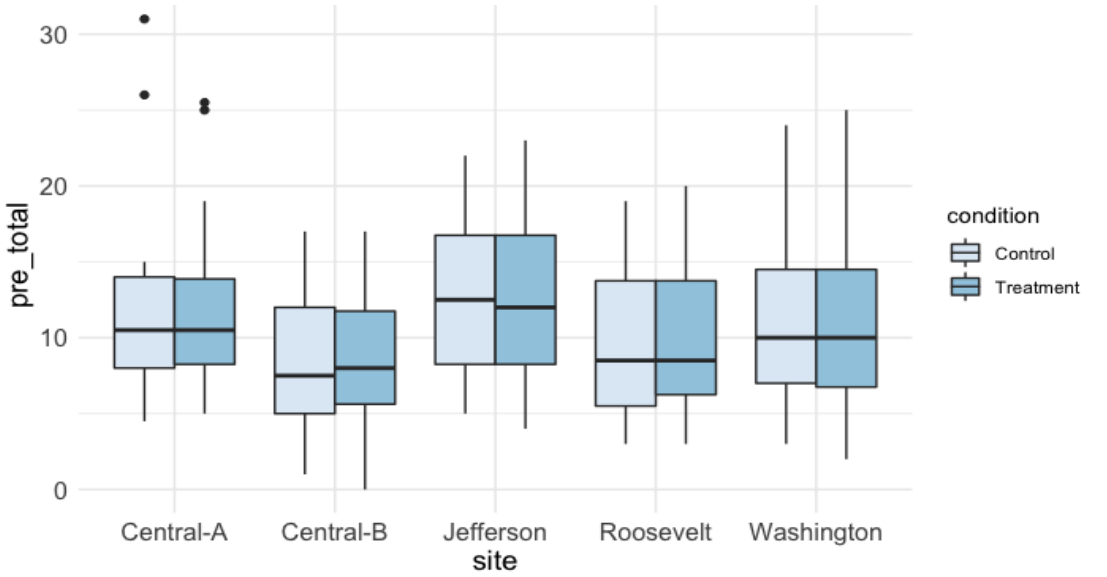
In addition, implementation data was collected. Cignition tracked the number of tutoring sessions with each student. Cignition also tracked how much each student played the game. Additional data about tutoring sessions and game sessions were recorded but is not discussed in this preliminary report.

Analysis Plan. The Digital Promise research team planned to analyze gain scores, which were posttest scores minus pretest scores. Using gain scores in a t-test produces equivalent results to an ANOVA with a pre-post and treatment-control contrasts and is simpler to understand. We also planned to look at each site individually in descriptive statistics and analyze the sites in a meta-analysis. For the meta-analysis, each site was considered one study. We planned to combine the results in a fixed-effects model to determine both an overall effect size and overall statistical significance.

Findings

Pretest scores in the Treatment and Control Group were equivalent in all sites (Figure 1).

Figure 1: Pretest scores match in each classroom



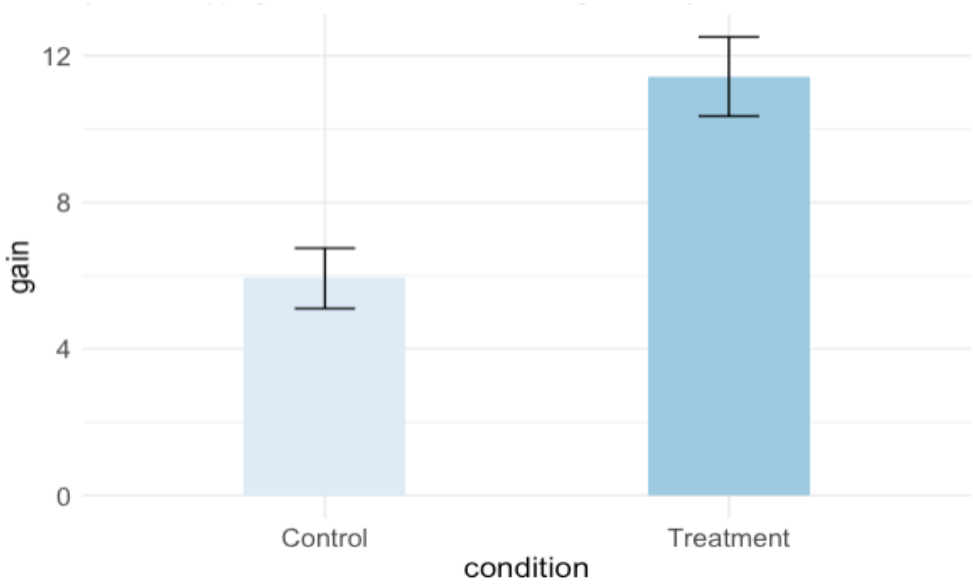
Students in the Treatment condition participated in a minimum of 12 and a maximum of 23 tutoring sessions. The mean number of sessions (18.2) varied by school, with the Washington

site having fewer sessions (15.6). Students in the Treatment condition logged into the game a mean of 27.4 times, as expected. Students in the Control condition logged into the game fewer times (mean 0.86 times; with 73.6% of students never logging in even though it was available to them).

Attrition was limited to 4 students (2 pairs). Total attrition was 3%. Differential attrition was zero (due to dropping pairs).

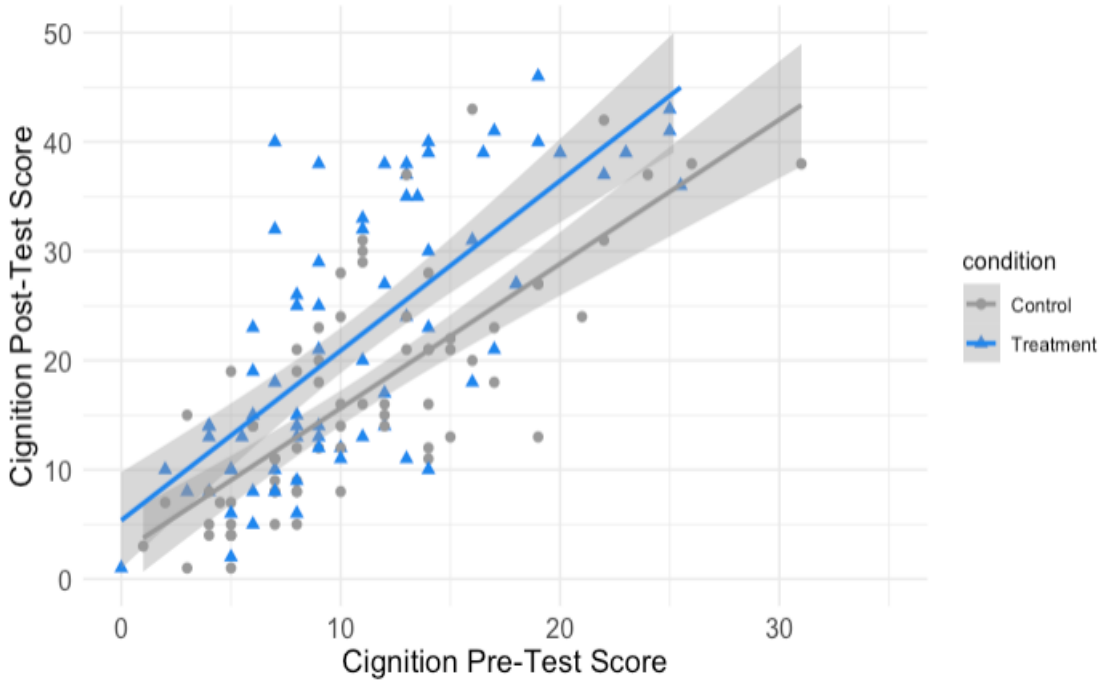
To describe the pattern of results, we first made a chart of the gain score for the Treatment vs. the Control group (Figure 2). The error bars on top of each bar show the standard error; if these error bars overlap, the result is NOT significant. (If the error bars do not overlap, the results may be statistically significant; a further test is required and performed below.)

Figure 2: Gains from Pretest to Posttest higher for Treatment



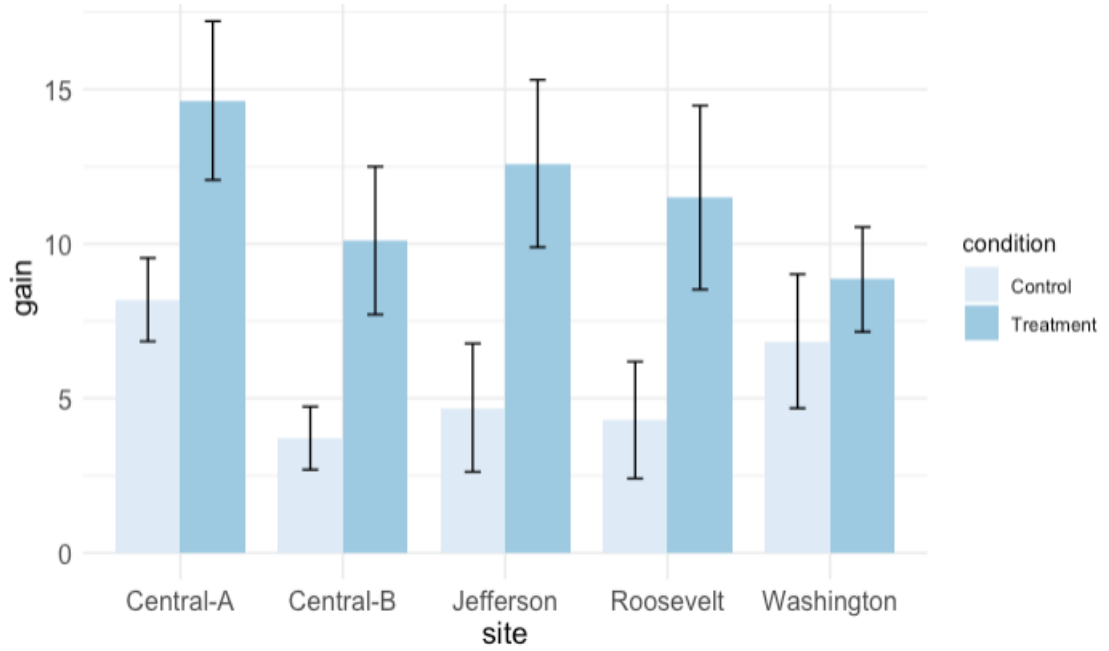
Similarly, we made a scatterplot of the posttest score vs. the pretest score for each group (Figure 3). Each point in this plot is one student and represents their two scores.

Figure 3: Relationship between the Posttest and the Pretest



Students with a higher dot for the same horizontal position performed better than would be expected given their pretest score. The regression line for the Treatment condition is higher than the regression line for the Control condition, suggesting that Treatment students learned more. We conducted a t-test on the aggregate gain scores, comparing conditions; the results were statistically significant (t = 4.0545, df = 132.66, p-value < .001).

Figure 4: Gain scores differences vary by site



We also plotted gain score comparisons for each of the five sites. The error bars for Washington clearly overlap, showing the contrast between conditions is NOT significant in

that site. The differences at other sites may be significant, but the experiment was not powered to detect differences on a site-by-site basis.

Finally, we conducted a meta-analysis of the five sites, using the dmetar package in R (Harrar et al, 2019). The effect size of the Treatment was estimated at g = .66 and was statistically significant (Table 1). This meta-analysis was run using gain scores. A statistical test for heterogeneity did not suggest a need to partition the sites, even though Washington's data appears different visually.

Study	Treatment Total	Mean	SD	Control Total	Mean	SD	Standardised Mean Difference	SMD	95%-CI	Weight
Central-A	18	14.64	10.9017	18	8.19	5.7218		0.72	[0.05; 1.40]	25.1%
Central-B	14	10.11	8.9533	14	3.71	3.8115		0.90	[0.12; 1.69]	18.8%
Jefferson	10	12.60	8.5661	10	4.70	6.5668		0.99	[0.05; 1.93]	13.0%
Roosevelt	10	11.50	9.4074	10	4.30	5.9824		0.87	[-0.05; 1.80]	13.4%
Washington	20	8.85	7.5761	20	6.85	9.7023		0.23	[-0.40; 0.85]	29.8%
Fixed effect model	**72**			**72**				**0.66**	**[0.32; 1.00]**	**100.0%**

Heterogeneity: I^2 = 0%, τ^2 = 0, p = 0.56

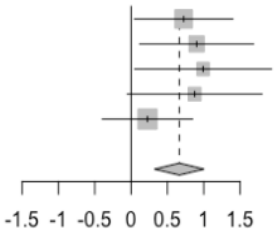


Table 1: Meta-analysis on gain scores in each site finds an overall effect size of 0.66. An analysis that corrects for a potential overestimate found an effect size of 0.45.

Due to software limitations, this meta-analysis used the pooled variance of gain scores, not pooled variance of raw scores. This rescales the effect and potentially overestimates the size of the effect. With consultation with an expert statistician (Elizabeth Tipton, Personal Communication, April 6, 2020), we also ran a meta-analytic model that corrected for the potential overestimate of effect size and found a corrected effect size of +.47 with a 95% confidence interval of [+.23, +.70]. As both of the estimated effect sizes fall within the [+.23, +.70] confidence interval, there's little reason to choose one as the best estimate of effect size. Further replications of the study will tell.

Discussion

We found that students who received online tutoring and played the related "Fog Stone Island" game learned more. Given the design of the study, the contribution of the tutoring component and gameplay component cannot be analytically separated. Further, the components were intended to reinforce each other; a strength of this tutoring approach is that the game contributes to the tutor's knowledge about the student and supports the student in gaining extra practice.

An effect size that falls in the 95% confidence interval of [+.23, +.70] is meaningful in educational research. Consider that the Pelligrini et al. (2018) meta-analysis of elementary math tutoring programs found g = .26 and was based largely on face-to-face tutoring studies. Thus, this study compares favorably to prior research on face-to-face mathematics tutoring with elementary students.

By way of meaningfulness, an effect size of .40 is commonly interpreted as corresponding to an additional year of instruction. If the effect size in this approach could be maintained for a full year, students would gain as much as they would in an additional year of math instruction. As many students are a year behind their peers in math when they are in fifth grade, this is a meaningful effect; it could allow students with weak prior knowledge to meet grade-level expectations.

Limitations of this study are as follows: The pretest and posttests were designed by Cignition, with consultation from our Digital Promise Global evaluation team. This was necessary because only a few fraction concepts could be addressed in the available time for the experiment. Effects might be smaller if a standardized test were used (e.g. because the test would cover all grade level expectations, not just this content). The number of students was also modest; effects may decrease at greater scale (e.g., due to regression to the mean). It is also worth considering that the Cignition approach is still being refined and improvements could increase the magnitude of the effect.

Strengths of the study include its design, which was a randomized controlled trial. Students were well-matched by prior test scores. Tutors were blind to pre- and post-test items. Attrition was low. The data analysis was conducted by an independent, external team.

Conclusion

Helping struggling students to learn fractions is important. Based on prior research one might anticipate that face-to-face tutoring would increase learning, but it was unclear whether online tutoring would work, especially with students who are in fifth grade. The data supported our hypothesis that it would work; the effect sizes in favor of online tutoring were encouraging. Combining tutoring and game components may have contributed both by informing tutors as to what issues students were encountering and by allowing tutors to assign targeted follow-up to students.

The analysis presented here is preliminary. Further research may consider mediating factors, for example, how students interacted with game as well as observed variations in the tutoring sessions. Further research may also generate hypotheses about why the results in one site appeared weaker than in the other sites. There is more analysis to do for a complete report.

These results are particularly timely now, as many students are at home under shelter-in-place orders due to COVID-19. The approach taken here is feasible for students at home, with tutors supporting them from a distance. Although not all low-income and student-of-color populations have sufficient technology and bandwidth at home, it would be possible to consider ways of lowering the requirements. Instead of computers, smart phones might be used. Also, some low-income students do have computers and bandwidth and they can be helped without provisioning new hardware or connectivity. Further, students may be able to get to a library or school with bandwidth, for example, by summertime. In this case, tutoring over the summer might help students prepare for their next grade level, partially making up for any weaknesses in instruction during Spring 2020.

References

Behr, M. J., Wachsmuth, I., Post, T. R., & Lesh, R. A. (1984). Order and equivalence of rational numbers: A clinical teaching experiment. *Journal for Research in Mathematics Education, 15*(5)*,* 323-341.

Booth, J. L., Newton, K. J., & Twiss-Garrity, L. K. (2014). The impact of fraction magnitude knowledge on algebra performance and learning. *Journal of Experimental Child Psychology, 118,* 110-118.

Empson S.B., Levi L., & Carpenter T.P. (2011). The algebraic nature of fractions: Developing relational thinking in elementary school. In J. Cai & E. Knuth (Eds.), *Early algebraization. Advances in mathematics education* (pp. 409-428). Springer, Berlin, Heidelberg.

Hackenberg, A. J. (2013). The fractional knowledge and algebraic reasoning of students with the first multiplicative concept. *Journal of Mathematical Behavior, 32,* 538-563.

Harrer, M., Cuijpers, P., Furukawa, T.A, & Ebert, D. D. (2019). *Doing meta-analysis in R: A hands-on guide.* https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R/. DOI 10.5281/zenodo.2551803.

Moss, J. (2005). Pipes, tubes, and beakers: New approaches to teaching the rational-number system. In M. S. Donovan & J. D. Bransford (Eds.), *How students learn: Mathematics in the classroom* (pp. 121-162). Washington, DC: National Academic Press.

National Research Council. (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: The National Academies Press. https://doi.org/10.17226/9822.

Pellegrini, M., Lake, C., Inns, A., and Slavin, R.E. (2018). *Effective Programs in Elementary Mathematics: A Best-Evidence Synthesis.* Retrieved from http://www.bestevidence.org/math/elem/elem_math_2018.htm

Rittle-Johnson, B., & Koedinger, K. (2009). Iterating between lessons on concepts and procedures can improve mathematics knowledge. *British Journal of Educational Psychology, 79*(3)*,* 483-500.

Saxe, G.B., Diakow, R., & Gearhart, M. (2013). Towards curricular coherence in integers and fractions: The efficacy of a lesson sequence that uses the number line as the principal representational context. *ZDM (International Journal on Mathematics Education), 45,* 343-364.

Siegler, R. S., Thompson, C., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive Psychology, 62,* 273-296.

Streefland, L. (1991). Fractions in realistic mathematics education: A paradigm of developmental research (1 ed. Vol. 8). Dordrecht, The Netherlands: Kluwer.

Thompson, P. W., & Saldanha, L. (2003). Fractions and multiplicative reasoning. In J. Kilpatrick & G. Martin (Eds.), *Research companion to the NCTM Standards* (pp. 95-113). Washington, DC: National Council of Teachers of Mathematics.