Efficacy of Zearn Math over two years in grades 3 to 5: An experiment in Texas

John F. Pane, Christopher Joseph Doss, Ivy Todd, Dorothy Seaman

RAND Corporation

Abstract: Zearn Math is a popular software platform for K-8 mathematics learning, designed to enable all students to successfully access grade-level content. RAND researchers collaborated with Zearn, the product's developer, to design this evaluation. Then RAND conducted the study independently, randomly assigning 64 schools in an urban Texas district to either supplement classroom instruction with Zearn Math in grades 3-5 for two years – or to continue with business-as-usual, which included various other supplemental technology products. High proportions of economically disadvantaged, Hispanic, English-learner, and below-proficient students made up the primary sample of 10,000+ students. The study preregistered two confirmatory research questions about Zearn Math's effects on Texas STAAR math assessment scores, for all students and students below proficient at baseline. Those results were positive but not statistically significant; equivalent to raising a control group student from the median to the 53rd or 54th percentile. Although this study did not yield confirmatory evidence that Zearn Math improves student learning, consistent positive signals across all estimated confirmatory and exploratory effects, including on the MAP adaptive mathematics assessment, suggest it holds promise to do so.

Introduction

Zearn Math is a popular platform for K-8 mathematics learning from Zearn, a non-profit organization. Zearn Math is offered as either a full curriculum, or schools can use the software as a complement to an existing curriculum. This study focuses on its use as a complement. Zearn designed the software to synchronize with the grade-level material currently being covered during teacher-led instruction, and when students struggle, to provide individualized, targeted remediation to scaffold learning of the grade-level material. This contrasts with other adaptive mathematics learning products that prioritize filling gaps in knowledge irrespective of grade level or the material currently being taught in class, to build a foundation of mathematics skills.

Although Zearn Math has shown promise in quasi-experimental evaluations discussed below, its efficacy for improving student math achievement has not yet been established in a well-powered, rigorous evaluation. A team of researchers from the non-profit RAND collaborated with Zearn to design a randomized experiment to evaluate Zearn Math's effects over two years on student mathematics achievement in grades 3 to 5. The RAND team then received a grant from the Institute of Education Sciences, U.S. Department of Education, to conduct the study independently of Zearn. This article focuses primarily on the design and implementation of the experiment and the quantitative student achievement results, along with a limited discussion of implementation. Future articles will address implementation in greater depth and examine cost and cost-effectiveness.

Background and context

The need to address lagging mathematics achievement in the U.S.

Lagging math achievement among K-12 students continues to concern educators and policymakers in the U.S. The focus on math is often justified, in part, by the modern economy's reliance on science, technology, engineering, and math (STEM) fields, all of which require a

foundation in math. Historically, the United States has performed relatively poorly on international assessments which has sparked fears that the United States may lose its competitive advantage in the global economy (Committee on STEM Education, 2018).

The COVID-19 pandemic has only exacerbated the concern for U.S. math achievement. Results from the 2022 National Assessment of Educational Progress (NAEP) indicated that decades of progress in mathematics achievement were lost to the pandemic (NAEP, 2022). Recent data from the 2024 NAEP revealed that setbacks remain: only 39 percent of fourth grade students were proficient in math, up three percentage points from 2022 but still two percentage points below pre-pandemic levels. Eighth grade results were even more concerning: 28 percent of students were proficient, up two percentage points from 2022 but still six percentage points below pre-pandemic levels (NAEP, 2024). These average statistics obscure wide disparities by race and socio-economic status. On the fourth grade NAEP, compared to their White or more-affluent peers, 32 percent fewer Black students, 24 percent fewer Hispanic students, and 31 percent fewer lower-income students were proficient; these gaps were 28, 23, and 27 percent in eighth grade (NAEP, 2024). Other large-scale assessment data confirm these trends. For example, during the pandemic NWEA MAP math achievement decreased as much as 0.26 standard deviations (SD) in math, gaps by race/ethnicity and socio-economic status widened, and recovery from these setbacks has been minimal (Lewis & Kuhfeld, 2022).

Late elementary school is a promising time to field interventions to improve math achievement. It is a critical juncture when students often struggle to master more complex number concepts such as the shift from whole numbers to fractions, decimals, percentages, ratios, and proportions (National Mathematics Advisory Panel, 2008; McMullen et al., 2015). As math instruction is highly sequential and mastery of the next concept often builds on mastery of

prior concepts, deficiencies in early math knowledge accumulate as students progress to middle school and beyond. For example, fractions that are first taught in third grade become foundational for success in algebra and the engineering and science classes that apply math to describe and understand natural phenomena (Booth and Newton, 2012; Siegler et al., 2012, Powell et al., 2019).

The promise of technology to improve mathematics achievement

Education technologies are often touted as a solution in situations where students are lagging academically and the variation in student skills in a classroom can be substantial. Supporters argue that well designed products can present material in novel and visual ways to increase student engagement and understanding. Computer adaptive technologies can home in on a student's skill level to differentiate instruction such that students struggling with more basic concepts and students ready for advanced concepts can both be presented with challenging material (Chatterji, 2018).

Escueta et al. (2020) found that this type of adaptive software is one of few types of education technologies that have relatively consistently generated positive effects in rigorous studies. However, students can experience these benefits only if they receive enduring opportunities to engage with the software and take up those opportunities. Field tests of such products often find they are used at very low dosage levels (e.g., Ikemoto et al., 2016; Karam et al, 2017; Phillips et al., 2020). Moreover, the students who stand to benefit most from the software tend to use it the least and receive the smallest benefits, if any. Holt (2024) calls this the "5 percent problem," named after the proportion of students who do receive high dosage, who also tend to already be high achievers.

The strategic question of how best to address gaps in student learning

Adaptive supplemental technology products generally take one of two approaches to help students address gaps in their knowledge of material taught in earlier grades. We refer to these as a ***foundational approach*** or a ***grade-level approach***.

Where gaps exist in student mastery of below-grade-level concepts, software taking a foundational approach tends to focus on systematically filling those gaps, to build a strong foundation upon which students can subsequently learn more complex topics. The approach has theoretical merit in the long run because filling gaps in more basic concepts can facilitate learning higher-level mathematics concepts that build upon them (e.g., Sweller, 1994; Pollock, Chandler, & Sweller, 2002). However, the approach does not directly support engagement with grade level content, potentially hindering students' opportunity to learn that content. Student work in the software might focus for the entire year on below-grade-level content that is not strongly connected to the topics of classroom instruction. Yet, the products that earned adaptive learning software its favorable status in Escueta's (2020) review used the foundational approach, and it remains the most common approach among products available today.

More recently, there has been growing attention in the field of education to focusing on grade-level content for all students, even those who are far behind, and providing scaffolding and support only to the extent needed to help each student accomplish grade-level work. This grade-level approach is sometimes referred to as acceleration (Lambert & Sassone, 2020; TNTP, 2021). The approach has theoretical merits of greater coherence between work within the software and in class, and greater equity in supporting all students to engage with grade-level work. But there is an assumption that the targeted remediation is sufficient to enable students to accomplish the grade-level work, even those with substantial gaps in more fundamental knowledge. Whereas software taking a foundational approach prioritizes mastery of a topic before moving to the next,

software taking a grade-level approach might switch to a new topic to keep up with classroom instruction even if the student has not finished the prior one.

Beyond the specifics of what content students work on, both approaches pose challenges for teachers that may affect how much time they allocate to the use of supplemental adaptive software. As previously mentioned, sufficient dosage has been a longstanding challenge to fully realizing the potential benefits of foundational-approach products. Teachers may limit dosage of such software if they have concerns about the dispersion of the content students are working on, or that many are not covering the grade-level material that will be tested on state accountability tests. Software using a grade-level approach could potentially mitigate those concerns; however, teachers using those products might limit dosage if they have concerns that focusing on grade-level work is not productive for students with very large gaps in prerequisite knowledge. The relative advantage of the two approaches at scale will likely depend not only on differences in content focus but also on differences in dosage achievable over extended timespans.

To our knowledge, this study is the first to rigorously evaluate an adaptive supplemental technology product, namely Zearn Math, that uses a grade-level approach. However, the study was not deliberately designed to isolate a contrast between the grade-level and foundational approaches. The grade-level approach is one of a composite of features that make up Zearn Math and are evaluated against a business-as-usual control group in this study. Nonetheless, as will be discussed later in this article, the business-as-usual control group used a variety of adaptive supplemental technology products that generally follow the foundational approach. These products were also available to the treatment group, although Zearn Math was the dominant product used there during the study. Thus, the difference in approaches may contribute to any group differences measured by the study.

The Zearn Math intervention

Zearn Math consists of independent digital lessons for students, digital and printed instructional materials, data and reports on student performance, and training and implementation support for teachers and administrators. It covers mathematics content for grades K-8, and provides guidance and support to align with mathematics standards in all U.S. states. It is primarily used as a digital supplement to an existing curriculum. Its lesson sequence and instructional approaches align with the Eureka Math curriculum in K-5 and Illustrative Math in 6-8, but it can be used as a supplement to any curriculum. The current study evaluates the efficacy of Zearn Math as a digital supplement in grades 3-5.

About 10 percent of teachers in a 2024 nationally representative survey endorsed Zearn Math among materials they used at least weekly (Schweig et al., 2024).

Digital lesson structure, sequencing, and recommended dosage

Teachers using Zearn Math assign lessons for students to complete on the platform. Lessons cover every concept of K-8 math. Upon logging in, students engage in guided lesson activities designed to enhance their understanding of mathematical concepts and procedures. All Zearn Math lessons include fluency games, videos led by on-screen teachers with pause points to solve math problems, and a closing mastery-based quiz called a Tower of Power. When students make a mistake in the Tower of Power, they receive targeted feedback and support to help them try again with a new problem. They get as many attempts at the Tower of Power as they need to demonstrate their mastery with a score of 100 percent, after which they can advance to the next lesson.

Zearn Math organizes instructional content by grade level. Each grade level consists of several “Missions,” or units, which each comprise several digital lessons. The software’s default settings begin students on Mission 1, Lesson 1. After students demonstrate understanding by

completing the lesson's Tower of Power, they progress to subsequent lessons and missions in sequential order (i.e., first through all the lessons of Mission 1 in order, then of Mission 2, and so on). As part of its grade-level approach, Zearn recommends that teachers keep students' digital lesson sequence assignment aligned with live instruction at the Mission level. For example, if whole class instruction progresses from Mission 1 to Mission 2 (or the equivalent unit in a non-Zearn curriculum), Zearn would recommend teachers re-assign all their students to the beginning of Mission 2, regardless of whether they have completed all the lessons from Mission 1.

Teachers can override the software's default lesson assignments. They can "assign" students to a different place in the curriculum, which resets students' default lesson path. For example, if a teacher assigned a fourth-grade student to Grade 3, Mission 2, Lesson 1, that student would follow the default progression from that point. Alternatively, teachers can "bookmark" lessons for students, which allows students to complete individual lessons outside of the assigned sequence but maintain their overall placement in the sequence. If the same example teacher had bookmarked one lesson from Grade 3, Mission 2, that student would first be directed to the bookmarked lesson, then upon completion would resume the lessons from the 4th grade sequence.

Zearn recommends students spend 90 minutes on Zearn Math and complete three on-or-above-grade-level digital lessons per week, adding up to at least 90 such lessons over the course of a school year.

Implementation support

Zearn provided implementation support to the treatment group schools throughout the study period. They offered beginning of the year trainings to teachers, instructional leaders, and building leaders in treatment group schools, which provided staff with an overview of the software and its features.

Zearn also provided implementation coaching to each treatment group school, which typically comprised bi-weekly calls between a Zearn coach and the school's implementation point-of-contact to review school usage data, highlight areas of success, and discuss any challenges hindering usage.

During the spring of each study year, Zearn facilitated challenges to motivate on-or-above-grade-level lesson completion. These challenges offered prizes to students, classrooms, and schools based on lesson completion. Zearn also shared print materials for bulletin boards or other common spaces that schools could use to celebrate Zearn Math usage.

Prior research on Zearn

As of April 2025, the What Works Clearinghouse (WWC) had not reviewed any studies of Zearn Math and had not rated its effectiveness as a product. Much of the prior evidence on Zearn Math is derived from observational or quasi-experimental studies. Morrison et al. (2019) studied Zearn Math's adoption in a subset of schools in a large urban district. Fifteen elementary schools opted to adopt Zearn Math in Grades 1-5 and were compared to non-adopting schools. Despite largely positive experiences with Zearn Math reported by teachers, HLM models estimated only a small and statistically insignificant relationship between Zearn Math use and math achievement on the MAP assessment. This lack of relationship was echoed for the state standardized assessment in math. However, the authors did find a strong correlation between using the Zearn platform for high amounts of time and assessment outcomes. Key limitations of this study are the convenience samples of adopters and students who used Zearn Math for high amounts of time, because controlling for observables cannot mitigate bias from unobserved variables.

A stronger differences-in-differences, cross-cohort design investigated Zearn Math adoption in Louisiana. Hashim (2024) limited the treatment sample to schools where at least half of

students used the Zearn Math software for any amount of time on at least half of the school year's instructional days. Hashim estimated a small but significant increase in state standardized test scores of 0.03 SD after one year of Zearn Math use. She also found suggestive evidence that Zearn Math's effects can grow with additional years of use. However, restricting the analysis to schools and students with higher usage reduces the generalizability of the results to the general population of Zearn Math schools and students who may use it less frequently. Moreover, estimates may have been biased as the author did not include school-specific secular achievement trends in her models.

Finally, Zearn has conducted its own research exploring how meeting Zearn Math usage goals (described earlier in this article) could be associated with math outcomes. Using data from Texas, Nebraska, and the District of Columbia, the authors matched students who met usage goals to students who had access to Zearn Math but did not meet those goals. The studies found that the groups meeting goals outperformed their peers on the Texas state assessment by 0.33 SD (Zearn, 2024b), the Nebraska state assessment by 0.40 SD (Zearn, 2022), and the iReady assessment by 0.27 SD (Zearn, 2024a). A key assumption of matching methods is that forming a comparison group equivalent to the treatment group on all observed variables makes them likely to be equivalent on unobserved variables, so that the only meaningful difference between the two groups is whether they received the treatment (in this case, meeting Zearn Math usage goals). This assumption is not likely to hold in these studies because there are observed differences in student behavior (the use of Zearn Math) between the treatment and comparison students after matching. These difference in behavior likely stem from differences in unobserved student, teacher, and school characteristics. For example, students who use Zearn Math more may be more motivated or in schools or classrooms that have differing resources or approaches to

learning math. These unobserved variables likely would lead to different student outcomes even in the absence of Zearn Math. Thus, it is likely the estimated differences in outcomes in these studies are upwardly biased estimates of the effect of meeting usage goals.

Researchers have conducted two small, randomized controlled trial evaluations of Zearn Math. First, Klopfenstein (2018) randomly assigned Zearn Math in either grades K, 2, and 4 or 1, 3, and 5 in six elementary schools in Colorado. This alternating grade design allowed each school to contribute to the treatment and control sample. The study analyzed SchoolCity assessments administered at baseline and end-of-year, estimating that pre-post gains in the treatment group were 0.16 SD larger than in the control group, though not statistically significant. We calculate that study was powered to detect effects of 0.34 SD or larger. More recently, Foster (2024) compared the effects of Dreambox Learning and Zearn Math in the southeastern U.S. by randomly assigning 112 students in classes taught by six teachers to one or the other product. Only 58 participating students completed posttests. The study did not detect a statistically significant difference in effects on math achievement as measured by the Research-based Early Math Assessment (REMA), although point estimates favored Dreambox Learning.

Overall, this body of evidence suggests potentially promising effects of Zearn Math, with the study most similar to this one showing potential effects as large as 0.16 SD. As can be expected, levels of Zearn Math usage are likely to be an important mediator of its effect.

Study setting

This study took place in a large urban school district in Texas. Sixty-four schools participated over a two-year period. All participating schools served students in grades 3-5. Most (n=57) enrolled pre-kindergarten through 5th grade students, and the remaining campuses served other grade configurations including PK-8 and K-5.

Between study years 1 and 2, the district underwent a change in administration. The new administration implemented several reform initiatives, which impacted district and school-level operations. These changes likely impacted the implementation of supplemental math technology products in both the treatment and control groups of this study.

One new initiative of the incoming administration was the creation of a school reform cohort. Reform cohort schools implemented instructional and organizational changes, which included new practices for scheduling, staffing, compensation, and performance appraisal for teachers and school leaders. These schools implemented a centrally mandated uniform schedule to organize and allocate instructional time. Teachers were required to regularly integrate specific practices in their instruction. School and district leaders conducted frequent walk-throughs and formal observations to monitor classroom teaching and to enforce implementation of the specified practices. One-fourth of the study schools (n=16; 9 treatment, 7 control), were in the reform cohort in year 2 of the study; they continued to comply with their assigned experimental condition.

Methods

Research Questions

We designed the study in collaboration with Zearn staff to address the following research questions about student outcomes:

RQ1. (Confirmatory) What is Zearn Math's cumulative effect over two academic years on student achievement on grade-level mathematics content, as measured by the Texas STAAR assessment?

RQ2. (Confirmatory) What is this cumulative two-year achievement effect for the subgroup of students who scored below grade-level proficiency at baseline?

RQ3. (Exploratory) What are the cumulative two-year achievement effects for other student subgroups based on grade level, economic disadvantage, race/ethnicity, gender, or English learner (classified in Texas as emergent bilingual learner)? What are the one-year effects?

RQ4. (Exploratory) What are the effects of Zearn Math on mathematics achievement growth, as measured by the MAP adaptive assessment, overall and for the above-listed subgroups?

We also posed questions about implementation, cost, and cost effectiveness, which we will address in future publications. In lieu of full implementation information, this paper includes a limited set of findings related to treatment group fidelity of implementation, and instructional contrasts between experimental groups.

Rationale and implications for confirmatory and exploratory designations

To increase transparency and credibility of this study and meet funder requirements, we clearly stated our primary research questions by designating the first two as confirmatory and preregistering them along with our analysis plan at the Registry of Efficacy and Effectiveness Studies (REES, undated), under Registry ID: 17280.1v1. We specified that we would apply the Benjamini-Hochberg (1995) method to control for false discovery for the two confirmatory tests. Had we included more analyses among the confirmatory tests, the statistical power of all of them would have been penalized by this method. Designating the remaining quantitative analyses as exploratory avoided excess power loss while retaining the ability to maximize learning from this study, with the caveat that exploratory results are not considered conclusive, even if statistically significant, until confirmed in a future study.

The STAAR assessment, as the state accountability test, is designed to measure student proficiency on Texas grade-level mathematics standards starting in grade 3. In collaboration with

Zearn we chose the STAAR as the outcome for confirmatory RQs 1 & 2, because it has high relevance for administrators, educators and families, and its focus on grade-level content might be well-suited to capturing effects of Zearn Math's grade-level approach to supporting all students to succeed at learning grade-level content.

Research design

Sixty-four schools within one urban school district in Texas participated in this study over the 2022-2023 and 2023-2024 school years. The district initially identified 128 elementary school campuses eligible for inclusion in the study. We randomly selected 64 of those schools to comprise the study sample. We used blocked randomization, stratifying by schoolwide percentages of students who were economically disadvantaged, English-learners, and proficient in mathematics in the 2020-2021 academic year. Within each block, half of the schools were assigned to the treatment group and half to the control group, resulting in 32 schools assigned to each condition.

Treatment group schools gained access to Zearn Math, received training for building leaders, instructional leaders, and teachers, and were asked to integrate Zearn Math into their instructional practice for the two-year study period. Control group schools were asked *not* to use Zearn Math and to continue with business as usual for the two years. In partnership with Zearn, the research team periodically monitored usage at treatment and control schools to ensure adherence to treatment group assignment.

Study population

The study's primary sample is defined to support the confirmatory research questions and most exploratory questions, which address two-year exposure to Zearn Math. This sample includes students in study schools who were enrolled in grades 3 or 4 at the beginning of the study and normally would have progressed to grades 4 or 5 the second year, and who did not

have significant cognitive disabilities. The sample was determined by enrollment on the date of the state's official enrollment snapshot at the beginning of study year 1 (October 28, 2022). No students were accepted into the primary sample after this date, i.e., no late joiners. The excluded students were the approximately two percent of students who met state eligibility for the alternate state mathematics assessment the following spring. State regulations specify that assessment is intended for students with the most significant cognitive disabilities who are receiving special education services. These students do not participate in general education classrooms where Zearn Math was implemented and have individualized education plans that control the instructional resources provided.

The primary sample included 10,577 students after excluding 235 students with significant cognitive disabilities. In the primary sample, 67 percent of students were Hispanic and 20 percent were Black, 83 percent were economically disadvantaged, 45 percent were English learners, and 9 percent received special education services. At the start of the study, 60 percent of students scored below grade-level proficiency on end-of-year mathematics assessments from the prior school year; 11 percent did not have scores reported from the prior year, indicating they were likely new to the district in year 1 of the study. See Table 1 for additional descriptive and group balance statistics for the primary sample.

We defined the teacher sample to include the mathematics teachers of record for the students in the sample. Teacher of record was operationalized as the individual educator responsible for assigning grades in the student's core mathematics course as of the district's fall enrollment snapshot; co-teachers, pull-out coaches, or aides were not included. As such, all teachers in the sample taught math to students in at least one of grades 3, 4, or 5. For the primary student sample, there were 343 teachers the first year and 502 teachers the second year, before

accounting for students without a teacher of record in the administrative data. Teacher counts were higher the second year because approximately 3 percent of primary sample students enrolled in a non-study school within the district that year. Across both years, teachers of primary sample students in study schools were 76 percent female, 39 percent Hispanic or Latino, 34 percent Black or African American, 20 percent White, and had an average of 9 years of teaching experience.

A secondary sample for exploratory one-year analyses also includes students enrolled in grade 5 in year 1, and a new cohort of grade 3 students who enrolled at the beginning of year 2. The analogous rules for determining enrollment and exclusion apply to this sample. This secondary sample included 20,996 students after excluding 506 students with significant cognitive disabilities.

Informed consent

After RAND selected the 64 schools for the study and before they were randomized to experimental groups, the district gave building leaders an opportunity to opt out of the study; none opted out. Because the study collected no identifiable student data, RAND's Institutional Review Board and the district's Research Review Board did not require us to obtain parental consent for study participation. We notified parents of their child's inclusion in the study by mailing notification letters to the study schools to send home with students. School staff who participated in surveys and interviews underwent informed consent procedures at the time of data collection.

Measures and data sources

As previously mentioned, the primary achievement outcome in this study is the STAAR state accountability test. Where available, the STAAR is also used as a baseline achievement measure. For students who were in grade 2 at baseline we use the Renaissance Star, a district-administered

adaptive standardized assessment. We also use the Renaissance Star for some exploratory analyses of year 1 outcomes, and the similar MAP for some exploratory analyses of year 2 outcomes.

The district used Renaissance Star and MAP for formative purposes, e.g., to help educators gain information on student strengths and weaknesses to guide instruction. Nonetheless, these assessments can estimate precise achievement scores across a broad range of mathematics content spanning kindergarten to high school. They accomplish this by using each student's performance on prior items to determine which subsequent items to present, drawing from large banks of items of varying grade level and difficulty. The adaptive nature of these assessments can make them more sensitive than STAAR to student learning of below- or above-grade-level content – learning that may be facilitated by Zearn Math and the products used by the business-as-usual control group in this study.

The school district provided de-identified administrative student data, which included date of enrollment at current school, date of school exit, grade level, prior math achievement, and demographic information: economic disadvantage, gender, age or date of birth, race/ethnicity, language learner status, gifted status, special education status. Each student in this dataset was assigned a unique persistent study identifier by the district research office to enable linkage across various data files and across years.

For prior math achievement of students enrolled in grades 4 or 5 at the start of the study, the district provided prior-year Texas STAAR Math assessment scores from grades 3 or 4, respectively. For these students, we determined below grade-level proficiency by not meeting or exceeding the state's "meets grade level" performance category threshold. Students enrolled in grade 3 for the study did not have a prior STAAR score because STAAR is not administered in

grade 2. For these students, the district provided prior-year Renaissance Star mathematics scores from the grade 2 end-of-year administration. Absent performance categories or linking studies for Renaissance assessments in grade 2, we used national percentile rankings on that assessment (Renaissance Learning, 2023), benchmarked against STAAR statewide percentile rankings. On the 2022 STAAR assessment, students enrolled in grade 3 who scored below the 74th percentile and students in grade 4 who scored below the 69th percentile did not meet grade level proficiency. Aligning with the values and trend of these STAAR statewide percentile rankings, we considered students who scored below the 75th percentile nationally on the Grade 2 Renaissance assessment to have scored below grade-level proficiency. Sensitivity tests indicated that outcomes analyses for the below-proficient subgroup were not substantively affected by the precise threshold we chose.

The Zearn Math software collected identifiable data on student and teacher use of the software. This included information such as the number of days and amount of time each student used Zearn Math, which digital lessons they worked on and completed, and indicators of struggle, as well as teacher uses of Zearn Math reports. Zearn sent this software log data to the district, which replaced the real student identifiers with the study identifiers to create a version of the dataset without identifiable student information but linkable to the student administrative data we received. They forwarded this deidentified dataset to us through a secure file transfer platform.

Data to inform implementation included software use data captured by the Zearn Math platform across the two study years, annual spring surveys of teachers in both experimental groups, and annual case study interviews at six treatment-group schools.

To gauge the extent to which students in the sample met recommended usage thresholds, we summarized platform data on time spent and lesson completion while using Zearn Math. For each semester of the study, we calculated average weekly time on the platform and tallied the sum and percent of students whose average weekly usage met Zearn's recommended weekly use target of 90 minutes. We also tallied the number and percent of treatment group students who met or exceeded Zearn's recommend lesson completion goal of 90 on-or-above-grade-level lessons.

Annual teacher surveys asked about instructional practices and technology use in math instruction. This included how teachers allocate math instructional time, the activities and strategies they employ in their math instruction, their data use practices, and their experiences with professional development. Additional questions sought to understand whether and how teachers in both groups use supplemental math technology products in their instruction, barriers they face, and their perceptions of the quality and utility those products. In year 1, we asked treatment and control group teachers about the supplemental math technology product they reported using most frequently. In year 2, the survey was modified to ask all treatment group teachers about Zearn Math, even the 12 percent who reported using a different product most frequently. In spring 2023, 222 respondents to the survey (55 percent of eligible teachers) included 129 treatment-group teachers (58 percent) and 93 control-group teachers (42 percent). In 2024, 323 responses (72 percent of eligible teachers) consisted of 172 treatment-group teachers (76 percent) and 151 control-group teachers (68 percent).

We also conducted annual case studies of six treatment group campuses, where we conducted semi-structured interviews of teachers and instructional leaders about their experiences using Zearn Math. We intentionally selected the six sites to represent diversity in school demographic

characteristics, baseline achievement, and early Zearn Math usage levels. However, this small sample is unlikely to be representative of the full sample of treatment-group schools. We use this data to gain richer insight that can complement the more systematic but rigid data collected through surveys.

The self-report data collected with surveys and interviews have additional limitations. Social desirability bias can arise if respondents over-report desirable practices and under-report undesirable practices. Relatively low survey response rates the first year, and differential response rates between experimental groups, can reduce the representativeness of the responses received.

Finally, we gathered contextual information about district policies in occasional conversations with staff members of the district's central office and of Zearn.

Analytic Methods

Quantitative analyses of student outcomes data

A model of the following form was used to analyze whole-sample student level outcomes:

$$Y_{itsd} = \beta_0 + \beta_1 Z_{sd} + \boldsymbol{X_{itsd}\beta_2} + \boldsymbol{T_{tsd}\beta_3} + \boldsymbol{S_{sd}\beta_4} + \gamma_d + \varepsilon_{itsd} \qquad (1)$$

where Y_{itsd} represents student's Texas STAAR score measured at the end of study year 2 of student *i*, with teacher *t*, in school *s*, in random assignment block *d*. Z_{sd} is an indicator for a school being randomized to use Zearn Math, and $\boldsymbol{X_{itsd}}$, $\boldsymbol{T_{tsd}}$, and $\boldsymbol{S_{sd}}$ are vectors of student, teacher, and school characteristics, respectively, mean-centered and standardized as appropriate. The parameter of interest, β_1, represents the effect of a school being assigned to implement Zearn Math.

Student characteristics include grade level, age, gender, race/ethnicity, economic disadvantage, English learner status, receipt of special education services, eligibility for gifted programming, prior-year mathematics achievement, and indicators for whether students repeated

or skipped grades between year 1 and year 2 of the study. Teacher characteristics include gender, race/ethnicity, years of teaching experience in the district, and aggregates of student characteristics. School characteristics included Title I status and aggregates of student characteristics. Though these vectors of characteristics are not required for unbiased estimates, they were included to increase precision.

Finally, γ_d are block fixed effects and ε_{itsd} is an individual level stochastic error term. Standard errors were clustered by school to account for the correlation of outcomes within schools, in accordance with research indicating that standard errors should be clustered at the unit of random assignment. We opted for ordinary least squares models with cluster-robust standard errors because the method makes fewer distributional assumptions and is more robust to real-world issues in education RCTs, such as noncompliance. For example, Schweig et al. (2020) showed that even small amounts of individual-level noncompliance in cluster-randomized trials can bias treatment effect estimates or standard errors calculated using hierarchical linear models.

For subgroup analyses, a variant of Model 1 interacts T_{sd} with moderator M_{itsd}, an indicator for membership in the subgroup of interest.

$$Y_{itsd} = \beta_0 + \beta_1 T_{sd} + \beta_2 Z_{sd} * M_{itsd} + \boldsymbol{X_{itsd}\beta_3} + \boldsymbol{T_{tsd}\beta_4} + \boldsymbol{S_{sd}\beta_5} + \gamma_d + \varepsilon_{itsd} \quad (2)$$

Here, β_2 represents the differential effect of Zearn Math for subgroup members relative to non-members, or the holdout group if is categorical. The total effect for subgroup members is represented by $\beta_1 + \beta_2$.

Missingness rates were low among the characteristics listed above: 11 percent of students in the primary sample were missing prior-year achievement scores, 12 percent were missing teacher years of experience, and 2 percent were missing teacher race/ethnicity and gender. We implemented multivariate imputation by chained equations using the mice package in R to fill in

missing values for these four variables. All covariates from the confirmatory model were included as predictors in the imputation model, except for teacher and school aggregates of student characteristics due to collinearity, as well as both the Texas STAAR and NWEA MAP raw scores. Student-level covariates were set as level 1 predictors in the imputation model, and teacher- and school-level covariates were set as level 2 predictors. Unique teacher IDs were used as the level 2 group identifier and in doing so, the mice package aggregated level 1 predictors at the teacher level. We imputed 10 datasets, each using 20 iterations of the chained equation process. Imputed values of the Texas STAAR and NWEA MAP outcomes were excluded from all confirmatory and exploratory models and treatment estimates, meaning that students missing those scores are dropped from the corresponding analysis and considered lost to attrition.

Statistical power

After applying the false discovery procedure, the study was powered to detect a main effect of 0.17 SD for RQ1, and a differential effect of 0.18 SD between below-proficient students and their peers above the proficiency threshold for RQ2. The rationale for powering for a main effect size slightly larger than the 0.16 SD effect estimated by Klopfenstein (2018) was that our study would measure benefits of Zearn Math that accumulate over two years. Powering for a smaller effect would have made the study larger and more expensive, and recruitment more challenging.

Analyses of implementation data

For the present manuscript, we focus on responses to survey items regarding characteristics of math instruction, noting areas where treatment and control group responses appear similar or appear to differ by a substantive amount. We also compare average teacher responses between study years 1 and 2 to investigate changes over time. We coded interview transcripts in Dedoose (SocioCultural Research Consultants, 2025) using a thematic coding scheme that incorporated

inductive and deductive codes. We analyzed coded text excerpts to compare similarities and differences within and across case study sites.

Results

Attrition and baseline balance

Table 2 displays attrition statistics for two-year analyses including the two confirmatory RQs. All schools randomized are represented for both STAAR and MAP outcomes, resulting in no cluster-level attrition. At about 17 percent, student-level attrition was low over the two years, as was differential attrition of 1 percent. These are well within WWC 5.0 standards for low attrition under cautious assumptions (WWC, 2022). These attrition results, along with our non-acceptance of late joiners and the intervention's low risk of bias due to joiners, make the study eligible to meet WWC standards without reservation (WWC, 2022).

Demonstrating baseline balance is not required by WWC for a randomized controlled trial to meet standards without reservation. Nonetheless, we report baseline balance as an additional indicator of the internal validity of the study. Table 1 shows that, except for the percentage of white students, standardized group differences across all baseline variables were less than 0.05. The group difference for white students was 0.07. None of these differences are statistically significant. Most importantly, baseline differences in baseline achievement and proficiency were 0.01 and 0.00, respectively.

Overall, these attrition and balance results demonstrate that the study execution retained the strong internal validity expected from a randomized controlled trial.

Student achievement results

Confirmatory analyses of effects on STAAR, the primary achievement outcome

We begin by reporting results for the two preregistered confirmatory research questions this study was designed to address. Table 3 shows that for all students in the analytic sample (RQ1),

students in schools randomized to use Zearn Math outperformed students in the control group on the Texas STAAR by a standardized effect of 0.07. This estimate is not statistically significant. When focusing on students who were not proficient at baseline, the standardized effect estimate of 0.10 is slightly larger but also not statistically significant. The study was unable to confirm positive effects of Zearn Math on the Texas STAAR for all students or the subgroup of students who were below proficient at baseline.

Exploratory analyses of effects on MAP, the secondary achievement outcome

We next discuss exploratory analyses that parallel the confirmatory analyses, but on the MAP assessment instead of STAAR. The district's use of the MAP assessment provides an opportunity to estimate effects of Zearn Math on a second, validated assessment of math achievement. As discussed earlier, the adaptive nature of MAP may make it more sensitive to student learning of content that diverges from the focal content for the student's grade level, such as below-grade level learning that students may accomplish while working with supplemental software like Zearn Math. Table 4 shows that students in schools randomized to use Zearn Math outperformed their control group peers by a standardized effect of 0.11. Focusing on just the students who were not proficient at baseline, we estimate an effect of 0.13, again slightly larger than the effect for all students. Both estimates are statistically significant, yet these are not the questions the study was designed to address and adherence to our preregistration means that these results should not receive the same weight as the confirmatory results. As exploratory analyses these MAP results require future prospective research to confirm them. Even if we had deemed these results confirmatory alongside the STAAR results, they would have been non-significant after adjustment for multiple hypothesis tests.

WWC synthesis of STAAR and MAP effects for all students

With its version 5.0 standards, WWC (2022) adopted a meta-analytic approach to assessing the evidence generated by a study, which stands in contrast to the primacy of preregistered confirmatory analyses we adopted. WWC's new method would calculate a meta-analytic average of the two whole-sample estimates – the estimated effects for all students on the STAAR and MAP assessments (row 1 in each of Tables 3 and 4). The WWC meta-analysis averages the two results and considers the correlation between the two outcomes in calculating the standard error. We calculated the meta-analytic average of the effects of the STAAR and MAP effect estimates as defined in the WWC Procedures and Standards Handbook, version 5.0 (WWC, 2022), and Table 5 displays the results. The estimated standardized effect of 0.09 is not statistically significant. This result reinforces that the study does not confirm a positive effect of Zearn Math on mathematics achievement.

Exploratory heterogeneity analyses

The rich administrative dataset offers an enticing opportunity to explore effects by subgroups of students based on demographic variables or baseline performance. Indeed, we specified exploratory research questions that imply dozens of such tests. There are several reasons to interpret these results with caution. First, the study was underpowered to detect differences in effects between subgroup members and nonmembers. Prospective power calculations determined that the minimum detectible standardized difference between subgroup members and non-members was at least 0.18. Second, when running so many statistical tests, the risk of spurious statistical significance is elevated. The risk can be mitigated by applying multiplicity corrections, which renders all the results non-significant. Third, we present estimated total treatment effects for subgroup members for ease of interpretation. In that manner of presentation, statistical significance evaluates the probability the estimate is different than zero, not that it differs from

subgroup non-members. Finally, these are exploratory analyses, implying they might be used to generate hypotheses for future study but not as already-confirmed results.

Tables 6 and 7 present subgroup results for the STAAR and MAP assessments, respectively. Each table displays estimated effects for subgroup by quartile of baseline achievement, by race/ethnicity, by gender, by grade level at randomization, and for economically disadvantaged and English language learner students.

In summary, we lack evidence of heterogeneity in effects across subgroups of students. Estimates for most subgroups are near the average effect, meaning subgroup differences are not detected. Divergences between subgroup members and non-members are nonsignificant and may be due to statistical noise.

Exploratory one-year effects on STAAR and formative assessments

Finally, we explore effects of Zearn Math on math achievement after students' first year of exposure to Zearn Math. This includes all students in grades 3-5 the first year of the study, and a new cohort of third graders the second year. The district switched adaptive formative assessments from the Renaissance Star in year 1 to the MAP in year 2. To analyze the combination of Star and MAP scores as outcomes, we standardized them against national norms (z-scores) and then combined them. The first two rows of Table 8 show that we estimated one-year effects of zero on the Texas STAAR, and 0.07 on the adaptive formative assessments. Neither estimate is statistically significant.

To complement this and ease comparisons with two-year effects, we also conducted one-year analyses that included only primary-sample students who also had the one-year outcome. In these analyses, presented in the last two rows of Table 8, we estimated one-year effects of 0.01 on the Texas STAAR, and 0.09 on the Renaissance Star adaptive assessment. Neither estimate is statistically significant. Comparing with the two-year results, for STAAR it appears that the

greater part of the estimated two-year effect may have occurred in year 2. This is not echoed in the formative assessment results, although the outcome measures were different each year (Star in year 1 and MAP in year 2).

Additional descriptive statistics and study results to facilitate WWC analysis

We conclude our presentation of student outcome results with ancillary information that WWC recommends reporting (WWC, 2021). These are presented in Tables 9 and 10.

Implementation Study Results

The extent to which students met goals for Zearn Math use

In study year 1, 96 percent of treatment group survey respondents reported using Zearn Math once per week or more, although only 78 percent reported that Zearn Math was their *main* supplemental technology product for math instruction. In year 2, a similar proportion (95 percent) of treatment group teachers reported using Zearn Math regularly but the proportion reporting it was their main supplemental technology product increased to 88 percent.

Table 11 summarizes software log data to display the number and percentage of treatment group students who met or exceeded Zearn goals for time using the software (90 minutes per week) and completion of on-or-above grade-level lessons (90 lessons per year). Only 18 percent of treatment group students met the annual lesson completion goal by the end of year 1. During the first semester, only nine percent of students completed 45 on-or-above grade-level lessons, but the pace of lesson completion increased the second semester, when 26 percent of students completed 45 lessons. In year 1, more students met time-use goals than lesson-completion goals, with a similar trend of increased use in the second semester.

Relative to year 1, there were substantial increases in both lesson completion and time using the software in year 2. The percentage of students completing 90 on-or-above-grade-level lessons increased to 50 percent from 18 percent the prior year. The bulk of this increase occurred

during final semester of year 2, with 65 percent of students completing 45 lessons versus 21 percent in first semester. Time use metrics reflect the lesson-completion increases, with 76 percent of students meeting or exceeding the goal to use the software for 90 minutes per week the final semester compared to 25 percent the first semester of year 2.

Across year 2, treatment-group students spent a median of 102 minutes per week using Zearn Math and completed a median of 90 on-or-above-grade-level lessons, meeting both of Zearn's usage goals of 90 minutes per week and 90 lessons per year. Table 12 breaks these metrics down by semester and quartile of baseline achievement, revealing several details: the yearlong results can be credited to use in the second semester that was much higher than recommended; within semesters, students across quartiles used the software for similar median amounts of time; and finally, despite the similarities in time using Zearn Math, students in lower quartiles completed, on median, fewer on-or-above-grade-level lessons than their counterparts in higher quartiles. The lower two quartiles of students did not, on median, fully meet the goal of 90 at-or-above-grade-level lessons, although the second quartile approached that level. Analyses not shown in the table suggest that lower-quartile students were using their time to attempt on-or-above-grade-level lessons, not to work on below-grade-level lessons. Apparently, lower-quartile students simply need more time to complete grade-level lessons than their higher-quartile peers.

Conditions influencing Zearn Math use each year

Here we present some high-level conditions that appeared to influence usage during the study, and how they changed from year 1 to year 2. In study year 1, only one of the six case study sites consistently met Zearn's usage goals. This site's existing routines and practices around math instruction, as described by the staff we interviewed, were unique among the sites we visited. They had a strong instructional leader with a clearly articulated vision for math instruction, which included the use of Zearn Math. They also had existing professional learning routines and

practices, including a culture around coaching, which the instructional leader leveraged to encourage teachers to use Zearn Math with their students. Interviewees at this site described adapting their instructional practices to incorporate Zearn Math in response to the guidance, coaching, monitoring, and feedback from the instructional leader.

In contrast, lack of strong endorsement of Zearn Math from school leadership in the other five study sites seemed to have played a role in their relatively lower usage during study year 1. In some sites, leaders offered guidance to teachers that conflicted explicitly with Zearn's usage goals. In others, the absence of guidance about Zearn Math led teachers to make their own decisions about how to allocate instructional time, which did not always include Zearn Math.

During interviews, teachers and leaders at case study sites raised some concerns about Zearn's appropriateness for specific student subgroups. Across four case study sites, teachers raised doubts about Zearn Math's effectiveness for struggling students. Across five sites, they raised concerns about Zearn Math's effectiveness for English learners because it provides instruction only in English. District administrators also mentioned this to us, suggesting the concern was more widespread than the case study schools. Some case study teachers went so far as to use products other than Zearn Math with English learners. Though these concerns were salient in interviews, they were less so in survey data. In both years, more than 80 percent of treatment group teachers said Zearn Math afforded the right level of challenge for the majority of students, and teachers in both experimental groups responded similarly on how well their main technology product met the needs of English learners. Responses for the latter question averaged about 4.7 on a scale of 1 to 7, where 1 was "completely inadequate" and 7 was "completely adequate." We hypothesize that the different modes of data collection could account for this difference –

interviews may afford teachers a greater opportunity than surveys to discuss challenges for student subgroups.

Teachers and building leaders from all case study sites reported changes in leadership behaviors in year 2, which coincided with the sharp increase in usage. Following guidance from the district, leaders of case study sites reported emphasizing Zearn Math expectations to encourage increased use, and monitoring use to follow up with teachers whose students were not meeting goals. Leaders at all the case study sites reported adjusting school schedules to accommodate time for Zearn Math in year 2, which they had not done in year 1. Teachers in five of the case study sites shared in interviews that the changes in how school leaders communicated about Zearn Math, combined with policy changes accommodating its use, led them to prioritize more than they had previously. The sixth, described above, was already prioritizing Zearn Math in year 1.

Interviewees across all the case study sites reported that district administration took a stronger role in instructional policies in year 2 compared to year 1, including steps to increase Zearn Math use. Administrators clearly articulated Zearn Math as a priority in their communications with building leaders. In case study interviews, leaders reported that the district began monitoring school-level use of Zearn Math and showed them which schools met or did not meet usage goals. They reported that this monitoring made them feel more accountable and motivated to meet those goals. Neither case study interviews nor conversations with district officials revealed any consequences for low use. Finally, in partnership with Zearn, the administration facilitated school, classroom, and student-level incentives and rewards for meeting usage goals.

In the surveys, treatment group teachers reported a decline from year 1 to year 2 in the severity of challenges to Zearn Math implementation. In year 1, competing priorities, lack of

professional development, and "the need to ensure [Zearn Math] contributes to student learning" were the most commonly cited barriers to Zearn Math implementation. The proportion of teachers who described those barriers as moderate or major challenges decreased from year 1 to year 2. A similar pattern was seen in the proportion of teachers rating a poor internet connection as a major or moderate challenge, decreasing from 25 percent to 8 percent of teachers from year 1 to year 2. Likewise, the proportion of teachers reporting too few devices as a major or moderate challenge declined from 21 percent to 7.

Though our survey did not ask teachers explicitly about facilitators to Zearn Math use, we hypothesize that the declines in barriers reported by teachers in year 2 reflect improved conditions for Zearn Math implementation that year.

Contrasts in math instructional practices between the Zearn Math and control groups

The annual survey asked teachers from both experimental conditions to report on their math instructional practices, how they allocate instructional time, how they use data, and their experiences with professional learning. Across both years, teachers in both groups mostly responded similarly, though we observed some differences in time allocation and data use practices.

In both years, teachers in both groups reported spending similar amounts of time on math instruction, including time spent on math-related supplemental technology products outside of the regular math period. Teachers in both groups reported similar rates of using math instructional practices such as having students relate new math content to other math content or explain their thinking. Likewise, they reported similar rates of using data to guide their instruction, such as to tailor instruction to individual students' needs, to group students within classes, or to identify topics for which students need extra review.

The survey asked teachers to report on the supplemental technology products they use regularly, defined as once per week or more on average. The most common products selected by control group teachers changed from Imagine Math, Kahoot, and Go Math! in year 1 to ST Math, i-Ready, and IXL in year 2. Schools not using Zearn Math were encouraged by the district to use ST Math in year 2. To our awareness, the adaptive products used heavily by the control group use a foundational approach. Among these products ST Math stands alone as having been reviewed by WWC and demonstrated evidence of effectiveness. Both years, teachers in both groups responded similarly regarding the adequacy of the product they used most often for: engaging students, being a good use of time, covering required standards, and supporting learners with diverse needs. Both years, most teachers in both groups rated their main product as at the right level of difficulty for a majority of their students.

Though teachers in both groups reported spending similar amounts of time on math instruction overall, treatment group teachers reported spending more time on independent technology work and less time on non-technology independent work relative to control group teachers. Treatment group teachers also reported greater use of data from their main product to monitor student progress than did control group teachers. Both of these differences were present both years.

The frequency, topics, and amount of time teachers reported engaging in professional learning was similar between groups both years. However, treatment group teachers in year 2 reported receiving coaching related to their supplemental math product (i.e., Zearn Math) more often than control group teachers.

Discussion

The study did not provide confirmatory evidence that Zearn Math has a positive effect on students' state mathematics assessment scores. The study was designed to estimate Zearn

Math's effect on student achievement of grade-level math content, measured by the Texas STAAR assessment after two years, (1) for all students, and (2) for the students who were below proficient at baseline. Those are the two confirmatory analyses the study preregistered. When designing the study, RAND and Zearn agreed that the state accountability test would be the focus of the confirmatory tests because producing effects on that type of assessment has strong policy relevance in the U.S. Both confirmatory results are positive but non-significant even before applying multiplicity corrections.

Exploratory analyses produce uniformly positive estimates of the effects of Zearn Math; evidence from these analyses may be useful for confirming a positive effect in the future. The larger effects estimated for MAP than for STAAR in this study are consistent with the possibility that Zearn Math's support and scaffolding help students to fill gaps in below-grade-level mathematics content. An adaptive assessment like MAP may be better able to measure learning of below-grade-level content than would the STAAR which was designed to prioritize measuring proficiency on grade-level content. A future experiment testing effects of Zearn Math on MAP or a similar adaptive assessment could hypothetically confirm that Zearn Math is better than the counterfactual for helping students learn math content conceived more broadly than just grade-level content.

Moreover, although the observed variability in subgroup effects could be statistical noise, when a noisy subgroup estimate aligns with theory or other qualitative data, it might warrant consideration for future work. For example, smaller estimated effects for English learners in this study might be considered alongside reports from teachers and administrators that Zearn Math lacked strong support for that subgroup. Maybe Zearn Math's efficacy could be enhanced with greater support for English learners, then tested in a new efficacy study with a population like the

population of this study, where nearly half of students were classified as English learners. Efficacy for both the subgroup and the overall study population could possibly improve.

Finally, this study can help to build the evidence base on Zearn Math by contributing data to future WWC-style meta-analyses or by helping to inform the details future study designs, such as sample size and power calculations, methods to ensure strong implementation like was seen in the final semester of this study, or adjusting how many years of exposure to Zearn Math to test.

The WWC meta-analysis that incorporates results from the MAP secondary achievement outcome concludes that Zearn Math produced uncertain effects on student mathematics achievement. The WWC meta-analysis takes an evidence synthesis approach to summarizing study results, providing an alternative to the primacy we have placed on the preregistered confirmatory analyses. The meta-analysis gives equal weight to the all-student analyses of effects on both the STAAR and MAP assessments in calculating a positive average effect that is not statistically significant. Our application of WWC's (2022) version 5.0 standards and procedures suggests that WWC would deem this study to have met WWC standards without reservation and to have found uncertain effects on mathematics achievement.

These results do not mean the study confirmed Zearn Math had no effect on student achievement. Importantly, this study's failure to confirm a positive effect does not mean there was not a positive effect or that it was too small to be meaningful. The evidence of promise emerging from this study can potentially lead to confirming a positive effect in further research.

Zearn Math's estimated effect is medium in magnitude and has the potential to accumulate over numerous years. The estimated effect of Zearn Math on STAAR was 0.07 SD, which is equivalent to an average control group student moving up 3 percentile points if they had been exposed to Zearn Math. The analogous interpretation for the WWC meta-analytic result

of 0.09 SD is an increase of 4 percentile points. Kraft (2020) classifies effects in this range as medium in magnitude. Those effects were measured over two years, but Zearn Math is available from kindergarten through eighth grade. If a student were exposed to Zearn Math over those nine years, its effect could hypothetically add up to something quite more substantial and educationally meaningful than the two-year effect estimated in this study.

Zearn Math's two-year estimated effect of 3 to 4 percentile points can be compared to effects measured in rigorous studies of other supplemental technology products intended for general education mathematics in K-8 or a subset of those grades. A May 2025 search of WWC's (undated) database identified five supplemental products, with Tier 1 to 3 evidence of positive effects on a broad standardized measure of mathematics achievement: ASSISTments, DreamBox Learning, Larson Pre-Algebra, Reasoning Mind, and ST Math. The estimated single-year effects for these five products ranged from 4 to 16 percentile points, with a median of 6. The range and median are the same if we confine the analysis to the three products that cover grades 3-5.

This study estimated Zearn Math's effect relative to a somewhat strong counterfactual. Participating schools had access to and used other supplemental math products that may already be more effective than other activities for independent student work. Any positive effect of Zearn Math is estimated over and above any effects of those products used by the control group. The same can be said regarding changes to instructional practice that occurred with the change in district administration. Because those changes occurred in both experimental groups, effects for Zearn Math are estimated in excess of any effects of those other changes.

School improvement may not hinge on a single "magic bullet" that demonstrates a large increase in student achievement, but rather by layering numerous changes that incrementally increase achievement. In that conception, changes such as the adoption of Zearn Math, that show

signals of positive effects may be warranted even if their estimated effects are positive but non-significant. What matters in an incremental improvement framework is not the effect of just one change, but the net effect of all of them, which this study was not designed to measure.

Effects of Zearn Math appear to be largely attributable to the software itself and not changes in classroom instruction by teachers. Other than the use of Zearn Math, the study found few differences in mathematics instruction between the treatment and control groups. Teachers in both groups generally reported similar amounts of total instructional time and similar instructional practices in both years of the study. Treatment group teachers did report spending relatively more time than control group teachers on independent work with technology – and less on independent work without technology, the apparent source of extra time with technology.

The study provides suggestive evidence in favor a grade-level approach for supplemental technology products. Two-thirds of students in the primary sample entered the study performing below grade-level proficiency. Such students make up the bottom two quartiles and the majority of the third quartile of baseline achievement. Even the lowest two quartiles of students completed a substantial number of on-or-above-grade-level lessons during year 2. The quartile medians of lessons completed were 59 and 81 for the lowest and second quartile, respectively, providing evidence that even students far below grade level can complete a substantial amount of grade-level work when given the opportunity. Although the study was underpowered to detect difference across quartiles in estimating the effects of Zearn Math, there is no sign the lower quartiles benefited less than their higher-performing peers. The supplemental math technology products that were used in the control group generally used a foundational approach, and Zearn Math's estimated effects are relative to whatever benefits those products

may have provided. The trends of positive effects for Zearn Math are thus suggestive that a grade-level approach might be advantageous.

Replicating the usage of Zearn Math seen in year 2 of this study could be a key to addressing the perennial problem of low usage of promising education technology products. Use of Zearn Math started well below recommended levels in year 1. This followed patterns seen in many previous evaluations of education technology products and characterized by Holt (2024) as the 5 percent problem. However, in year 2 of this study, the medians of student time using Zearn Math and completing on-or-above-grade-level lessons met goals of 90 minutes per week and 90 lessons per year. This was achieved by exceeding usage goals by a wide margin in the final semester of the study. Challenges to Zearn Math use also decreased that year, as reported by teachers on surveys. A change in district administration between study years, along with accompanying changes to policies and practices – such as clear communication from administrators to support strong implementation, and incentives toward that objective – likely played a role in the increased usage. The mechanisms by which usage goals were met might help guide other districts toward meeting goals for technology use, perhaps at a more stable rate across the full implementation period.

Conclusion

Although this study does not provide confirmatory evidence that Zearn Math improves grade-level achievement, consistent positive signals across confirmatory and exploratory results suggest that Zearn Math has promise for improving student learning and is not likely to be detrimental. Inability to confirm a positive effect does not mean we have confirmed there is no effect, and lack of any negative signals among the multitude of effects we estimated implies the study found no evidence suggesting that Zearn Math was harmful to student learning. While awaiting stronger evidence of positive effects, these signs of promise might be given some

tentative weight among other considerations in making decisions about whether to use Zearn Math.

Acknowledgements

The authors acknowledge the contributions of Zearn staff to the design and implementation of this study. We also acknowledge RAND staff who contributed to the research design and study implementation and analysis, including Julia Kaufman, Miray Tekkumru-Kisa, Julia Szabo, Tara Blagg, Allyson Gittens, Eupha Jeanne Daramola, Melanie Rote, Lauren Covelli, Anton Wu, Karen Christianson, Armenda Bialas, and Zhan Okuda-Lim. We are grateful for the educators and administrators in the participating school district for agreeing to participate in and provide data for this study. We also appreciate the wisdom offered by external advisors John Friedman, Paul von Hippel, Cristofer Price, Elizabeth Tipton, and Vivian Wong. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A230170 to the RAND Corporation. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological), 57(1), 289-300.

Booth, J.L., & Newton, K.J. (2012). Fractions: Could they really be the gatekeeper's doorman? *Contemporary Educational Psychology, 37*(4), 247-253.

Chatterji, Aaron K. 2018. Innovation and American K–12 Education. *Innovation Policy and the Economy*, 18, 27–51. doi:10.1086/694406

Committee on STEM Education. (2018). *Charting a Course for Success: America's Strategy for STEM Education*. Retrieved from https://trumpwhitehouse.archives.gov/wp-content/uploads/2018/12/STEM-Education-Strategic-Plan-2018.pdf

Escueta, M., Nickow, A. J., Oreopoulos, P., & Quan, V. (2020). Upgrading Education with Technology: Insights from Experimental Research. Journal of Economic Literature, 58(4), 897-996. doi:10.1257/jel.20191507

Foster, M.E. (2024). Evaluating the impact of supplemental computer-assisted math instruction in elementary school: A conceptual replication. *Journal of Research on Educational Effectiveness, 17*(1), 94-118.

Hashim, S. (2024). Measuring the Efficacy of Zearn Math in Louisiana. *AERA Open,* doi:10.1177/23328584241269825

Holt, L. (2024). The 5 Percent Problem. Education Next.

Ikemoto, G. S., Steele, J. L., & Pane, J. F. (2016). Poor Implementation of Learner-Centered Practices: A Cautionary Tale. Teachers College Record (Yearbook), 118(13), 1-34.

Karam, R., Pane, J. F., Griffin, B. A., Robyn, A., Phillips, A., & Daugherty, L. (2017). Examining the implementation of technology-based blended algebra I curriculum at scale. Educational Technology Research and Development, 65, 399-425. doi:10.1007/s11423-016-9498-6

Klopfenstein, K. (2018). The Impact of Integrated Blended Learning in Elementary Mathematics. unpublished manuscript.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. Educational Researcher, 49(4), 241–253. doi:10.3102/ 0013189X20985448

Lambert, M., & Sassone, J. (2020). Accelerate, Don't Remediate: An Instructional Framework for Meeting the Needs of the Most Vulnerable Students after COVID School Closures. Journal for Leadership and Instruction, 19(2).

Lewis, K., & Kuhfeld, M. (2022). Progress towards pandemic recovery: Continued signs of rebounding achievement at the start of the 2022–23 school year: NWEA. https://www.nwea.org/uploads/2022/12/CSSP-Brief_Progress-toward-pandemic-recovery_DEC22_Final.pdf

McMullen, J., Laakkonen, E., Hannula-Sormunen, M., & Lehtinen, E. (2015). Modeling the Developmental Trajectories of Rational Number Concept(s). *Learning and Instruction, 37*, 14-20.

Morrison, J. R., Wolf, B., Ross, S. M., Risman, K. L., & McLemore, C. C. (2019). *Efficacy Study of Zearn Math in a Large Urban School District*. Center for Research and Reform in Education, Johns Hopkins University.

National Assessment of Educational Progress. (2022). The Nation's Report Card. U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. Reading and Math Results. Retrieved from: https://nces.ed.gov/nationsreportcard/assessments/

National Assessment of Educational Progress. (2024). The Nation's Report Card. U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. Reading and Math Results. Retrieved from: https://nces.ed.gov/nationsreportcard/assessments/

National Mathematics Advisory Panel. (2008). *The Final Report of the National Mathematics Advisory Panel: Foundations for Success.* Washington: U.S. Department of Education.

Phillips, A., Pane, J. F., Reumann-Moore, R., & Shenbanjo, O. (2020). Implementing an Adaptive Intelligent Tutoring System as an Instructional Support. Educational Technology Research and Development, 68, 1409–1437. doi:10.1007/s11423-020-09745-w

Pollock, E., Chandler, P., & Sweller, J. (2002). Assimilating complex information. Learning and Instruction, 12(1), 61-86. doi:https://doi.org/10.1016/S0959-4752(01)00016-0

Powell, S. R., Gilbert, J. K., & Fuchs, L. S. (2019). Variables Influencing Algebra Performance: Understanding Rational Numbers is Essential. *Learning and Individual Differences, 74*, Article 101758. doi:10.1016/j.lindif.2019.101758

REES (undated). Registry of Efficacy and Effectiveness Studies https://sreereg.icpsr.umich.edu/sreereg/about

Renaissance Learning. (2023). Renaissance Star Math: Enterprise Benchmarks and Cut Scores. Wisconsin Rapids, WI: Renaissance Learning.

Schweig, J., Pane, J. F., & McCaffrey, D. F. (2020). Switching Cluster Membership in Cluster Randomized Control Trials: Implications for Design and Analysis. Psychological Methods, 25(4), 516-534. doi:10.1037/met0000258

Schweig, J., Pandey, R., Grant, D., Kaufman, J. H., Steiner, E. D., & Seaman, D. (2024). American Mathematics Educator Study: 2024 Technical Documentation and Survey Results. Retrieved from https://www.rand.org/pubs/research_reports/RRA2836-4.html

Siegler, R. S., Duncan, G. J., Davis-Kean, P. E., Duckworth, K., Claessens, A., Engel, M., Susperreguy, M. I., & Chen, M. (2012). Early Predictors of High School Mathematics Achievement. *Psychological Science, 23*(7)*,* 691–697. https://doi.org/10.1177/0956797612440101

SocioCultural Research Consultants, L. (2025). Dedoose Version 10.0.25, cloud application for managing, analyzing, and presenting qualitative and mixed method research data. Los Angeles, CA. Retrieved from www.dedoose.com

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. Learning and Instruction, 4(4), 295-312. doi:10.1016/0959-4752(94)90003-5

TNTP. (2021). Accelerate, Don't Remediate: New Evidence from Elementary Math Classrooms. Retrieved from https://tntp.org/publication/accelerate-dont-remediate/

van Ginkel, J. R. (2019). Significance Tests and Estimates for R2 for Multiple Regression in Multiply Imputed Datasets: A Cautionary Note on Earlier Findings, and Alternative Solutions. Multivariate Behavioral Research, 54(4), 514–529.

What Works Clearinghouse. (2021). WWC Reporting Guide for Study Authors. Institute of Education Sciences, U.S. Department of Education. https://ies.ed.gov/ncee/wwc/ReportingGuide

What Works Clearinghouse. (2022). What Works Clearinghouse Procedures and Standards Handbook, Version 5.0. In: Institute of Education Sciences, U.S. Department of Education. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/Final_WWC-HandbookVer5_0-0-508.pdf

What Works Clearinghouse. (undated). WWC Publication Search. Institute of Education Sciences, U.S. Department of Education. Accessed at https://ies.ed.gov/ncee/WWC/Search/Products

Zearn. (2022). Nebraska students experience large gains in math learning with Zearn: 2022 Nebraska State Assessment Results. Retrieved from: https://about.zearn.org/insights/zearn-impact-statewide-study-nebraska

Zearn. (2024a). DCPS students see significant math gains with consistent Zearn usage: i-Ready diagnostic interim assessment results. Retrieved from: https://about.zearn.org/insights/zearn-impact-study-dcps

Zearn. (2024b). Zearn Math Quasi-Experimental Efficacy Study: Prepared for the Texas Education Agency. Retrieved from: https://about.zearn.org/insights/zearn-impact-study-texas

Tables

Table 1: Descriptive statistics and group balance for primary sample

Student Characteristics	Zearn Math Mean	Control Group Mean	Standardized Group Difference	p-value
Number of students	5,349	5,228		
Baseline Math Achievement:				
Grade 2 Renaissance Star Scale Score	907	905	0.01	0.87
Grade 3 Texas STAAR Scale Score	1425	1419	0.01	0.88
Percent of Students Below Proficiency	67%	68%	0.00	0.99
Percent of Students Missing Score	11%	12%	-0.02	0.69
Student Race/Ethnicity:				
Percent Black Students	20%	20%	-0.01	0.93
Percent Hispanic Students	66%	68%	0.00	0.99
Percent White Students	9%	6%	0.07	0.41
Percent Female Students	50%	50%	0.01	0.62
Percent Low Income Students	82%	84%	-0.02	0.85
Percent EL Students	44%	45%	0.00	0.95
Percent Special Education Students	8%	9%	-0.04	0.25
Percent Gifted Students	11%	11%	-0.03	0.68
Student Age	9.33	9.30	0.04	0.23

Note: Standardized group differences are Hedges *g* values calculated from ordinary least-squares models that control for randomization block, with standard errors clustered by school.

Table 2: Sample accounting

	Overall sample	*Zearn group*	*Control group*
Schools randomized	64	32	32
Schools analyzed	64	32	32
School attrition rate	0.0%	0.0%	0.0%
School differential attrition		0.0%	
Texas STAAR Assessment			
Students present at baseline	10,577	5,349	5,228
Students present at posttest	8,797	4,478	4,319
Students analyzed	8,797	4,478	4,319
Student attrition	1,780	871	909
Student attrition rate	16.8%	16.3%	17.4%
Student differential attrition		1.1%	
NWEA MAP Assessment			
Students present at baseline	10,577	5,349	5,228
Students present at posttest	8,807	4,475	4,332
Students analyzed	8,807	4,475	4,332
Student attrition	1,770	874	896
Student attrition rate	16.7%	16.3%	17.1%
Student differential attrition		0.8%	

Table 3: Confirmatory effects on STAAR, the primary achievement outcome

	Outcome	g	SE	p	95% Confidence Interval	n Stu.
All students	Texas STAAR	0.07	0.06	0.22	-0.04 – 0.19	8,797
Not Proficient at Baseline	Texas STAAR	0.10	0.08	0.20	-0.05 – 0.26	5,859

Note: Because unadjusted p-values are non-significant, the pre-specified multiplicity corrections make no substantive difference.

Table 4: Exploratory effects on MAP, the secondary achievement outcome

	Outcome	g	SE	p	95% Confidence Interval	n Stu.
All students	MAP	0.11	0.04	0.01	0.02 – 0.19	8,807
Not Proficient at Baseline	MAP	0.13	0.05	0.01	0.03 – 0.24	5,863

Note: p-values have not been corrected for the multitude of statistical tests.

Table 5: Authors' WWC synthesis of STAAR and MAP effects for all students

	Outcome	g	SE	p	95% Confidence Interval
All students	STAAR and MAP	0.09	0.05	0.07	-0.01 – 0.19

Table 6: Heterogeneity of effects on STAAR

	Outcome	g	SE	p	95% Confidence Interval	n Stu.
1st Quartile at Baseline (Lowest)	Texas STAAR	0.11	0.10	0.28	-0.09 – 0.31	2,172
2nd Quartile at Baseline	Texas STAAR	0.09	0.10	0.37	-0.10 – 0.27	2,092
3rd Quartile at Baseline	Texas STAAR	0.08	0.09	0.36	-0.09 – 0.26	2,308
4th Quartile at Baseline (Highest)	Texas STAAR	0.10	0.08	0.21	-0.06 – 0.25	2,226
Low Income	Texas STAAR	0.07	0.07	0.26	-0.06 – 0.20	7,220
English Learner	Texas STAAR	0.04	0.08	0.64	-0.12 – 0.20	4,059
Black	Texas STAAR	0.22	0.08	0.01	0.06 – 0.39	1,514
Hispanic	Texas STAAR	0.03	0.07	0.65	-0.10 – 0.17	6,091
White	Texas STAAR	0.15	0.12	0.21	-0.09 – 0.39	716
Female	Texas STAAR	0.07	0.06	0.23	-0.05 – 0.19	4,352
Male	Texas STAAR	0.07	0.06	0.24	-0.05 – 0.20	4,445
Grade 3 at Baseline	Texas STAAR	0.11	0.07	0.15	-0.04 – 0.25	4,299
Grade 4 at Baseline	Texas STAAR	0.04	0.07	0.55	-0.10 – 0.18	4,498

Note: p-values have not been corrected for the multitude of statistical tests.

Table 7: Heterogeneity of effects on MAP

	Outcome	g	SE	p	95% Confidence Interval	n Stu.
1st Quartile at Baseline (Lowest)	MAP	0.17	0.08	0.03	0.02 – 0.32	2,173
2nd Quartile at Baseline	MAP	0.13	0.07	0.08	-0.02 – 0.27	2,094
3rd Quartile at Baseline	MAP	0.09	0.07	0.20	-0.05 – 0.24	2,310
4th Quartile at Baseline (Highest)	MAP	0.14	0.07	0.04	0.00 – 0.28	2,232
Low Income	MAP	0.12	0.05	0.01	0.03 – 0.21	7,228
English Learner	MAP	0.09	0.05	0.10	-0.02 – 0.19	4,062
Black	MAP	0.19	0.07	0.01	0.05 – 0.33	1,516
Hispanic	MAP	0.07	0.05	0.15	-0.03 – 0.17	6,096
White	MAP	0.17	0.11	0.13	-0.05 – 0.38	717
Female	MAP	0.11	0.04	0.01	0.02 – 0.20	4,360
Male	MAP	0.10	0.05	0.03	0.01 – 0.20	4,447
Grade 3 at Baseline	MAP	0.12	0.05	0.02	0.02 – 0.22	4,304
Grade 4 at Baseline	MAP	0.10	0.05	0.06	0.00 – 0.21	4,503

Note: p-values have not been corrected for the multitude of statistical tests.

Table 8: Exploratory one-year effects on STAAR and formative assessments

Sample	Outcome	g	SE	p	95% Confidence Interval	n Stu.
All available	STAAR	0.00	0.04	0.97	-0.07 – 0.08	19,913
All available	Star or MAP	0.07	0.04	0.10	-0.01 – 0.15	19,495
Primary only	STAAR	0.01	0.05	0.90	-0.08 – 0.10	8,717
Primary only	Star	0.09	0.04	0.05	0.00 – 0.18	8,515

Note: The “all available” sample includes grade 3-5 students joining the study in year 1, and a new cohort of grade 3 students joining in year 2. The “primary only” sample includes only students who are in the primary sample for two-year analyses and have the outcome score; they are students who were in grades 3 or 4 in year 1. p-values have not been corrected for the multitude of statistical tests.

Table 9: Additional descriptive statistics for the primary analytic sample

	Zearn group					*Control group*				
Measure	n_{stu}	n_{sch}	*Unadjusted mean*	*Adjusted mean*	*Unadjusted standard deviation*	n_{stu}	n_{sch}	*Unadjusted mean*	*Adjusted mean*	*Unadjusted standard deviation*
Pretest Z-score	5,349	32	0.02	N/A	0.97	5,228	32	-0.01	N/A	1.02
STAAR Posttest Score	4,478	32	1,608	1,606	170	4,319	32	1,592	1,594	171
MAP Posttest Score	4,475	32	215	214	18	4,332	32	212	212	19

Notes: Pretest combines scores from two separate assessments (Grade 2 Renaissance STAR and Grade 3 Texas STAAR), thus pretest descriptive statistics are presented as z-scores rather than raw scores. The two posttests report scores on continuous developmental scales. Adjusted posttest means are calculated from linear regression model-based predictions of the posttest score with the treatment indicator set to 1 for all students (Zearn group adjusted mean), and with it set to 0 for all students (Control group adjusted mean). Adjusted mean prediction models also control for student pretest score, baseline characteristics of students, teachers, and schools, and school randomization block, and include students with imputed values for baseline covariates. Posttest scores are not imputed.

Table 10: Additional study results

Model	n_{stu}	n_{sch}	*Estimated Zearn effect*	*Standard error*	*t*	*Degrees of freedom*	*p*	*Effect size (Hedges' g)*	*Model* R^2	*Unadjusted school-level ICC*	*Unadjusted teacher-level ICC*	*Baseline-outcome correlation*
STAAR Score	8,797	64	12.31	10.26	1.20	24.77	0.24	0.07	0.54	0.07	0.15	0.65
MAP Score	8,807	64	1.85	0.81	2.28	24.48	0.03	0.10	0.57	0.08	0.16	0.70

Notes: Both models control for student pretest score, baseline characteristics of students, teachers, and schools, and school randomization block. Models account for clustering at the school level using cluster-robust variance estimation. Model R^2 is the average of R^2's across imputed data sets (van Ginkel, 2019). Unadjusted intra-cluster correlations (ICC) are calculated using random effect models with no covariates.

Table 11: Students meeting Zearn Math usage goals

Year	Semester	Students completing 90 lessons per year		Students averaging 90 minutes per week		Students completing 45 lessons per semester	
		n	percent	n	percent	n	percent
1	whole year	899	18				
	1			719	14	465	9
	2			1,823	36	1,289	26
2	whole year	2,181	50				
	1			1,079	25	911	21
	2			3,300	76	2,839	65

Note: Lessons completed includes only on-or-above-grade-level lessons.

Table 12: Median Zearn Math use during year 2, by baseline achievement quartile

Quartile	Semester 1		Semester 2		All of year 2	
	Minutes per week	Lessons completed	Minutes per week	Lessons completed	Minutes per week	Lessons completed
1 (lowest)	50	11	137	44	99	59
2	53	17	147	61	105	81
3	55	21	144	71	104	97
4 (highest)	55	28	137	84	100	118

Note: Lessons completed includes only on-or-above-grade-level lessons. The whole-year medians for the entire sample were 102 minutes per week and 90 lessons.