The Uncertain Role of Educational Software in Accelerating Student Learning: Regression discontinuity evidence from three local education agencies

Sidrah Baloch
Insight Education Group

Thomas J. Kane
Harvard Graduate School of Education
Center for Education Policy Research at Harvard University

Ethan Scherer
Center for Education Policy Research at Harvard University

Douglas O. Staiger
Dartmouth College

Abstract

Educators must balance the needs of students who start the school year behind grade level with their obligation to teach grade-appropriate content to all students. Educational software could help educators strike this balance by targeting content to students' differing levels of mastery. Using a regression discontinuity design and detailed software log and administrative data, we compare two versions of an online mathematics program used by students in three education agencies. We find that although students assigned the modified curriculum did progress through content objectives more quickly than students assigned the default curriculum, they did not perform better on pre- and post-objective quizzes embedded in the software, and most never progressed far enough to reach the grade-level content. Furthermore, there was no statistically significant effect of the modified curriculum on formative test scores. These findings suggest policymakers and practitioners should exercise caution when assigning exclusively remedial content to students who start the school year behind grade level, even though this is a common feature of many math educational software programs.

---

* This publication is based on research funded by the Bill & Melinda Gates Foundation. The findings and conclusions contained within are those of the authors and do not necessarily reflect positions or policies of the Bill & Melinda Gates Foundation. We thank the local education agencies involved in Proving Ground for co-generating the hypotheses of this research and providing the administrative and educational technology data, as well as the education technology vendor for generating and testing an augmentation to their software specifically built for the intervention. The authors are also thankful for feedback and advice from Dan Thal, leadership at the educational technology company and local education agencies, and conference participants at AEFP. Any errors of fact or interpretation are our own.

I. Introduction

On the 2019 NAEP assessment, 19 percent of fourth-grade students and 31 percent of eighth-grade students scored below the basic achievement level in mathematics. These patterns are even more striking when broken down by race, with 35 and 27 percent of fourth-grade Black and Hispanic students and 53 and 43 percent of eighth-grade Black and Hispanic students respectively scoring below the basic achievement level in math. Moreover, despite extensive efforts to improve math learning, these rates have not changed significantly in the last 15 years.

Previous research has focused on four types of policies for students performing below grade level: grade retention, summer school, increased instructional time, and tutoring. Rigorous analyses show that younger students identified for summer school and potential grade retention tend to experience increases in test scores initially, with few positive or negative effects later (Matsudaira, 2007; Jacob & Lefgren, 2009; Eren et al., 2017; Schwerdt et al., 2017). In contrast, the effects of retaining students in eighth grade depend on context, with some observed negative effects likely related to social stigma (Jacob & Lefgren, 2009; Eren et al., 2017). Studies on the effects of increased instructional time find initial positive test score impacts that fade out over time (Taylor, 2014), but also positive impacts on college entrance tests, credits earned, high school graduation, and college enrollment (Cortes et al., 2015). However, Taylor (2014) has found some suggestive evidence that additional math instructional time crowds out instructional time in other subjects like English, music, and art. There is strong causal evidence for the large positive impacts of tutoring on student learning outcomes, especially in earlier grades and for low-income students (Nickow et al., 2020; Dietrichson et al., 2017). Tutoring is most effective in high doses, one-on-one or in very small groups, and when delivered by teachers, paraprofessionals, or highly trained personnel (Robinson et al., 2021), making it a potentially expensive undertaking for districts. Kraft & Falken (2021) estimate that a K-12 tutoring program would cost about $1,000 per student, with a majority of the cost going toward staffing expenses. There are significant time and monetary costs associated with summer school, grade retention, and increased instructional time as well. Taylor (2014) estimates that Miami-Dade County Public Schools required a 15% increase in math teachers across the district to support their remedial class intervention, which could be particularly costly in a tight teacher market.

Support for students performing below grade-level has gained increased importance in the context of the COVID-19 pandemic. Early estimates of unfinished learning from spring 2021 indicate significant decreases in student learning rates (Kuhfeld et al., 2021). These are likely underestimates––a supplementary analysis indicates that students who participated in testing during the 2020-21 school year were less racially diverse and higher performing than average. In addition, policymakers believe the effect of the pandemic will linger for years. For example, embedded in the American Rescue Plan Elementary and Secondary School Emergency Relief Fund is $1.2 billion of funding for evidence-based summer enrichment programs as well as $21 billion for evidence-based initiatives to address the impact of lost instructional time (U.S. Department of Education, 2021). Thus, now more than in the past, there is an acute need to address the learning needs of students working to catch up to grade-level content.

Educational software programs may provide a time- and cost-efficient way to assist students who are academically behind their peers. These are targeted programs designed to develop a particular skill in coordination with a teacher's regular curriculum. They have the potential to "personalize" education and meet each student's performance level, allowing teachers to effectively instruct students with varying abilities and strengths in one classroom. One possible way to elevate student achievement is to incorporate below-grade-level content into an existing

curriculum using educational software, allowing students to progress and learn at their own pace while remaining in the same classroom as their grade-level peers. The theory of change for this approach is that reteaching or solidifying understanding of foundational below-grade-level concepts increases students' self-efficacy in the subject and ultimately accelerates grade-level learning compared to the default condition in which students do not receive any below-grade-level scaffolding. If effective, this practice could help eliminate the social stigma of holding students back, as discussed by Jacob and Lefgren (2009), as well as potentially save school systems the high financial and staffing costs of interventions such as summer school, grade retention, and additional remedial coursework.

In this study, we take advantage of detailed software log data linked to student-level administrative records to estimate the causal effects of providing low-performing students with below-grade-level content via an online math program. Our pilot draws upon approximately 2,800 students enrolled in grades 3-6 in three local education agencies, including traditional district schools and charter schools. We employ a regression discontinuity approach to compare two versions of the program. Students with prior test scores below a designated cutoff were assigned a modified version of the curriculum with below-grade-level content before the grade-level curriculum, while those above the cutoff received only the default grade-level curriculum. We examine whether providing students with this modified curriculum accelerated their completion of learning objectives, improved pre- and post-objective quiz scores, and raised subsequent standardized math test scores.

The results of our quasi-experimental analysis suggest that while students assigned the modified curriculum made faster progress through learning objectives in the program, most of them did not ultimately advance to the grade-level material that followed. They also did not perform better on the most proximal measure of pre- or post-objective quizzes embedded in the software, nor did they perform better on standardized assessments. These results occur even though most of the modified curriculum students were working on easier below grade-level content that they should have been taught in the prior year. Specifically, we find that students who used the modified curriculum completed about 0.09 more objectives per hour than students who used the default curriculum, representing an approximately 22-percent increase in objectives completed per hour before the modified curriculum was introduced. Yet, since students in the modified curriculum group received 6-7 below-grade-level objectives, their higher rates of progress were not enough to advance them to the grade-level content, and in fact they were 70 percent less likely to complete any grade-level objectives at all.

While not conclusive, our study suggests that policymakers and practitioners should exercise caution when considering the use of educational software that place students onto exclusively remedial content to support and encourage students who begin the school year behind grade level. While education software often complements regular grade-level instruction, our observation that so few students who received the modified curriculum were ever able to work on grade-level concepts in the program raises the question of whether assigning students exclusively remedial content at the beginning of the school year only perpetuates the challenges associated with being behind their grade-level peers. Our study is the first, to our knowledge, to link detailed software log data with district administrative data to assess and quantify whether educational software could provide an alternative to other strategies for supporting below-grade-level students, such as grade retention, summer school, increased instructional time, and tutoring.

II. Background and Intervention

A. Background

This study was conducted in three school systems (consisting of both traditional school districts and charter management organizations) located in the Pacific and East South Central regions of the United States. All three sites at the time of the study used a math educational software that focuses on elementary and middle grades (K-8). The software is almost entirely visual and utilizes very little language-based instruction, but it is structured like a textbook, with students progressing through a certain number of learning objectives during the course of a specific grade-level curriculum. While teachers can modify the order of objectives within the grade-level curriculum, they do not have the ability to add above- or below-grade-level objectives to the syllabus.

The software provider recommends a weekly usage level (in minutes) and an approximate amount of content that students should cover each week in order to complete a majority of the curriculum by the end of the year. However, the mode of delivery of the software (e.g., large school-based labs, mobile laptop carts for classroom) varies by site, and often by school within one site. At the beginning of this study, all three sites had been implementing the software for several years, particularly in their younger grades, with 60 to 80 percent of all classrooms using the software.

In our past work with this group of school systems, we used a matched comparison method to estimate the impact of the math software program on student achievement in previous years of implementation. We found that students who increased their weekly usage of the software from fall to spring experienced significant math score gains on state accountability tests. However, on average, students with lower prior achievement were using the software for less time each week than students with higher prior achievement. School system leaders hypothesized that low-achieving students were less engaged with the software because they were struggling to work through math content that was geared toward their assigned grade level rather than their true ability level. Based on this hypothesis, these three sites developed an intervention strategy in coordination with the software vendor to help increase the software usage and math achievement of students with lower test scores.

B. Intervention and Assignment Strategy

The pilot study took place over the course of the 2016-17 school year. The software provider designed modified versions of the regular syllabi for grades 3-6. For each grade level, the provider identified foundational learning objectives from earlier grade levels and placed them at the beginning of the grade-level curriculum. For example, the fourth-grade modified curriculum began with key activities from the second- and third-grade levels before introducing the fourth-grade content. In contrast, the default fourth-grade curriculum contained only the fourth-grade content. The goal of this intervention was to provide additional scaffolding and support to students who otherwise would have likely struggled to complete their grade-level activities in the software program.

We identified students in grades 3-6 in all three participating sites to receive the modified curriculum based on their prior spring math assessment scores. Students who scored below a particular cutoff (roughly corresponding to the bottom quartile of achievement across the entire sample) were assigned to a treatment group that would receive the modified curriculum, while students who scored above the cutoff were assigned to a control group that would receive the

default grade-level curriculum.[1] The software provider pushed out the correct versions of the curriculum directly to students, reducing the likelihood that individual teachers would switch the order of the objectives or otherwise change the design of the modified or default curriculums.

III. Data, Descriptive Statistics, and Hypotheses

A. Data Sources

This study uses student-level data from two sources. Student demographic characteristics, school and class enrollment information, and formative and summative test scores are provided by the local education agencies. The running variable is derived from students' prior spring math test score. For most students in grades 4-6, this is their spring 2016 state assessment. For third-grade students and students in a state that did not administer an assessment in 2016, their spring formative assessment is used instead.[2] State test scores are standardized by each state's reported means and standard deviations. The formative test scores are standardized using national norming studies provided by the company administering the assessment. The standardized test scores are then converted to percentiles and centered at the cut scores to create a common running variable for all students.

The math software vendor provides data on students' usage of the software, including the number of minutes they spend using the program each week, completion of individual objectives, performance on quizzes taken before and after each objective, and their assigned version of the program (modified curriculum or default curriculum). We aggregate these data to the student level to calculate each student's average weekly minutes of usage, number of objectives completed per hour[3], and pre- and post-quiz scores for the time periods before and after the modified curriculum was turned on for the treatment group.[4] These data are then merged with the demographic, enrollment, and test score data described above.

B. Descriptive Statistics

Table 1 presents summary statistics on all students eligible for participation in the study.[5] There are two analytic samples described in this table. The software sample includes students from all three sites. One site did not administer formative tests in the fall, so the test score sample is restricted to the two sites with complete test score information. Both samples are

[1] Note that we identified the test score cutoffs for each grade level across all students in the participating sites. However, within each site, schools had to opt in to participate in the study. Thus, while there is a clear cutoff score for the design of our regression discontinuity model, the test score percentile distribution we observe around the cutoff does not appear uniform.

[2] Students at one site did not take a state assessment in 2016, but they did take the same formative assessment as students in other sites. We created concordance tables with the other sites' test score data to map the state assessment against the formative assessment and translate the state assessment cut scores for each grade level to the formative assessment's scale.

[3] We divide the number of lessons completed per hour to normalize the completion across classrooms and sites since students spent different amounts of time on the application

[4] Modified curriculum start dates varied by site and ranged from early October 2016 to early December 2016. However, start dates were consistent for students within a site. All students used the default curriculum in the weeks before the modified curriculum was turned on for the treatment group students.

[5] Participation varied by site. In some sites, only certain schools, grade levels, or classrooms were included in the study.

largely comprised of English language learners, Hispanic students, and students receiving free or reduced-price lunch. The majority of students possess all three of these characteristics.

In the weeks prior to the introduction of the modified curriculum, students used the math software for an average of 59 minutes per week, below the vendor's recommended level of 90 minutes per week. Average usage following the start of the modified curriculum is somewhat lower, at 53 minutes per week. In the pre-treatment period, students received average pre-quiz scores of 0.46 and average post-quiz scores of 0.55 (on scales of 0 to 1). Following the start of the modified curriculum, students received average pre-quiz scores of 0.59 and average post-quiz scores of 0.72, representing an overall increase from the fall. Students completed about one-third of an objective per hour in both the pre- and post-treatment periods.

Students' fall and spring formative math and ELA scores represent their academic ability at the beginning and end of the school year.[6] On average, students in our sample demonstrate a 0.77 standard deviation increase in math achievement and a 0.60 standard deviation increase in ELA achievement over the course of the 2016-17 school year. Compared to national norms, they end the year slightly above average in math performance and slightly below average in ELA performance.

C. Hypotheses

This pilot study follows an earlier analysis of the implementation and impact of the math instructional software program in 13 education agencies, including the three sites in our sample. The results of this investigation indicate that usage of the software varies significantly by prior achievement: on average, students with lower prior test scores use the software for less time each week than students with higher prior test scores. At the same time, increased usage of the software is associated with significant gains in math test scores, and though not statistically different, the magnitude of these gains is largest for students in the bottom quartile of prior achievement. We presented these results to the 13 sites, and three of them were interested in testing an intervention using the educational software to improve engagement and achievement outcomes for their lowest performers. The data presented informed their hypotheses as they co-created a modified version of the math software in partnership with the vendor. Based on anecdotal feedback, these sites hypothesized that a key mechanism driving the lower usage of low-performing students was that these students lacked the knowledge to complete their grade-level content, resulting in lower self-efficacy, frustration, and disengagement with the software overall. They hypothesized that students would spend more time using the software if allowed to complete objectives more closely aligned with their true math ability before beginning their grade-level material. Thus, their proposed intervention was a modified version of the software that gradually ramps students up to their grade-level objectives.

Unlike other math educational software programs, the program in this study does not rely on diagnostic tests to place students in different curriculum levels. By default, students are assigned a syllabus that corresponds to their grade level and contains a set number of activities aligned to their grade-level math concepts. The sites' first hypothesis was that providing below-grade-level content to low-performing students would improve their engagement with the software, strengthen their understanding of the material, and increase the rate at which they completed learning objectives. Then, once they reached the grade-level content, they would have

[6] The growth in test scores from fall to spring is expected. The formative test is vertically scaled, and the scores from both testing periods are standardized against the spring national norms to place both the fall and spring tests on the same scale.

more confidence and stronger foundational content knowledge, which would allow them to accelerate their learning of grade-level material in the software. Detailed student-level log data made available by the educational software company provides strong proxies to measure these outcomes. First, students' average weekly minutes of software usage proxies for engagement. Second, pre- and post-objective quiz scores measures students' proximal cognitive understanding of the content. Finally, the number of objectives completed per hour represents students' learning progress, which the sites expected would improve in response to the modified curriculum, both because the below-grade-level content would fall within students' optimal learning zone and because students would be more engaged with the software. Then, when students completed the below-grade-level objectives, they would perform better on grade-level pre- and post-quizzes and experience an acceleration of grade-level content completion compared to the comparison group.

The sites' second hypothesis was that deeper engagement and stronger foundational content knowledge in response to the modified curriculum would increase students' distal math test scores at the end of the school year. This aligned with our previous findings that increased usage of the software is associated with math test score gains, particularly for low-achieving students. We would not expect to see a corresponding impact on ELA test scores because the math software is a visual program that is designed to be mostly language independent. Accordingly, we examine ELA scores as a placebo outcome in our analysis.[7]

IV. Methodology

We employ an intent-to-treat regression discontinuity technique to estimate the impact of the modified curriculum on average weekly minutes of software usage, pre- and post-objective quiz scores, number of objectives completed per hour, access to grade-level objectives, and formative math and ELA test scores.[8] Students who score below a particular cutoff on their prior spring math assessment are assigned the modified curriculum, which includes key content from earlier grade levels, while students who score above the cutoff receive their default grade-level curriculum. Equation (1) describes the relationship of the treatment indicator T_i to the cut score c:

$$T_i = \begin{cases} 0 \; if \; c > 0 \\ 1 \; if \; c \leq 0 \end{cases} \tag{1}$$

We begin by confirming that there is no discontinuity in the running variable or covariates around the assignment cutoff. We then use a standard regression discontinuity model, shown in equation (2), to estimate the effect of the modified curriculum on outcome Y of student i. In this model, T_i is an indicator of whether the student was assigned the modified curriculum. $(X_i - c)$ is a linear function of the running variable X_i, prior math percentile score, centered at the cut score c. Z_i is a vector of student demographics and pre-treatment covariates. α_i represents site-fixed effects.

[7] This does not hold in the other direction: there are plausible explanations for how this intervention could decrease ELA scores, e.g., by increasing time spent on math at the expense of time spent on ELA. Also, non-cognitive factors, like persistence, could be enhanced by struggling through math problems that would also have an effect on ELA scores, but we would expect these to be small in magnitude.

[8] A small number of students assigned to the treatment group did not use the modified curriculum. We do not conduct a fuzzy regression discontinuity analysis to address noncompliance at this stage.

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 (X_i - c) + \beta_3 (X_i - c) * T_i + Z_i + \alpha_i + \varepsilon_i \quad (2)$$

The coefficient β_1 is the main estimate of the discontinuous effect of the modified curriculum at the test score cutoff. We focus on fitting this model within a bandwidth of 15 percentile points above and below the cutoff; however, we also run sensitivity tests with a range of bandwidths, including the "optimal" non-parametric bandwidth selected by a mean squared error method. Our final sensitivity test involves running a non-parametric model across the same range of bandwidths. The non-parametric model uses a triangular kernel to place more weight on observations closer to the cutoff. In all cases, we calculate robust standard errors clustered at the school-grade level.

V. Results

A. Density and Covariate Balancing Tests

We begin by examining the density of the running variable to confirm there is no discontinuity at the assignment cutoff. Figure 1 shows the share of students in the larger software usage sample with prior spring math percentile scores 20 points above and below the cutoff. As expected, there is no evidence of a discontinuity at the threshold. Next, we apply our standard regression discontinuity model to examine whether any of the pre-treatment covariates vary discontinuously at the cutoff. Table 2 shows that there are no discontinuities in student demographics, average weekly minutes of software usage, number of objectives completed per hour, or fall formative test scores at the cutoff. These results confirm that the treatment group and control group are demographically and academically similar close to the cutoff. However, by chance, students assigned the modified curriculum had higher pre- and post-objective quiz scores during the pre-treatment period. As such, we would expect the post-treatment results to be biased upward. To partially address these concerns, all models control for baseline performance.

B. Impact Estimates

Our main outcomes of interest are students' average weekly minutes of software usage in the post-treatment period (a measure of engagement), their average pre- and post-quiz scores on grade-level objectives in the post-period (a proximal measure of content mastery), their number of objectives completed per hour in the post-period (a measure of progress), and their spring formative math assessment scores (a distal measure of academic achievement). We also examine whether students ever accessed the grade-level objectives after completing the below-grade-level objectives to test the hypothesis that providing students with below-grade-level content would accelerate their grade-level learning. Finally, we analyze the impact of the modified curriculum on students' spring formative ELA scores as a placebo indicator. Observing a discontinuity in ELA scores at the cutoff would be concerning; we do not expect the modified curriculum to have a meaningful effect on English learning or achievement because the math software is mostly language independent.

Figures 2-4 illustrate the unadjusted non-parametric regression discontinuity models for these outcomes within a bandwidth of 15 percentile points above and below the cutoff. In the top left panel of figure 2, the average weekly software usage of students directly below the cutoff (the modified curriculum group) appears higher than that of students directly above the cutoff (the

default curriculum group), suggesting that students who received the modified curriculum spent more time using the software each week than those who received the default curriculum. The top right and bottom left panel of figure 2 shows that the modified curriculum did not appear to have much impact on pre- and post-objective quiz scores within the software, the most proximal measure of content mastery. The left panel of Figure 3 shows a positive discontinuity for number of objectives completed per hour, indicating that students who received the modified curriculum completed more objectives per hour than students who received the default curriculum. However, the right panel of figure 3 shows that these students were much less likely to ever reach the grade-level objectives that followed. Finally, figure 4 shows small positive discontinuities for both formative math and ELA test scores.

Table 3 reports the impact estimates of the modified curriculum using the linear model outlined in equation (2) within the same bandwidth of 15 percentile points. A positive coefficient signifies a positive effect of the modified curriculum, while a negative coefficient signifies a negative effect. After controlling for student demographics, pre-treatment covariates, and site differences, the modified curriculum's effects on average weekly minutes of usage and pre- and post-objective quiz scores are not statistically different from zero. The confidence intervals on the null effects are relatively precisely estimated. We do, however, observe a positive and statistically significant effect of the modified curriculum on the number of objectives completed per hour. Students who receive the modified curriculum complete 0.09 more objectives per hour than students who receive the default curriculum. However, despite making more progress through their below-grade-level objectives compared to the progress through grade-level content made by students assigned the default curriculum, students who received the modified curriculum were over 70 percentage points less likely to work on even one grade-level objective in the post-treatment period.

We do not observe a statistically significant effect of the modified curriculum on either math or ELA test scores. In our earlier analytic work on the impact of this software program (outside the scope of this paper), we observed a 0.04 standard deviation increase in math test scores for the average student who increased his or her fall to spring usage by an average of about 30 minutes per week. Because we did not observe a significant increase in the average weekly usage of modified curriculum students, it is not surprising that we do not identify a significant effect on test scores.[9]

C. Sensitivity Tests

Table 4 shows the impact estimates for all three outcomes under two different model specifications and across multiple bandwidths, including the "optimal" non-parametric bandwidth for each outcome, calculated using the mean squared error approach outlined by Calonico et al. (2014). The parametric regression discontinuity model is the one outlined in equation (2); the non-parametric model is included as a sensitivity check. Though the magnitudes of the coefficients on objectives completed per hour and access to grade-level objectives differ across models, their directions are consistent, providing suggestive evidence

[9] We conduct a similar analysis using state-administered summative test scores for students in grades 4-6 in three sites and find that the effect on math scores is close to zero and insignificant (-0.01 standard deviations) and the effect on ELA scores is negative and statistically significant (-0.26 standard deviations). One possible explanation for the apparently negative impact of the modified curriculum on ELA scores is that teachers diverted time away from ELA instruction for students who received the modified curriculum in order to give them more time to work through their extended math software curriculum.

that the modified curriculum has a positive effect on students' progress through the software yet a negative effect on their ability to reach grade-level content. This table also illustrates the bias-variance tradeoff inherent in the bandwidth selection decision for a regression discontinuity analysis such as this one. Closer to the cutoff, our estimates are less biased but also less precise. Farther away from the cutoff, we gain precision but also increase bias in our estimates because students are less comparable.

VI. Discussion

In this study, we use a regression discontinuity design with unique educational software log data linked to traditional administrative data to causally investigate whether assigning students remedial work prior to grade-level material through a math educational software program can provide teachers an efficient way to support below-grade-level students. While educational software promises to "meet students where they are," allowing teachers to instruct students of varying levels more optimally, there is no existing causal evidence on the effectiveness of this strategy. Our work demonstrates that, at a minimum, the strategy of providing students with remedial content needs further refinement to achieve that goal. Moreover, our model of providing descriptive and causal data to a network of agencies to inform hypotheses and work in partnership with educational software vendors to pilot and test could provide a fruitful roadmap for the future.

We find that, among students quasi-randomly assigned to receive a modified mathematics curriculum containing material from earlier grade levels, there is no statistically significant effect on engagement, proxied by minutes on the software, no positive effect on proximal pre- or post-objective quizzes, and no statistically significant increase in math test scores. These patterns are particularly concerning for two reasons. First, students in the modified curriculum group tended to perform 10 to 15 points better on the pre- and post-quiz scores prior to the beginning of the intervention. This suggests a potential upward bias in our treatment group, even though we control for the student's pre-curriculum quiz performance in our models. Yet the effect on post-curriculum quizzes can't be distinguished from zero. Second, the students enrolled in the modified curriculum were receiving significantly easier content than their peers who received the default curriculum. Thus, on these measures, there is no evidence that providing the modified curriculum improved student engagement or understanding the material, despite receiving content that should have been covered in the prior year.

Our quasi-experimental design does find an acceleration of content completion, but also provides detailed data that providing remedial content might not be beneficial. We observe an acceleration in the rate of content completion of 0.09 objectives per hour, a 22-percent increase from baseline, but this acceleration is insufficient to bring approximately 70 percent of students back to the first grade-level objective. Thus, the modified curriculum students did not end the year at a similar place in the syllabus as the default curriculum students. The acceleration in learning was not sufficient to cover the material assigned to them, suggesting that this remedial approach to instruction could perpetuate a cycle in which students provided below-grade-level instruction are never able to catch up to their grade-level peers. These findings are congruent with a recent descriptive study published by The New Teacher Project, which found that assigning students below-grade-level content could be harmful (TNTP, 2018). A follow-up study using descriptive Zearn data showed that starting students on grade-level content and then interspersing below-grade-level content at key points prior to the grade-level content demonstrated significantly better results compared to the "ramp-up" style curriculum that we tested (TNTP, 2021). Alternatively, the amount of below-grade-level content a student receives

may be an important factor. The software program in our study assigned students 6-7 key below-grade-level concepts before resuming the grade-level curriculum, but perhaps fewer below-grade-level objectives could be assigned, even if students have not mastered all the foundational concepts. While these emerging hypotheses should be rigorously tested, it does indicate the promise of educational software could be fulfilled with further experimentation surrounding the curriculum design.

Our partnership and collaborative approach provide one way these hypotheses could be tested. The current pilot took place in the context of a larger project that created a network of district and charter agencies to collectively learn about the implementation and impact of educational software. By convening this network of partners, the sites, in collaboration with the authors identified challenges, identified potential solutions, and then co-designed a test of the potential solution with an educational software partner. Implementing continuous improvement networks at scale such as these in a partnership with educational software vendor could allow for specific software strategies to be causally tested and monitored, accelerating our understanding of this sector as well as other potential interventions. Our work provides a proof point that networks like these can help the field learn about what works for whom.

Our work offers several lessons for the design of future evaluations of assisting students who start the school year behind their peers. First, while we were able to leverage rich administrative and log data it could be helpful to obtain additional data on student engagement apart from their usage of the software, such as attendance data. This could help us better understand if the minutes are indicative of allowing the students more time to work on the software or by an actual improvement in these students' overall engagement with school. Adding these metrics to our analysis will help us gain a deeper understanding of how or if this intervention worked. Second, another important piece to understanding the benefits of instructional software as a tool for remedial education would be identifying the costs of the software, including the licenses, hardware, network creation, and high-quality teaching required for effective implementation. While we have begun to think through these factors, in order to really weigh the pros and cons of additional class time versus additional educational software usage, we need to obtain better data on many of these costs. Third, there is a learning curve to participating in these improvement networks, thus we suggest that districts and policymakers commit to them in order to reap the benefits of the setup cost.

By documenting mixed findings with a math instructional software in response to the inclusion of below-grade-level material, this paper adds to a growing literature on educational software in supplementing and improving instruction for students who are otherwise at risk of falling behind (see, e.g., Escueta et al., 2017). How or if these software programs can be used to address this particular issue remains an understudied area in the economics of education.

References

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica, 82*(6), 2295-2326.

Cortes, K. E., Goodman, J. S., & Nomi, T. (2015). Intensive math instruction and educational attainment long-run impacts of double-dose algebra. *Journal of Human Resources*, *50*(1), 108-158.

Dietrichson, J., Bøg, M., Filges, T., & Jørgensen, A. K. (2017). Academic interventions for elementary and middle school students with low socioeconomic status: A systematic review and meta-analysis. *Review of Educational Research, 87*(2), 243-282.

Eren, O., Depew, B., & Barnes, S. (2017). Test-based promotion policies, dropping out, and juvenile crime. *Journal of Public Economics*, *153*, 9-31.

Escueta, M., Quan, V., Nickow, A. J., & Oreopoulos, P. (2017). *Education technology: An evidence-based review* (No. w23744). National Bureau of Economic Research.

Jacob, B. A., & Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics*, *86*(1), 226-244.

Jacob, B. A., & Lefgren, L. (2009). The effect of grade retention on high school completion. *American Economic Journal: Applied Economics*, *1*(3), 33-58.

Kraft, M. A., & Falken, G. T. (2021). *A blueprint for scaling tutoring across public schools* (EdWorkingPaper NO. 20-335). Annenberg Institute at Brown University.

Kuhfeld, M., Ruzek, E., Lewis, K.,&McEachin, A. (2021). Technical appendix for: Learning during COVID-19: Reading and math achievement in the 2020-21 school year. NWEA.

Matsudaira, J. D. (2008). Mandatory summer school and student achievement. *Journal of Econometrics*, *142*(2), 829-850.

Nickow, A. J., Oreopoulos, P., & Quan, V. (2020). *The impressive effects of tutoring on PreK-12 learning: A systematic review and meta-analysis of the experimental evidence* (EdWorkingPaper No. 20-267). Annenberg Institute at Brown University.

Robinson, C. D., Kraft, M. A., Loeb, S., & Schueler, B. E. (2021). *Accelerating student learning with high-dosage tutoring* (EdResearch for Recovery Design Principles Series). Annenberg Institute at Brown University.

Schwerdt, G., West, M. R., & Winters, M. A. (2017). The effects of test-based retention on student outcomes over time: Regression discontinuity evidence from Florida. *Journal of Public Economics*, *152*, 154-169.

Taylor, E. (2014). Spending more of the school day in math class: Evidence from a regression discontinuity in middle school. *Journal of Public Economics*, *117*, 162-181.

TNTP. (2018). *The Opportunity Myth: What Students Can Show Us About How School Is Letting Them Down—and How to Fix It*. https://tntp.org/assets/documents/TNTP_The-Opportunity-Myth_Web.pdf

TNTP. (2021). *Accelerate, Don't Remediate: New Evidence from Elementary Math Classrooms*. https://tntp.org/assets/documents/TNTP_Accelerate_Dont_Remediate_FINAL.pdf

U.S. Department of Education (2021, April 26). *U.S. Department of Education Launches National Summer Learning & Enrichment Collaborative to Help Students Most Impacted by the Pandemic*. https://www.ed.gov/news/press-releases/us-department-education-launches-national-summer-learning-enrichment-collaborative-help-students-most-impacted-pandemic

Workman, E., 2014. Third grade reading policies. Technical Report, CO: Education Commission of the State of Denver.

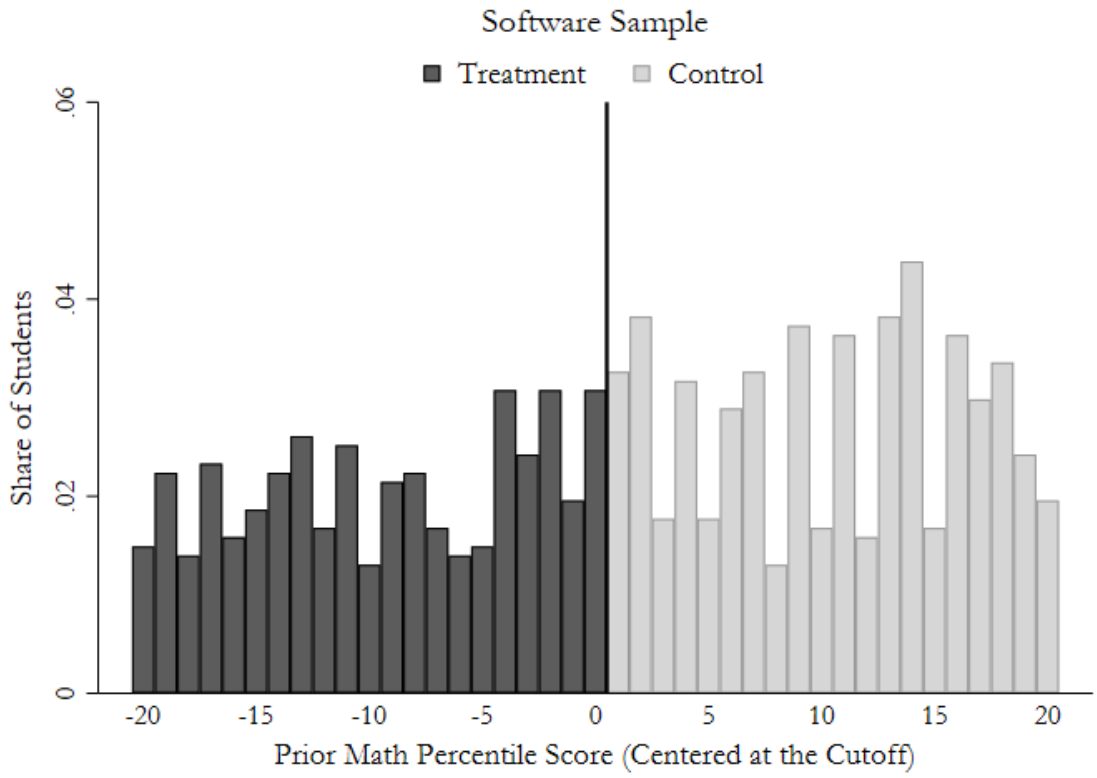
Tables and Figures

Table 1
Summary Statistics

	Software Sample (Three Sites)			Test Score Sample (Two Sites)		
	Mean	**SD**	**N**	**Mean**	**SD**	**N**
Demographics						
Male	0.49	0.50	2,852	0.50	0.50	2,194
ELL	0.70	0.46	2,852	0.76	0.42	2,194
Hispanic	0.70	0.46	2,852	0.77	0.42	2,194
FRPL	[Not available for one site]			0.79	0.41	2,194
Pre-treatment covariates						
Pre-treatment average weekly minutes of usage	58.61	29.09	2,852			
Pre-treatment average pre-quiz score	0.46	0.31	2,852			
Pre-treatment average post-quiz score	0.55	0.39	2,852			
Pre-treatment objectives completed per hour	0.32	0.35	2,852			
Fall math std. score				-0.75	0.96	2,194
Fall ELA std. score				-0.70	1.13	2,194
Outcomes						
Post-treatment average weekly minutes of usage	52.53	23.95	2,852			
Post-treatment average pre-quiz score	0.59	0.22	2,852			
Post-treatment average post-quiz score	0.72	0.24	2,852			
Post-treatment objectives completed per hour	0.34	0.17	2,852			
Spring math std. score				0.02	1.02	2,194
Spring ELA std. score				-0.10	1.07	2,194

Note: The growth in test scores from fall to spring is expected. The formative test is vertically scaled, and the scores from both testing periods are standardized against the spring national norms to place both the fall and spring tests on the same scale.

Figure 1
Density of the Running Variable



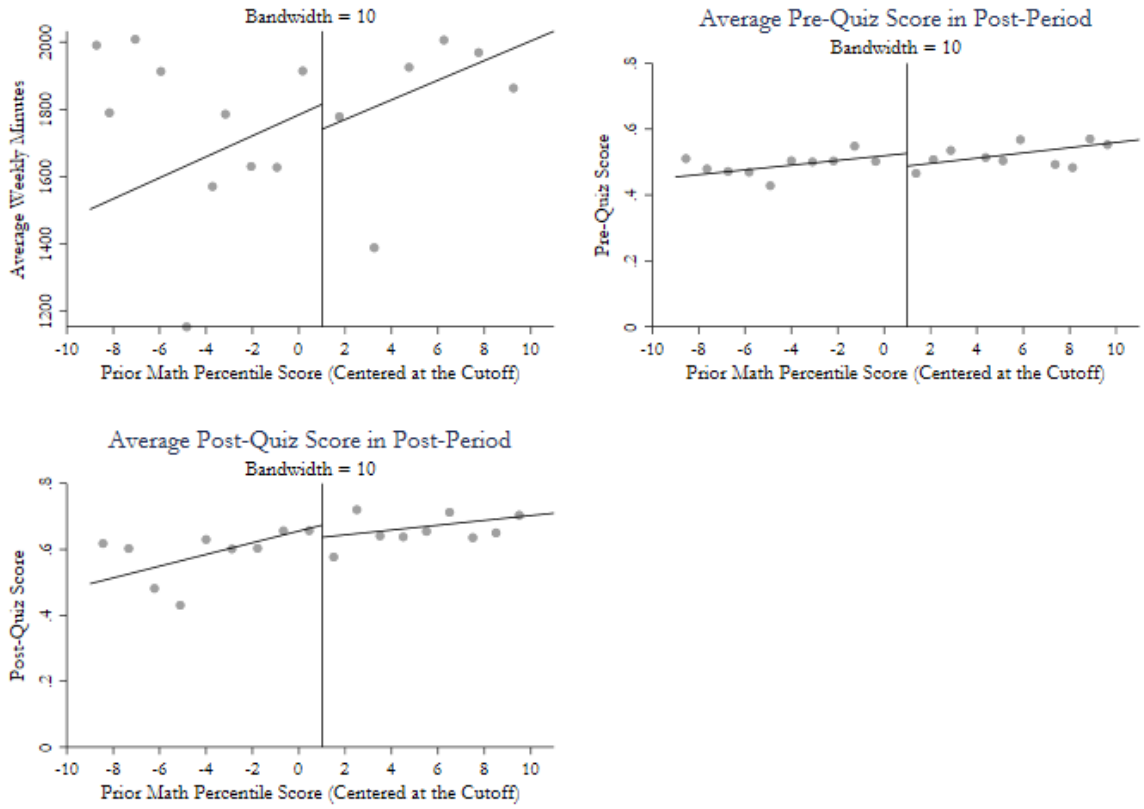
Note: Percentiles were identified across all students at all three sites. However, not all students participated in the pilot study. As such, we do not observe a uniform distribution of test scores percentiles around the cutoff.

Table 2
Covariate Balance around the Cutoff

	Bandwidth = 15 percentiles	
	RD Estimate	**N**
Demographics		
Male	-0.02 (0.07)	782
ELL	0.07 (0.06)	782
Hispanic	-0.03 (0.04)	782
FRPL	-0.01 (0.06)	571
Pre-treatment covariates		
Pre-treatment average weekly minutes of usage	-0.28 (2.93)	782
Pre-treatment average pre-quiz score	**0.09**** (0.04)	782
Pre-treatment average post-quiz score	**0.14***** (0.05)	782
Pre-treatment objectives completed per hour	-0.04 (0.07)	782
Fall math std. score	0.05 (0.08)	571
Fall ELA std. score	0.08 (0.12)	571

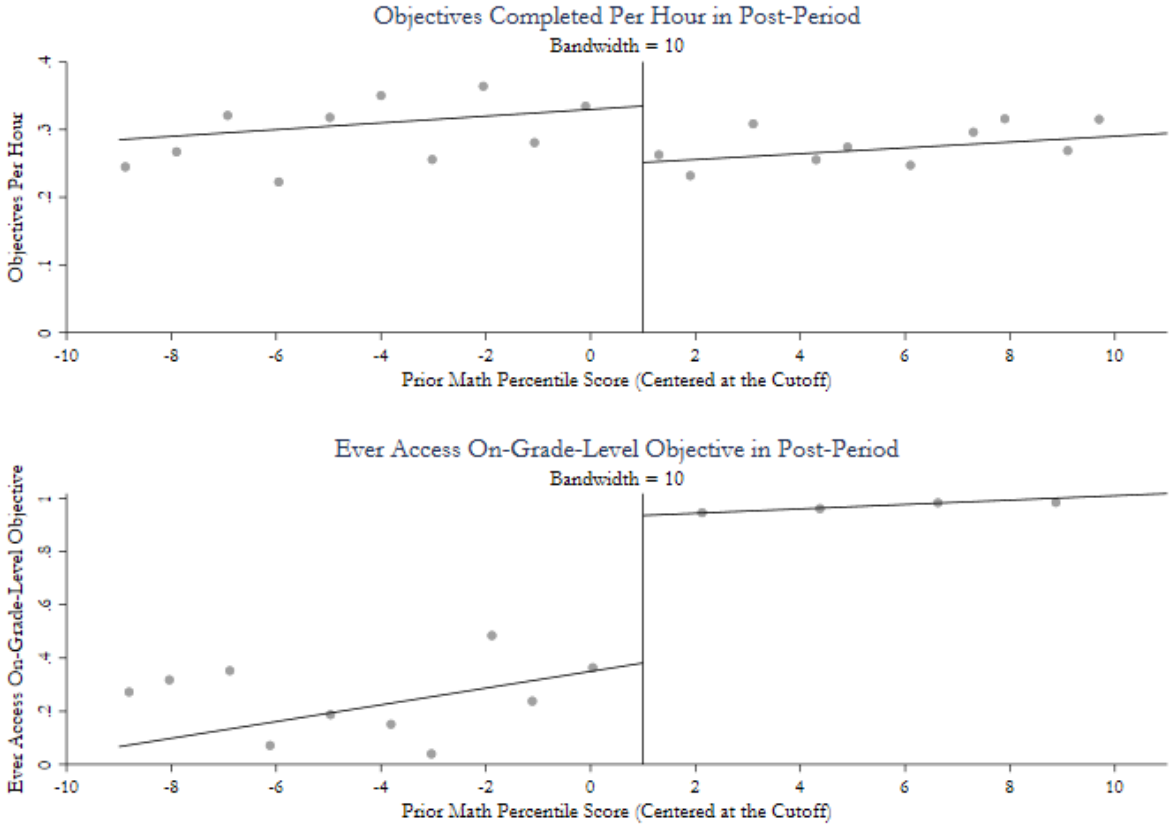
Notes: Parametric regression discontinuity model. School-grade-clustered standard errors are in parentheses. The sample for the male, ELL, Hispanic, and pre-treatment usage analyses is the software usage sample (three sites). The sample for the FRPL and fall test score analyses is the test score sample (two sites).
*** p<0.01, ** p<0.05, * p<0.1

Figure 2
Impact of the Modified Curriculum on Post-Treatment Average Weekly Minutes of Usage, Average Pre-Quiz Score and Average Post-Quiz Score



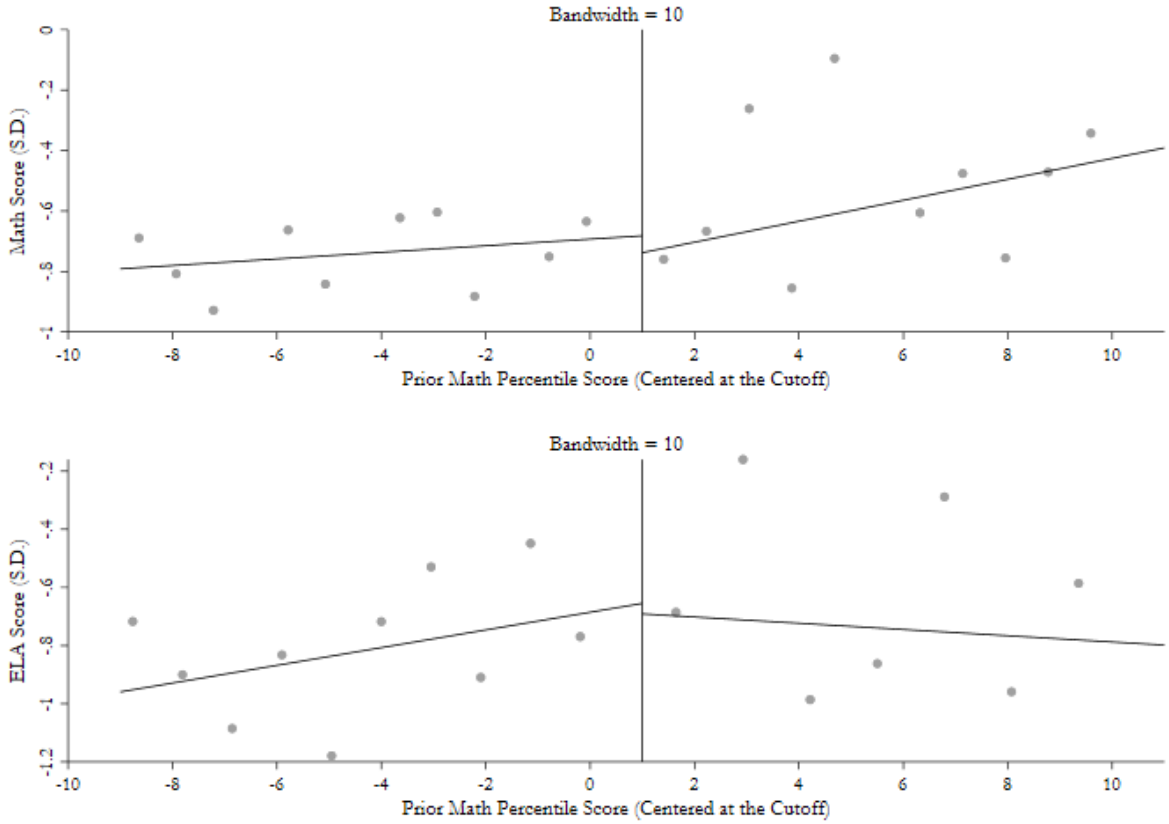
Notes: Unadjusted non-parametric regression discontinuity model with triangular kernel weighting.

Figure 3
Impact of the Modified Curriculum on Post-Treatment Objectives Completed per Hour and Ever Accessing Grade-Level Objectives in Post-Treatment Period



Notes: Unadjusted non-parametric regression discontinuity model with triangular kernel weighting.

Figure 4
Impact of the Modified Curriculum on Spring Formative Math and ELA Scores



Notes: Unadjusted non-parametric regression discontinuity model with triangular kernel weighting.

Table 3
Impact of the Modified Curriculum on Student Outcomes

	Bandwidth = 15 percentiles	
	RD Estimate	**N**
Outcomes		
Post-treatment average weekly minutes of usage	3.61 (2.38)	782
Post-treatment average pre-quiz score	0.01 (0.03)	782
Post-treatment average post-quiz score	-0.02 (0.03)	782
Post-treatment objectives completed per hour	**0.09***** (0.02)	782
Ever accessing grade-level objectives in post-treatment period	**-0.73***** (0.06)	782
Spring math std. score	0.06 (0.08)	571
Spring ELA std. score	0.04 (0.09)	571

Notes: Parametric regression discontinuity model including student demographics, pre-treatment covariates, and site-fixed effects. School-grade-clustered standard errors are in parentheses. The sample for the post-treatment usage analysis is the software usage sample (three sites). The sample for the spring test score analyses is the test score sample (two sites).
*** p<0.01, ** p<0.05, * p<0.1

Table 4
Impact Estimates for All Model Specifications

	Bandwidth (percentiles)				
	20	**15**	**10**	**5**	**Optimal (19)**
Post-treatment average weekly minutes of usage					
Parametric RD estimate	2.78 (2.25)	3.61 (2.38)	4.69 (3.40)	**7.63**** (3.77)	2.45 (2.33)
Non-parametric RD estimate	1.62 (3.24)	3.46 (3.65)	6.22 (4.26)	**10.99*** (5.62)	2.97 (3.49)
N	1,026	782	518	292	958

	Bandwidth (percentiles)				
	20	**15**	**10**	**5**	**Optimal (15)**
Post-treatment average pre-quiz score					
Parametric RD estimate	0.02 (0.02)	0.01 (0.03)	0.00 (0.02)	0.02 (0.03)	0.01 (0.03)
Non-parametric RD estimate	0.01 (0.03)	0.00 (0.03)	0.02 (0.03)	0.01 (0.05)	0.01 (0.03)
N	1,026	782	518	292	801

	Bandwidth (percentiles)				
	20	**15**	**10**	**5**	**Optimal (19)**
Post-treatment average post-quiz score					
Parametric RD estimate	-0.01 (0.03)	-0.02 (0.03)	-0.04 (0.04)	0.01 (0.05)	-0.01 (0.03)
Non-parametric RD estimate	-0.02 (0.03)	-0.02 (0.04)	0.01 (0.04)	0.02 (0.06)	-0.01 (0.03)
N	1,026	782	518	292	1,005

Continued on next page.

	Bandwidth (percentiles)				
	20	**15**	**10**	**5**	**Optimal (18)**
Post-treatment objectives completed per hour					
Parametric RD estimate	**0.09*** ** (0.02)	**0.09*** ** (0.02)	**0.11*** ** (0.02)	**0.09*** ** (0.03)	**0.09*** ** (0.02)
Non-parametric RD estimate	**0.09*** ** (0.02)	**0.09*** ** (0.02)	**0.09*** ** (0.03)	**0.08** ** (0.04)	**0.07*** ** (0.02)
N	1,026	782	518	292	958

	Bandwidth (percentiles)				
	20	**15**	**10**	**5**	**Optimal (14)**
Ever accessing grade-level objectives in post-treatment period					
Parametric RD estimate	**-0.70*** ** (0.06)	**-0.73*** ** (0.06)	**-0.72*** ** (0.07)	**-0.65*** ** (0.10)	**-0.73*** ** (0.06)
Non-parametric RD estimate	**-0.66*** ** (0.06)	**-0.64*** ** (0.07)	**-0.57*** ** (0.09)	**-0.53*** ** (0.12)	**-0.63*** ** (0.08)
N	1,026	782	518	292	696

	Bandwidth (percentiles)				
	20	**15**	**10**	**5**	**Optimal (13)**
Spring math std. score					
Parametric RD estimate	0.11 (0.07)	0.06 (0.08)	0.08 (0.08)	0.10 (0.12)	0.05 (0.08)
Non-parametric RD estimate	0.04 (0.08)	0.02 (0.09)	0.00 (0.10)	0.07 (0.14)	0.01 (0.09)
N	737	571	377	216	506

Continued on next page.

	Bandwidth (percentiles)				
	20	**15**	**10**	**5**	**Optimal (12)**
Spring ELA std. score					
Parametric RD estimate	0.09 (0.07)	0.04 (0.09)	-0.05 (0.10)	-0.18 (0.15)	-0.04 (0.09)
Non-parametric RD estimate	0.00 (0.07)	-0.05 (0.09)	-0.09 (0.11)	-0.19 (0.14)	-0.09 (0.10)
N	737	571	377	216	438

Notes: Parametric and non-parametric regression discontinuity models include student demographics, pre-treatment covariates, and site-fixed effects. Non-parametric model includes triangular kernel weights. School-grade-clustered standard errors are in parentheses. Optimal bandwidths for each outcome are determined using a mean squared error method (Calonico et al. 2014). The sample for the post-treatment usage analysis is the software usage sample (three sites). The sample for the spring test score analyses is the test score sample (two sites).
*** p<0.01, ** p<0.05, * p<0.1

Appendix

Figure A1
Impact of the Modified Curriculum on Post-Treatment Average Weekly Minutes of Usage at Different Bandwidths

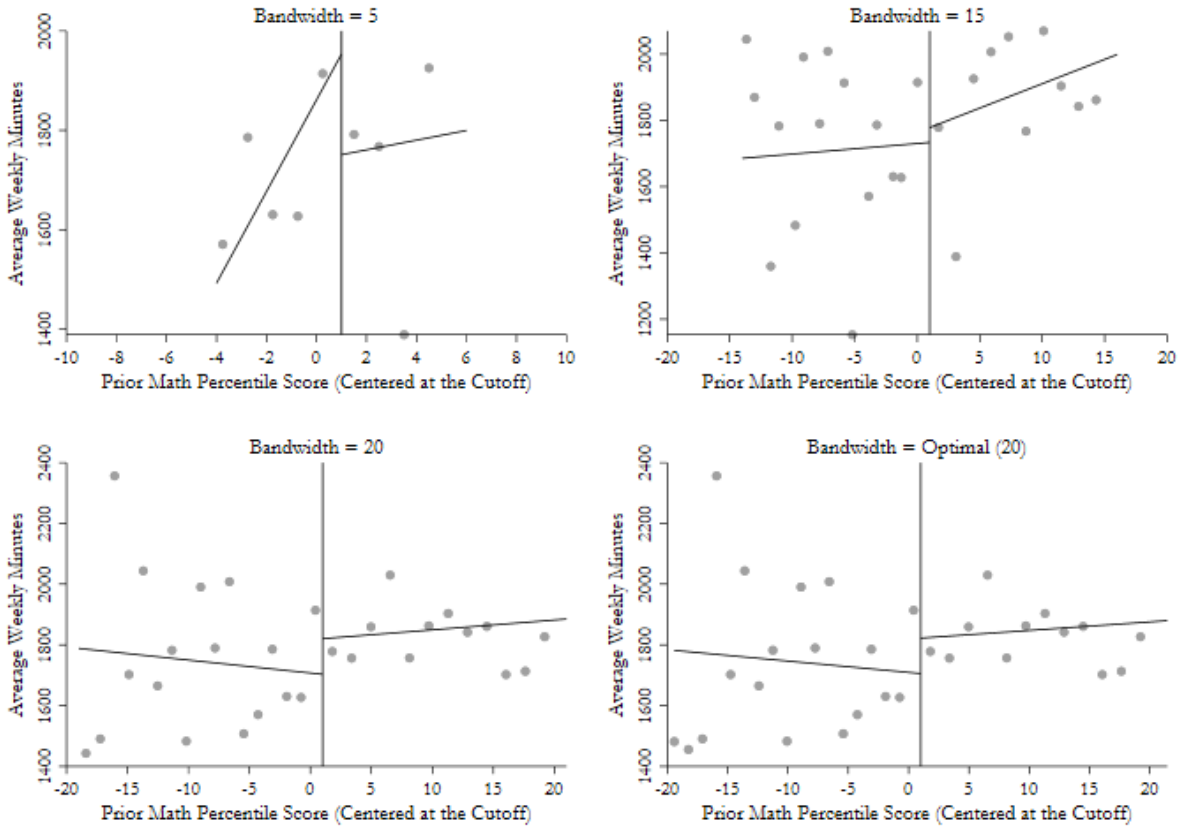


Figure A2
Impact of the Modified Curriculum on Post-Treatment Average Pre-Quiz Score at Different Bandwidths

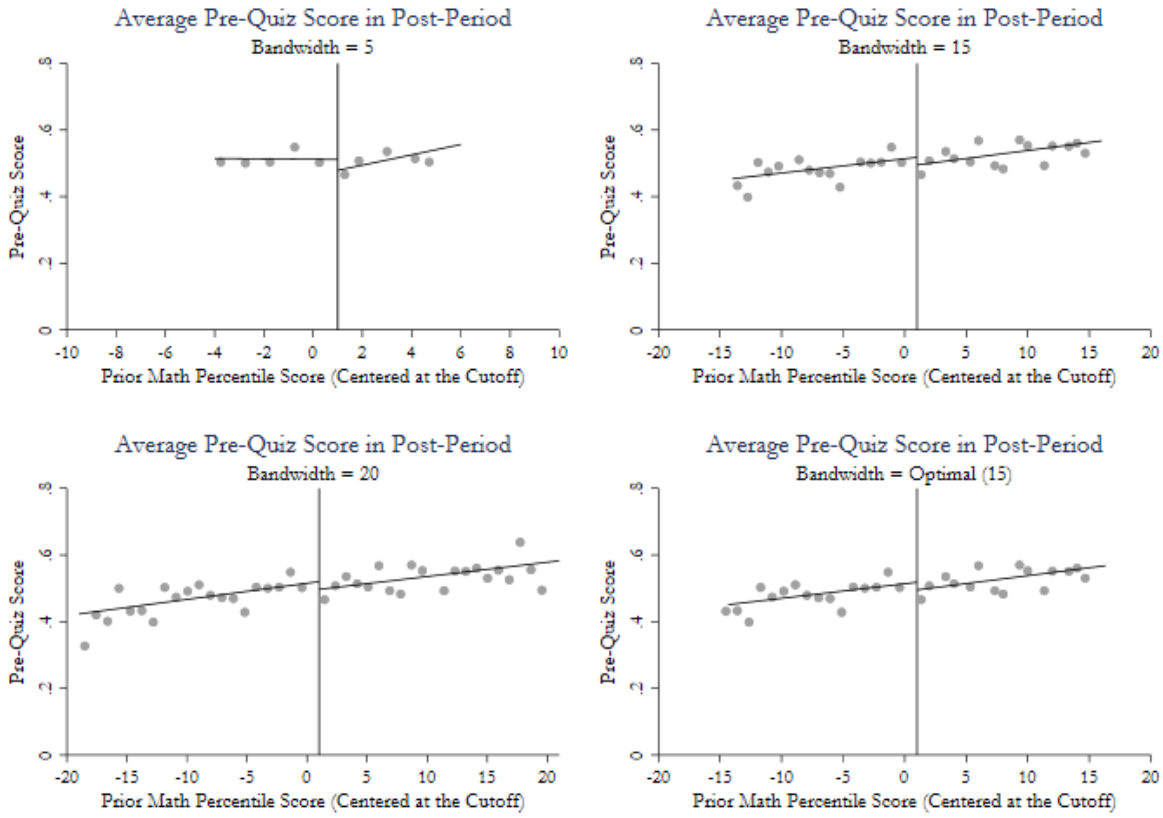


Figure A3
Impact of the Modified Curriculum on Post-Treatment Average Post-Quiz Score at Different Bandwidths

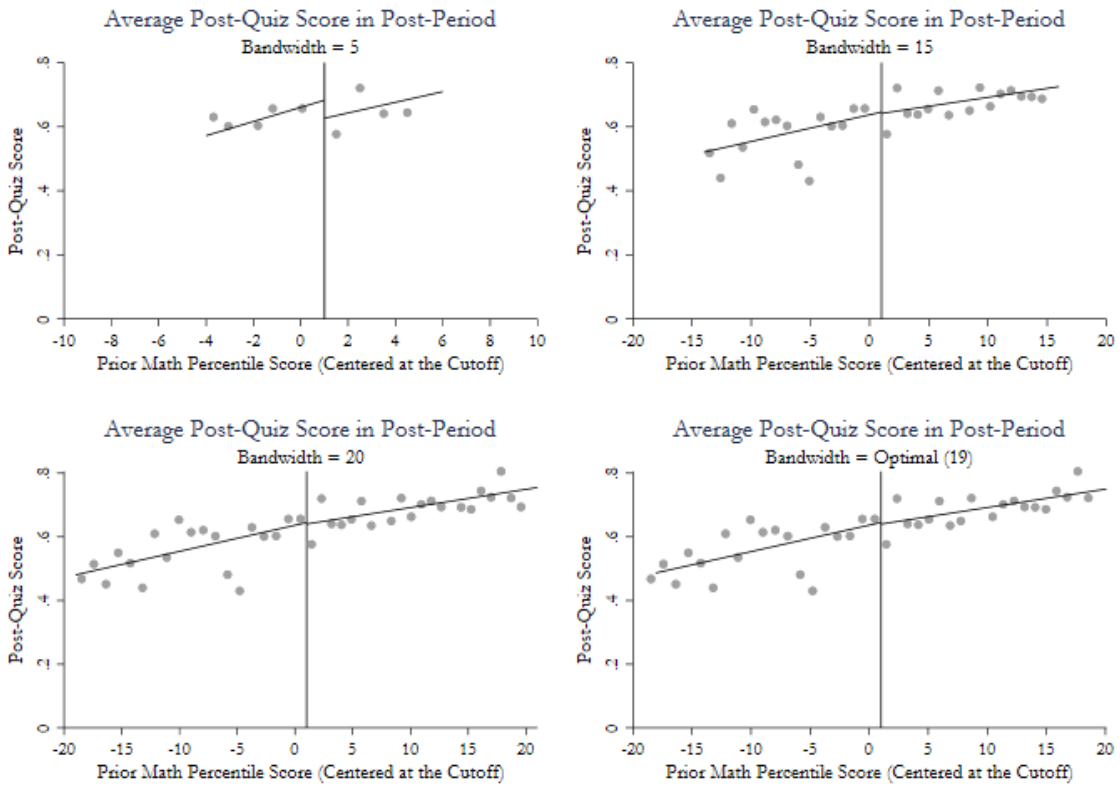


Figure A4
Impact of the Modified Curriculum on Post-Treatment Objectives Completed per Hour at Different Bandwidths

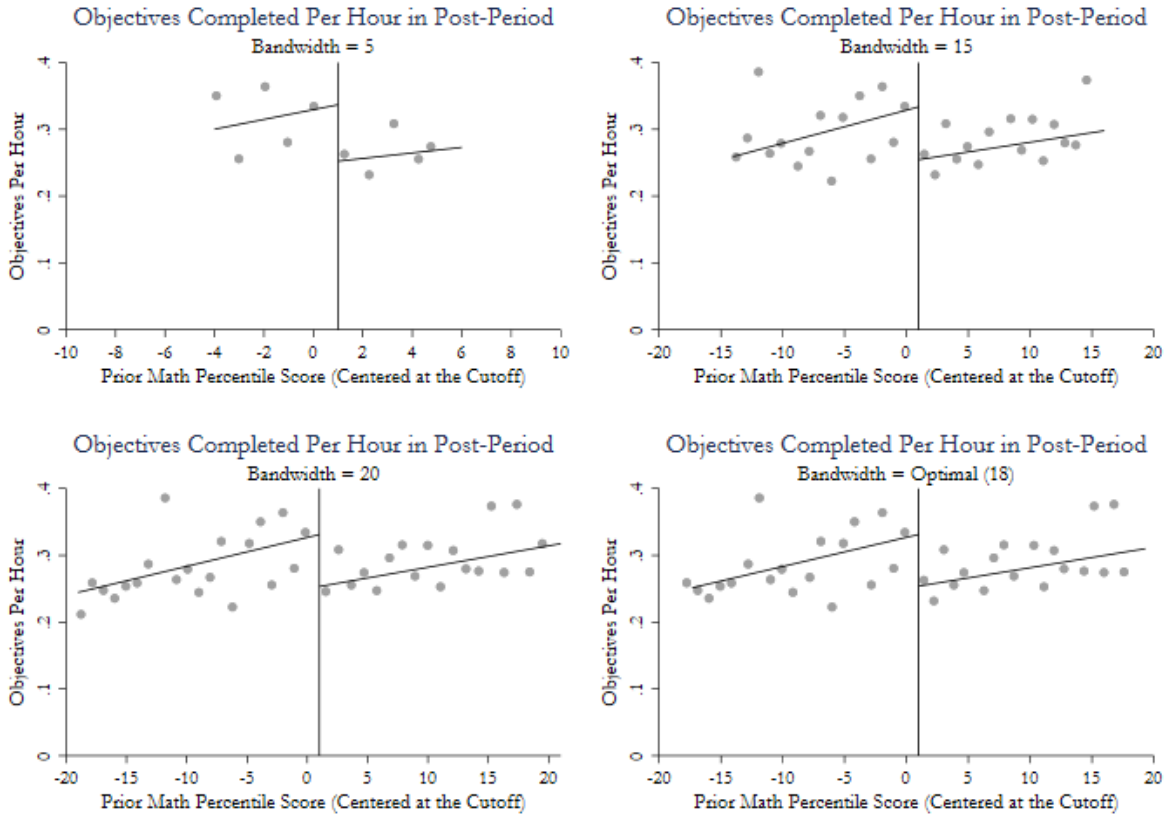


Figure A5
Impact of the Modified Curriculum on Ever Accessing Grade-Level Objectives in the Post-Treatment Period at Different Bandwidths

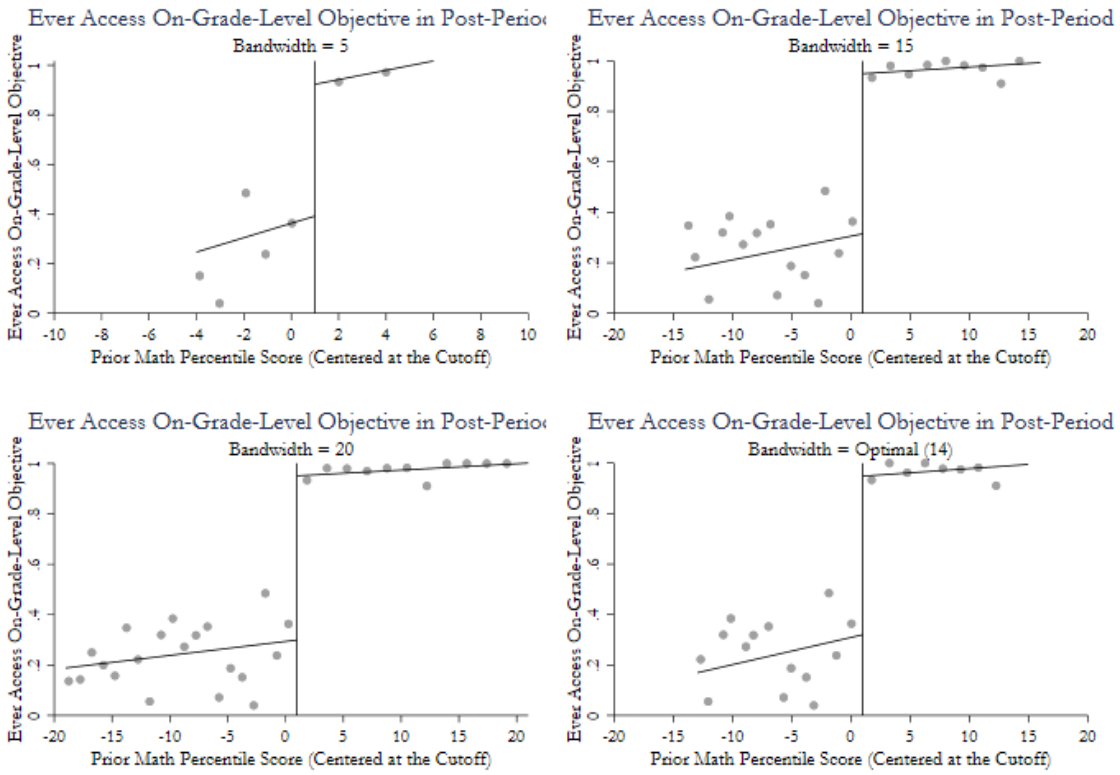


Figure A6
Impact of the Modified Curriculum on Spring Formative Math Scores at Different Bandwidths

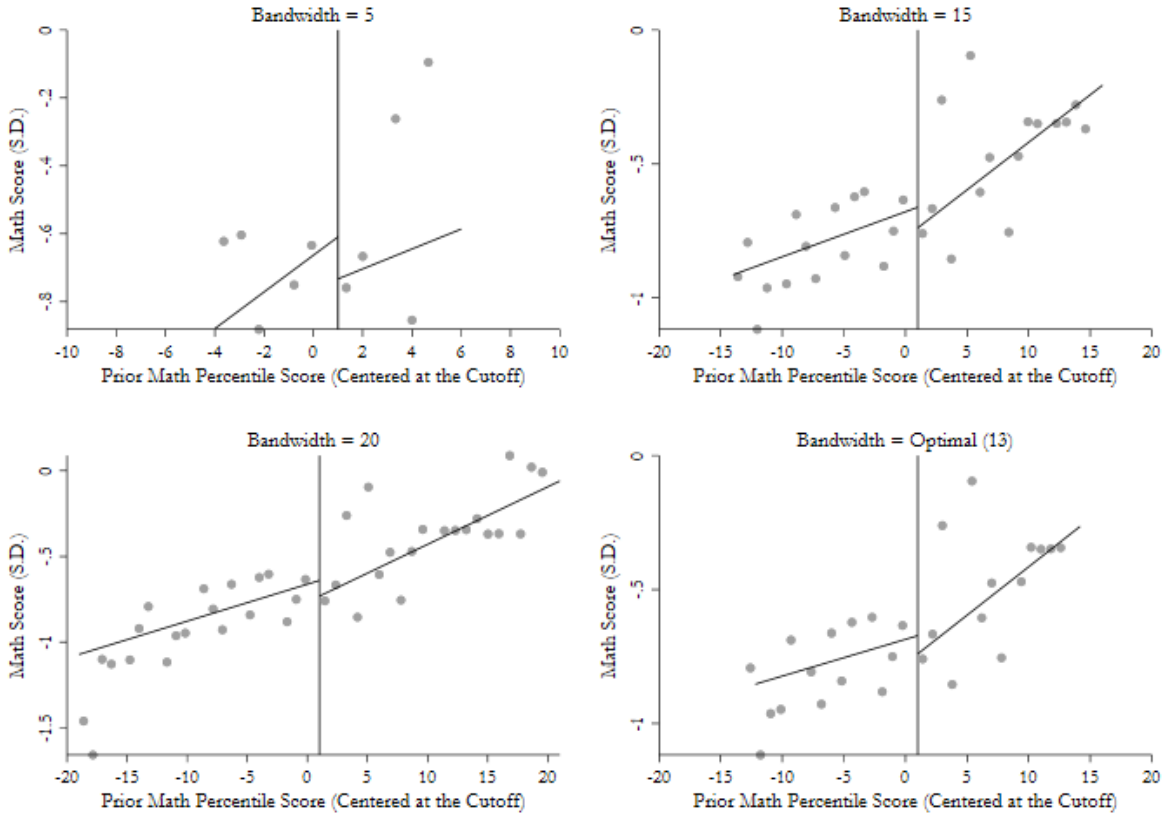


Figure A7
Impact of the Modified Curriculum on Spring Formative ELA Scores at Different Bandwidths