ZEARN

TECHNICAL APPENDIX

Efficacy analysis of Zearn Math: Findings from implementation in Tennessee

December 2023

Jessica Rickel, M.A.
Zearn

Alisa Szatrowski, Ph.D.
Zearn

Abstract

Using Coarsened Exact Matching (CEM), this analysis examined the impact of Zearn Math for students in grades 4-8 in Tennessee. Students who completed 3+ Zearn Math lessons per week during the 2021-2022 and/or 2022-2023 school year were compared to similarly matched peers with no usage. Findings showed that, on average, students who used Zearn Math grew an additional 7.2 scale score points on the Tennessee Comprehensive Assessment Program (TCAP) compared to non-users. Students who started “Below Expectations” and used Zearn Math grew an additional 13.8 scale score points compared to non-users. Zearn Math users also improved their Performance Level at higher rates and had a higher percentage of students meeting proficiency on the post TCAP. While students who completed 3 lessons per month, 1-2 lessons per week, and 2-3 lessons per week had higher growth than non-users, students demonstrated strongest gains when averaging 3+ lessons per week. Additionally, students who used Zearn Math in 2021-2022 retained additional growth a year after they discontinued use.

Table of Contents

List of Tables

2022-2023

Introduction

Zearn is the 501(c)(3) nonprofit educational organization behind Zearn Math, a top-rated math learning platform used by 1 in 4 elementary-school students and by more than 1 million middle-school students nationwide. This report summarizes findings from an efficacy analysis of Zearn Math implemented in Tennessee. The goal of this study was to isolate the impact of Zearn Math on student achievement through quasi-experimental matching methods that facilitate causal inference.

This efficacy analysis was conducted in Tennessee, a state with 967,356 students across grades K-12. The student body is 30% economically disadvantaged, 13% students with disabilities, 8% English learners (ELs), 20% chronically out of school, 24% Black, and 13% Hispanic (Tennessee Department of Education, 2022).

Analyses assessed the impact of Zearn Math usage during the 2021-2022 and 2022-2023 school years. In grades 4-8, there were 1,485 consistent Zearn Math users, those who completed 3+ lessons per week or approximately 90+ lessons per year, who could be matched to assessment data from the 2021-2022 and/or 2022-2023 school years (see Appendix A Table A1 for fidelity usage by year).[1]

This study was designed to meet the rigorous standards set by the What Works Clearinghouse (WWC) and qualify as an Every Student Succeeds Act (ESSA) evidence-based intervention. The study uses quasi-experimental matching methods to create baseline equivalency between treatment and control groups along major confounding factors (see Appendix B for more information).

Zearn Usage in Tennessee

Zearn Math launched for some districts in the 2021-22 school year, with additional districts beginning usage in the 2022-2023 school year. In the 2021-2022 school year, the platform engaged approximately 85,000 Tennessee students, and by the 2022-2023 school year, participation nearly doubled.

In 2022-2023, 147,000 K-8 students in 147 districts across all 8 CORE regions engaged with Zearn, representing more than 20% of Tennessee K-8 students. Districts used the platform flexibly across various learning contexts including tutoring, intervention, summer programs, and before/after school.

[1] Tennessee only used Zearn Math with fidelity in grades 7 and 8 during the 2022-2023 school year. School year 2021-2022 only included students in grades 4-6.

Methodology

This paper contains three analyses of the impact of Zearn Math on students' academic growth. The first analysis is a dose-response analysis that examines the impact of Zearn Math usage at 4 thresholds: 3 lessons per month (as recommended by TN ALL Corps), 1-2 lessons per week, 2-3 lessons per week, and 3+ lessons per week (Zearn recommended) on students' academic growth as measured by changes in scale score on the TCAP between pre-scores from spring of the year prior to use and post-scores at the end of the year of Zearn Math usage.

The second analysis examines the impact of consistent Zearn Math usage (defined as the recommended 3+ lessons per week), disaggregated by demographic and academic subgroups and measured by additional growth metrics including changes in Performance Level and changes in percent proficient.

The final analysis is on long-term retention of academic gains for consistent Zearn Math users which looks at the extent to which consistent Zearn Math users from 2021-2022 retained additional learnings through 2022-2023 despite discontinuing Zearn Math usage.

All analyses in this study used quasi-experimental matching techniques to isolate the impact of Zearn Math on student achievement. Consistent Zearn Math users were matched to non-users[2] on starting math and English Language Arts (ELA) Tennessee Comprehensive Assessment Program (TCAP) scores, along with nine student characteristics. The goal of matching was to create 1:1 pairings between similar students, differing primarily on Zearn Math usage during the 2021-2022 and/or 2022-2023 school years. The outcome under investigation was the average treatment effect on the treated as controls were selected to match individuals in the treatment group.

In order to see maximum benefit from Zearn Math, students are advised to complete three or more digital lessons per week during the school year. Therefore, the treatment groups for the one-year impact analysis and the long-term retention analysis were composed of students who consistently used Zearn Math during the 2021-2022 and/or 2022-2023 school years, operationalized as an average of three or more digital lessons per week; 90 or more digital lessons per year. The control group was selected from other students in Tennessee with no Zearn Math usage, operationalized as completing 0

[2] Non-users are those who completed 0 lessons and have accounts on Zearn Math, with the exception of the long-term retention analysis in which non-users are defined as those completing fewer than one lesson per week (i.e., fewer than 30 lessons per year).

lessons per year.[3]

Drawing causal inference from observational data is challenging because factors that impact a person's likelihood to receive an intervention may also impact their outcomes. Therefore the differences in outcomes observed between individuals may not be caused by the intervention itself, but by other confounding factors that imbalance the treatment and control groups (Stuart, 2008; Iacus et al., 2011).

Matching methods were used to balance the composition of confounding factors between individuals who consistently used Zearn Math (the treatment group) and a comparison group of individuals who had no Zearn Math usage (the control group). This is done to isolate the difference in outcomes from the intervention itself, separate from any impact due to potential confounding factors.

This efficacy analysis used a two-step Coarsened Exact Matching (CEM) method with optimal matching to create a control group that was as similar as possible to the treatment group of consistent Zearn Math users. CEM is a technique that simulates block sampling by matching students on covariates, demographic and academic factors that may be related both to a student's likelihood of using Zearn Math consistently and their academic performance (Blackwell et al., 2009; Iacus et al., 2011). The effectiveness of matching is conditional on the ability of observable factors to capture the selection process that sorted individuals into treatment and control. Models that do not capture major factors may produce biased estimates.[4]

Using CEM, treatment students were put into matching strata with control students that were in the same academic year,[5] same grade, and within 5 scale score points on both the math and ELA pre

[3] This definition of treatment and control does not use an intention-to-treat (ITT) framework that would include in the treatment all students that had been offered Zearn Math (McCoy, 2017). While the ITT approach is the most efficacious for identifying the impact of a program under real-world implementation constraints, the goal for this study was to understand the impact of fidelity usage in the hopes of increasing fidelity usage of the platform across schools. This efficacy analysis examines the impact of Zearn Math, implemented with fidelity, vs. no usage. The implications of Zearn's approach are discussed further in the limitations section.

[4] This potential for bias does not exclude a study from meeting WWC's Group Design Standards with Reservations as long as baseline equivalency can be established. According to WWC: "In QED studies, confounding is almost always a potential issue due to the selection of a sample, because some unobserved factors may have contributed to the outcome. The WWC accounts for this issue by not allowing a QED study to receive the highest rating" (What Works Clearinghouse, 2020).

[5] All treatment students in the long-term retention analysis, used Zearn Math with fidelity during the 2021-2022 school year and discontinued usage during the 2022-2023 school year.

TCAP.[6] Then, within strata, treatment students were matched to control students with whom they shared at least six of nine other demographic and academic characteristics: district, school, gender, race, economic disadvantage, English learner status, special education status, gifted status, and chronically out of school status.[7]

This optimal matching method utilized Bertsekas' auction algorithm to produce combinatorial optimization such that treatment individuals were matched to others closest to them in the control pool and, when controls were the best-fit match for more than one treatment individual, the pairing went to the individual from whom the next best pairing was the farthest (Bertsekas, 1981; Rosenbaum, 2020).[8]

If a treatment student had no match within their academic year,[9] grade, and score strata with whom they shared at least six characteristics, they were excluded from the analysis. The caliper that limited match difference to no more than three characteristics was selected to maximize inclusion in the sample, prevent biasing through uneven patterns of exclusion and still ensure similarity between groups.

For more information on Zearn's methodological approach, see ***Efficacy Analysis Methodology: Zearn's approach to Coarsened Exact Matching***.

Difference of Means

Once consistent Zearn Math users were matched to a similar group of non-users, a difference of means analysis was conducted to quantify the impact of Zearn Math on student achievement. Means were calculated for treatment and control groups overall as well as for groups disaggregated by starting

[6] For the CEM one-year analysis and dose response analysis, the 2021-2022 school year pre scores are from the spring '21 TCAP; the 2022-2023 school year pre scores are from the spring '22 TCAP. For the long-term analysis, all pre scores are from the spring '21 TCAP.

[7] Chronically out of school is defined by Tennessee as missing 10% or more school days (Tennessee Department of Education, 2022a).

[8] In other words, if Control Student A was the best match for Treatment Student 1 and Treatment Student 2, sharing 6 out of 9 characteristics with each, Control Student A could still only be matched with either Treatment Student 1 or Treatment Student 2. If the next best match for Treatment Student 1, Control Student B, shared 4 characteristics, and the next best match for Treatment Student 2, Control Student C, shared 5 characteristics, then Treatment Student 1 would be matched with Control A and Treatment Student 2 would be matched with Control C. In this way, the algorithm of optimal matching balances the closeness of any individual match with its impact on the closeness of the overall group match.

[9] Academic year does not apply for long-term analysis. See footnote 5 for further explanation.

math Performance Level and demographic subgroup.[10] Because pairs of consistent Zearn Math users and non-users were allowed to mismatch on up to three demographic characteristics, subgroups did not always align on starting scale scores. Therefore, differences in achievement by demographic subgroup were reported as difference-in-difference, rather than as raw scores.

Academic growth was measured as the change in math scores between the pre and post TCAP administrations.[11] The TCAP has four Performance Levels: Below Expectations (1), Approaching Expectations (2), Met Expectations (3), and Exceeded Expectations (4). Students scoring at "Met Expectations" or "Exceeded Expectations" are considered proficient (Tennessee Department of Education, 2023). Outcomes are reported in terms of change in scale score, change in Performance Level, percent of students reaching proficiency on the post TCAP, and change in this percent between pre and post assessment administrations.

Difference in means t-tests were run on the average academic growth of the treatment group vs. the average academic growth of the control group to determine if the impact of Zearn Math was statistically significant. Given *SD*=standard deviations and *n*=number of observations per group, t-tests were conducted as:

$$t = \frac{mean_{treatment} - mean_{control}}{\sqrt{\frac{SD^2_{treatment}}{n_{treatment}} + \frac{SD^2_{control}}{n_{control}}}}$$

Effect size was calculated with *Cohen's d* which divides the difference in means between treatment and control by the pooled standard deviations:

$$Cohen's\ d = \frac{mean_{treatment} - mean_{control}}{pooled\ SD}$$

[10] For the CEM one-year analysis, results disaggregated for 2021-2022 and 2022-2023 are located in Appendix A Tables A14, A15, A17, A18, A20, A21 and A24-A27.

[11] For the CEM one-year analysis and dose response analysis, the 2021-2022 school year pre scores are from the spring '21 TCAP and post scores are from the spring '22 TCAP; the 2022-2023 school year pre scores are from the spring '22 TCAP and post scores are from the the spring '23 TCAP. For the long-term analysis, all pre scores are from the spring '21 TCAP and post scores are from the spring '23 TCAP.

CEM Dose-Response Analysis

This analysis examines the impact of Zearn Math when used at four lesson completion thresholds: 3+ lessons per month yet less than one lesson per week (i.e., 25-29 lessons per year, the TN ALL Corps recommendation), 1-2 lessons per week (i.e., 30-59 lessons per year), 2-3 lessons per week (i.e., 60-89 lessons per year), and 3+ lessons per week (i.e., 90+ lessons per year, Zearn's recommended usage). Using CEM, students in each threshold were matched to a control group of students who completed zero lessons but had Zearn accounts. Matching between treatment and control students was done separately for each treatment threshold, but the same demographic and score calipers were used to limit the distance between matches.

Findings represent the average difference in achievement between students using at each threshold and their matched pair from the control pool of students. There are notable implications of separately matching each threshold to a control group of students. First, the dosage thresholds generally do not have identical control groups. Rather the control groups represent the best match for students at that usage threshold. This accommodates situations in which the demographic and starting proficiencies differ between groups thereby maintaining the quasi-experimental robustness that is derived from comparing similar treatment and control students.

In addition, conducting separate rounds of matching means the same student may appear as a control for students in more than one threshold if that student is the best match for students in both groups. While matching treatment-to-control within a defined treatment benchmark is done without replacement, matching is done to the same pool of controls so this duplication may occur. In other words, findings represent the outcomes of four separate quasi-experiments in which matching criteria are standardized across all four.

As noted in the "Methodology" section, using CEM, treatment students were put into matching strata with control students that were in the same academic year, same grade, and within 5 scale score points on both the math and ELA pre TCAP. Then, within strata, treatment students were matched to control students with whom they shared at least six of nine other demographic and academic characteristics: district, school, gender, race, economic disadvantage, English learner status, special education status, gifted status, and chronically out of school status. This matching criterion was used for all analyses in this paper (see Appendix A Table A2 for details on the matched samples used for analysis and Tables A3-A5 and A9-A11 for sample breakdown).

Scale Score Growth

Across all usage thresholds, Zearn Math users gained more scale score points than matched non-users between the pre- and post-TCAP administrations. Consistent Zearn Math users, those who completed the recommended 3+ lessons per week, saw the largest gains relative to non-users. This group of students gained 7.2 additional scale score points on the TCAP, more than two-times the additional

gains of students who completed 2-3 lessons per week, an exponential increase (see Appendix A Tables A8 and A22 for findings from the difference in means analysis and Results Table 1).

RESULTS TABLE 1

Growth by Zearn Math Dosage

Scale score gains for consistent Zearn Math users (Treatment) vs. non-users (Control), by Zearn Math dosage

	Zearn Users	Non-Users	Difference
3 lessons per month (TN ALL Corps recommended)	3.4	1.6	1.8
1-2 lessons per week	3.9	1.5	2.4
2-3 lessons per week	3.6	0.5	3.1
3+ lessons per week (Zearn recommended)	6.7	-0.5	7.2

CEM One-Year Impact Analysis

The next analysis examines in more depth the impact of Zearn Math for consistent users. Results are disaggregated by starting Performance Level and demographic and academic subgroups. Impact is also explored as changes in percent proficient in addition to changes in scale score points. Although there were slight differences in starting scores among some demographic subgroups, all groups met baseline equivalency to be included in this analysis (see Appendix A Tables A9-A11 for sample breakdown and Table A12 for a comparison between sample and state demographics).[12] [13]

Scale Score Growth

Across all students and each starting Performance Level, consistent Zearn Math users gained more scale score points than matched non-users between the pre and post TCAP administrations. On average, consistent Zearn Math users gained 6.7 scale score points, while matched non-users lost 0.5 of a scale score point, a difference of 7.2 points (effect size = .23; see Appendix A Table A22 for findings

[12] Subgroups of female, male, Black and/or Hispanic, and economically disadvantaged had baseline differences < .05 of a standard deviation, which satisfies baseline equivalence without adjustment, according to WWC. All other subgroups had baseline differences < .25 of a standard deviation, satisfying baseline equivalency with a difference-in-difference adjustment (2022). Subgroup analyses were not conducted for English learners due to the subgroup having less than 10 students (see Appendix A Table A13 for full details on baseline equivalence).

[13] The sample population of fidelity users differs from the state population, having a proportionally smaller population of students from disadvantaged backgrounds. The implications of this difference are discussed in the limitations section.

from the difference in means analysis and Results Table 2). Gains were highest among consistent Zearn Math users who started the year "Below Expectations." These students gained 21.3 scale score points, while non-users who started the year "Below Expectations" gained 7.5 points, a difference of 13.8 points (effect size = .42; see Appendix A Table A22 for findings from the difference in means analysis and Results Table 2).

RESULTS TABLE 2

Growth Across TCAP Performance Levels

Scale score gains for consistent Zearn Math users (Treatment) vs. non-users (Control), by starting Performance Level

	Consistent Users	Non-Users	Difference
All Students	6.7	-0.5	7.2
Below Expectations	21.3	7.5	13.8
Approaching Expectations	10.3	1.0	9.3
Met and Exceeded Expectations	1.1	-3.3	4.4

NOTE: Summary report presentation of this data may vary slightly from the detailed chart above due to rounding for visual simplification.

Across all subgroups, consistent Zearn Math users gained scale score points between the pre and post assessment administrations, while non-users lost points. Differences were even larger across traditionally disadvantaged subgroups of students including Black and/or Hispanic students, economically disadvantaged students, students with disabilities, and students chronically out of school. Notably, students with disabilities with consistent Zearn Math usage grew the most, gaining 11.6 scale score points, while matched non-users lost 4.6 points, a difference of 16.3 points (effect size = .52; see Appendix A Table A23 for findings from the difference in means analysis and Results Table 3).

RESULTS TABLE 3

Growth Across Subgroups

Scale score gains for consistent Zearn Math users vs. non-users, by subgroup

	Consistent Users	Non-Users	Difference
All Students	6.7	-0.5	7.2
Male	7.1	-0.7	7.8
Female	6.4	-0.3	6.7
Black and/or Hispanic	9.3	-3.1	12.4
Economically disadvantaged	6.8	-2.2	9.0
Students with disabilities	11.6	-4.6	16.3
Gifted	**	**	**
Chronically out of school	5.2	-5.7	10.8

NOTE: Summary report presentation of this data may vary slightly from the detailed chart above due to rounding for visual simplification.
**Excluded due to lack of statistical significance. Full results are available in Appendix A Table A23.

Performance Level Mobility

Mobility models compared the change in Performance Level for treatment and control students based on their starting Performance Level. Across all Performance Levels, consistent Zearn Math users maintained or increased their Performance Levels at higher rates than non-users. Notably, among consistent Zearn Math users who started at the “Below Expectations” level, 64% improved their Performance Level, compared to only 43% of non-users. Similarly, among consistent Zearn Math users who started at the “Approaching Expectations” level, 44% improved their Performance Level, compared to only 30% of non-users (see Results Table 4).

RESULTS TABLE 4

Change in TCAP Performance Level

Post TCAP Performance Level change relative to starting TCAP Performance Level

Starting TCAP Performance Level	Zearn usage	Decrease	Maintain	Increase
Below Expectations	Consistent Users	0%	36%	64%
	Non-Users	0%	57%	43%
Approaching Expectations	Consistent Users	6%	49%	44%
	Non-Users	14%	56%	30%
Met and Exceeded Expectations*	Consistent Users	25%	75%	0%
	Non-Users	29%	71%	0%

*Of the students starting at "Met Expectations", 22% of Zearn Math users moved up to "Exceeded Expectations" compared to 18% of non-users.

Across all students and each subgroup, consistent Zearn Math met proficiency ("Met Expectations" and "Exceeded Expectations") on the post TCAP at higher rates than non-users, despite having the same starting proficiency rates. Across all students, 63.9% of Zearn Math users met proficiency, compared to 55.7% of non-users, a difference of 8.1%. Differences were even more pronounced for traditionally disadvantaged subgroups of Black and/or Hispanic students and economically disadvantaged students, with consistent Zearn Math users reaching proficiency at 1.2-1.4x the rate of non-users (see Results Table 5).[14] Further details are located in Appendix A Table A19.

[14] The percentages displayed in Results Table 5 reflect the mean-adjusted post values. First, an adjusted spring baseline mean was calculated as the average percent of Zearn Math users and non-users starting at proficiency ("Met Expectations" or "Exceeded Expectations"). Then the actual percentage-point change in those reaching proficiency between pre and post was added to the adjusted baseline mean to obtain the adjusted post percentages for treatment and control students. This corrected for subgroups that met baseline equivalence but did not have identical starting percentages of proficiency, so that outcomes could be compared accurately (see Appendix A Table A19 for the unadjusted and adjusted pre and post values and further details).

RESULTS TABLE 5

Proficiency Rate Across Subgroups

Percent of students reaching proficiency on the post TCAP for consistent Zearn Math users vs. non-users, by subgroup*

	Consistent Users	Non-Users	Difference	Relative Likelihood
All Students	63.9%	55.7%	8.1%	1.15
Male	65.9%	56.9%	9.0%	1.16
Female	61.7%	54.5%	7.2%	1.13
Black and/or Hispanic	53.4%	37.9%	15.6%	1.41
Economically disadvantaged	49.0%	40.5%	8.5%	1.21
Students with disabilities	**	**	**	**
Gifted	**	**	**	**
Chronically out of school	**	**	**	**

*The values displayed reflect the adjusted percentages (see footnote 14 for a description of this calculation).
**Excluded due to lack of statistical significance for change in the percent reaching proficiency. Full results are available in Appendix A Table A23.
NOTE: Summary report presentation of this data may vary slightly from the detailed chart above due to rounding for visual simplification.

Across all students and each subgroup, consistent Zearn Math users had larger gains in the percent of reaching proficiency between the pre and post TCAP administrations than non-users. On average, consistent Zearn Math users saw an 11.8 percentage point increase, while non-users saw a 3.6 percentage point increase, a difference of 8.1 percentage points. Notably, Black and/or Hispanic students with consistent Zearn Math usage saw the largest gains. These students gained 18.7 percentage points, compared to only 3.1 percentage points for non-users, a difference of 15.6 percentage points (see Results Table 6).[15] Full results are available in Appendix A Tables A22 and A23.

[15] For each subgroup in treatment and control, the percent reaching proficiency on the pre TCAP was subtracted from the percent reaching proficiency on the post TCAP. This change is depicted in Results Table 6. If the percent within a subgroup was the same for pre and post, the change listed in Results Table 6 would be 0. The pre and post percentages are reported in Appendix A Table A19.

RESULTS TABLE 6

Change in Proficiency Rate Across Subgroups

Change in percent of students reaching proficiency between the pre and post TCAP for consistent Zearn Math users vs. non-users, by subgroup

	Consistent Users	Non-Users
All Students	11.8%	3.6%
Male	13.9%	4.9%
Female	9.6%	2.4%
Black and/or Hispanic	18.7%	3.1%
Economically disadvantaged	11.9%	3.4%
Students with disabilities	**	**
Gifted	**	**
Chronically out of school	**	**

**Excluded due to lack of statistical significance. Full results are available in Appendix A Table A23.

Two-Year Retention of Academic Gains

Results from the one-year CEM analysis show promising evidence that using Zearn Math consistently has a robust impact on student growth for Tennessee students. In this section, models were run to examine the extent to which consistent Zearn Math users from 2021-2022 school year retained additional learnings through 2022-2023, despite being low- or non-users during the 2022-2023 school year.

As states consider investments in education interventions to catch students up and move them forward in math, it is important that students retain their gains in learning beyond a year. Therefore, longitudinal success of Zearn Math in increasing student learning, even beyond a student's time on the learning platform, is an important measure of efficacy.

This analysis focused on the two-year outcomes of a matched sample of students with consistent Zearn Math usage during the 2021-2022 school year and low or no usage during the 2022-2023 school year, and a similar group of students with low or no usage during the 2021-2022 and 2022-2023 school year.[16] Additionally students needed to have pre-math and ELA TCAP scores from spring '21, and post-math TCAP scores from spring '22 and spring '23.

[16] Non-users are defined as those who completed fewer than 30 lessons per year and have an account on Zearn Math.

For this analysis, the matched sample was selected using the same matching process, calipers and criteria as the CEM one-year efficacy analysis. In total, 177 students with consistent Zearn Math usage during the 2021-2022 school year met the criteria for inclusion in the matching pool. Of the 177 consistent Zearn Math users, all but 7 were matched, creating a sample of 170 matched pairs.[17] Treatment and control groups differed by an average of 1.8 demographic factors, 0.32 in starting math and 0.34 in starting ELA scale score points.[18] The 7 students excluded from the study, due to lack of match, did not concentrate in any demographic category that would bias the sample (see Appendix A Table A28 for breakdown of sample demographics). Although there were slight differences in starting scores among some demographic subgroups, all groups met baseline equivalency to be included in this analysis.[19]

Students with consistent Zearn Math usage in 2021-2022 maintained and even increased their growth, relative to low- or non-users, a year after they stopped using Zearn Math. In this two-year impact sample, students who consistently used Zearn Math during the 2021-2022 school year, ended the 2021-2022 school year 7.9 points higher than matched low- or non-users (similar to the 7.2 scale score point difference measured in the sample used for the one-year impact analysis). At the end of the 2022-2023 school year, students who consistently used Zearn Math during the 2021-2022 school year ended the 2022-2023 school year 10.2 points higher than matched low- or non-users (effect size=.31; see Appendix A Table A31 for findings from the difference in means analysis and Results Table 7).

RESULTS TABLE 7

Additional Scale Score Growth Retained After Discontinuing Use

Additional growth in scale score points retained by students who consistently used Zearn Math during the 2021-2022 school year, relative to matched peers with no usage

	EOY 2021	EOY2022
All Students	7.9	10.2

[17] Note: The sample for this analysis is <350.

[18] Mean pretest math scores differed by 0.32 points. This is less than 0.05 of a standard deviation of the combined means. According to WWC, “Baseline differences less than or equal to 0.05 standard deviations in absolute value automatically satisfy the baseline equivalence standard and do not require statistical adjustment” (WWC, 2022, p. 53). See Appendix A Table A29 for full details on baseline equivalence.

[19] Subgroups of female, male, and economically disadvantaged had baseline differences < .05 of a standard deviation, which satisfies baseline equivalence without adjustment, according to WWC (2022).Subgroup analyses were not conducted for Black and/or Hispanic students, English learners, students with disabilities, gifted, or students chronically out of school due to these subgroups having less than 10 students (see Appendix A Table 29 for full details on baseline equivalence).

Conclusion and Limitations

This analysis provides promising evidence of Zearn Math's positive impact on student achievement. A dose-response analysis of the impact of Zearn Math on student achievement at four lesson completion thresholds found that, across all usage thresholds, Zearn Math users had larger academic gains than matched non-users as measured by the TCAP. Consistent Zearn Math users, those who completed the recommended 3+ lessons per week, saw the largest gains relative to non-users.

A second analysis examining additional impacts of consistent Zearn Math usage found that, in addition to positive changes in student performance overall, students who started below standards, Black and/or Latino students, economically disadvantaged students, students with disabilities, and students chronically out of school who consistently used Zearn Math saw even larger gains[20] than the average student. The finding that Zearn Math impacts all students positively, but is associated with even more growth among those starting below standards or traditionally disadvantaged students, further substantiates findings from efficacy analyses of Zearn Math's impact in other states or districts (2022a, 2022b, 2022c; Szatrowski, 2022a, 2022b, 2022c; Szatrowski et al., 2022; Rickel, 2023; Rickel et al., 2023).

A separate analysis showed that students who used Zearn Math in 2021-2022, then discontinued use for 2022-2023, maintained and even increased their additional growth relative to matched non-users at the end of 2022-2023. This means that the benefits from Zearn Math held even beyond a student's time on the platform. This supports that students who learn on the Zearn Math platform better retain their learnings as measured by the TCAP.

By matching students closely on starting TCAP scores in both math and ELA, grade, and nine demographic and academic factors, treatment and control groups were similar along major confounding characteristics. This technique better isolated the impact of Zearn Math usage as an explanatory factor for differences in academic growth and performance than less rigorous correlational analyses and meet the rigorous standards set by What Works Clearinghouse and ESSA evidence-based intervention guidelines. For both students overall and traditionally disadvantaged subgroups, Zearn Math usage appears to drive higher levels of academic growth.

Despite the strong findings from this analysis, there are some limitations. While quasi-experimental methods allow researchers to control for observed confounders, there is a possibility that unobserved confounders mediate the relationship between Zearn use and academic performance. Eliminating this limitation entirely would require implementation of a randomized controlled trial for Zearn usage.

This study was conducted on a sub-population of students in Tennessee. It is possible that the impact of Zearn Math in other locations, or across a larger number of students, might show a different effect

[20] Refers to either gains in scale score or change in the percent of students reaching proficiency.

size, whether larger or smaller. This sample may not be completely representative of Tennessee as a whole. For instance, in the one-year impact analysis, students from traditionally disadvantaged groups such as Black and Hispanic students, economically disadvantaged students, English learners, students with disabilities, and students chronically out of school were underrepresented in the sample relative to the total population across Tennessee, which is a factor that may impact outcomes. It is also possible that there are features specific to the districts represented in this sample that facilitate large gains with Zearn Math usage that may not be present in other schools or districts. The geographic specificity of this study may limit the generalizability to a more nationally representative population.

This study's findings of Zearn Math's efficacy align with those from other district and state efficacy analyses (Zearn 2022a, 2022b, 2022c; Szatrowski, 2022a, 2022b, 2022c; Szatrowski et al., 2022; Rickel, 2023, Rickel et al., 2023). With robust methods and the expansion of efficacy studies to multiple districts across the country, continued replication of trends and findings will provide even stronger evidence of Zearn Math's efficacy moving forward. Zearn plans to continue this work over the coming months and years.

References

Bertsekas, D.P. (1981). A new algorithm for the assignment problem. *Math Programming*, *21*, 152–71. https://doi.org/10.1007/BF01584237

Blackwell, M., Iacus, S., King, G., & Porro, G. (2009). cem: Coarsened exact matching in Stata. *Stata Journal*, *9*(4), 524-546. https://doi.org/10.1177/1536867X0900900402

Iacus, S. M., King, G., & Porro, G. (2011). Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association*, *106*(493), 345–361. https://doi.org/10.1198/jasa.2011.tm09599

McCoy, C.E. (2017). Understanding the intention-to-treat principle in randomized controlled trials. *The Western Journal of Emergency Medicine*, *18*(6), 1075–1078. https://doi.org/10.5811/westjem.2017.8.35985

REL Midwest. (2019). *ESSA tiers of evidence: What you need to know*. Regional Educational Laboratory at American Institutes for Research. https://ies.ed.gov/ncee/edlabs/regions/midwest/pdf/blogs/RELMW-ESSA-Tiers-Video-Handout-508.pdf

Rickel, J. (2023). *Efficacy analysis of Zearn Math: Findings from implementation in a large Southern district.* Zearn. https://webassets.zearn.org/Implementation/LargeSouthernDistrictTechnicalPaper.pdf

Rickel, J., Szatrowski, A., & Rosemond, C. (2023). *Efficacy analysis of Zearn Math: Findings from implementation in a large Southern district.* Zearn. https://webassets.zearn.org/Implementation/DCPSTechnicalPaper.pdf

Rosenbaum, P. R. (2020). Modern algorithms for matching in observational studies. *Annual Review of Statistics and Its Application*, *7*(1), 143-176. https://dx.doi.org/10.1146/annurev-statistics-031219-041058

Stuart, E. A., & Rubin, D.B. (2008). Matching with multiple control groups with adjustment for group differences. *Journal of Educational and Behavioral Statistics*, *33*(3), 279–306. https://doi.org/10.3102/1076998607306078

Szatrowski, A. (2022a). *Efficacy analysis of Zearn Math: Findings from a large Midwestern urban district.* Zearn.

https://webassets.zearn.org/Implementation/MidwesternUrbanDistrictTechnicalPaper.pdf

Szatrowski, A. (2022b). *Efficacy analysis of Zearn Math in Nebraska*. Zearn. https://webassets.zearn.org/Implementation/NebraskaTechnicalAppendix.pdf

Szatrowski, A. (2022c). *Efficacy analysis of Zearn Math in Tulsa*. Zearn. https://webassets.zearn.org/Implementation/TulsaTechnicalPaper.pdf

Szatrowski, A., Rickel, J., & Rosemond, C. (2022). *Efficacy Analysis of Zearn Math in DC Public Schools.* Zearn. https://webassets.zearn.org/Implementation/DCPSTechnicalPaper.pdf

Tennessee Department of Education. (2022). *About the State.* TDOE. https://tdepublicschools.ondemand.sas.com/state/about

Tennessee Department of Education. (2022a). *Performance Indicators.* TDOE. https://tdepublicschools.ondemand.sas.com/state/performance/truancy

Tennessee Department of Education. (2023, June 29). *TDOE Releases 2022-23 State-Level TCAP Results* [Press Release]. TDOE. https://www.tn.gov/education/news/2023/6/29/tdoe-releases-2022-23-state-level-tcap-results-.html

What Works Clearinghouse. (2020). *What Works Clearinghouse standards handbook, version 4.1*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance (NCEE). https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Standards-Handbook-v4-1-508.pdf

What Works Clearinghouse. (2022). *What Works Clearinghouse procedures and standards handbook, version 5.0*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance (NCEE). https://ies.ed.gov/ncee/wwc/Docs/referenceresources/Final_WWC-HandbookVer5.0-0-508.pdf

Zearn. (2022a). *Consistent Zearn usage dramatically reduces learning loss.* Zearn. https://webassets.zearn.org/Implementation/Zearn_Impact_for_Students_Below_Grade_Level.pdf

Zearn. (2022b). *Economically disadvantaged students, multilingual learners, and Black and Latino students experience double-digit increases in proficiency on 2022 end-of-year assessments with*

consistent Zearn usage. Zearn. https://webassets.zearn.org/Implementation/EfficacyResearch_Zearn_Impact_for_Student_Subgroups.pdf

Zearn. (2022c). *For students who scored below grade level in math, consistent Zearn usage tied to 2 grade levels of growth in 2 years of pandemic learning—nearly double the gains of curriculum alone.* Zearn. https://webassets.zearn.org/Implementation/Zearn_Impact_for_Students_Below_Grade_Level.pdf

Appendix A

Table A1

Consistent Zearn Math users (i.e., 3+ lessons/week, approximately 90+ lessons/year), by grade and year*		
	2021-2022	2022-2023
Grade 4	248	335
Grade 5	131	224
Grade 6	25	445
Grade 7	No students included in the analysis	106
Grade 8	No students included in the analysis	76

*There were 1,485 unique fidelity users across both years, however, 105 students used Zearn Math w/fidelity for two years, therefore totals add up to 1,590.
NOTE: not all students were matched and included in the analysis.

Table A2

Matched sample details by Zearn Math dosage						
	Grade 4-8 users	Grade 4-8 users w/TCAP scores	Matched pairs	Demographic difference	Math TCAP pre-score difference (scale score points)	ELA TCAP pre-score difference (scale score points)
Threshold						
3 lessons per month (TN ALL Corps recommended)	3,909	2,076	1,982	1.69	0.19	0.01
1-2 lessons per week	10,484	6,416	6,126	1.71	0.02	0.00
2-3 lessons per week	4,155	2,774	2,673	1.71	0.08	0.09
3+ lessons per week (Zearn recommended)	2,578	1,590	1,537	1.83	0.10	0.05

Table A3

Dose Response 25 to 29 lessons: breakdown of sample matching characteristics, 2021-2022 and 2022-2023		
	Treatment	**Control**
Total N's	1,982	1,982
Pre-scores (Spring '21 or Spring '22 assessment scores)		
Math scale score	321.71	321.89
ELA scale score	331.62	331.63
Starting Performance Level (N's)		
Below	488	485
Approaching	867	851
Met	506	517
Exceeded	121	129
Grade Level (N's)		
Grade 4	465	465
Grade 5	522	522
Grade 6	361	361
Grade 7	373	373
Grade 8	261	261

Table A4

Dose Response 30 to 59 lessons: breakdown of sample matching characteristics, 2021-2022 and 2022-2023		
	Treatment	**Control**
Total N's	6,126	6,126
Pre-scores (Spring '21 or Spring '22 assessment scores)		
Math scale score	324.84	324.86
ELA scale score	333.80	333.81
Starting Performance Level (N's)		
Below	1,390	1,408
Approaching	2,584	2,545
Met	1,664	1,696
Exceeded	488	477
Grade Level (N's)		
Grade 4	1,392	1,392
Grade 5	1,399	1,399
Grade 6	1,500	1,500
Grade 7	885	885
Grade 8	950	950

Table A5

Dose Response 60 to 89 lessons: breakdown of sample matching characteristics, 2021-2022 and 2022-2023		
	Treatment	**Control**
Total N's	2,673	2,673
Pre-scores (Spring '21 or Spring '22 assessment scores)		
Math scale score	330.08	330.16
ELA scale score	337.68	337.59
Starting Performance Level (N's)		
Below	463	455
Approaching	1,131	1,123
Met	837	851
Exceeded	242	244
Grade Level (N's)		
Grade 4	589	589
Grade 5	438	438
Grade 6	672	672
Grade 7	524	524
Grade 8	450	450

Table A6

Dose Response: pre and post TCAP scale score means, across all students, 2021-2022 and 2022-2023*							
	Treatment pre	Treatment post	Control pre	Control post	Starting mean difference	Pooled SD	Difference in SDs**
All Students							
25 to 29 lessons: math scale score	321.71	325.07	321.89	323.47	-0.19	35.78	-0.01
30 to 59 lessons: math scale score	324.84	328.75	324.86	326.39	-0.02	35.90	0.00
60 to 89 lessons: math scale score	330.08	333.72	330.16	330.69	-0.08	35.20	0.00

*For the 2021-2022 school year, pre scores are from spring '21 and post scores are from spring '22. For the 2022-2023 school year, pre scores are from spring '22 and post scores are from spring '23.
**According to WWC, baseline differences <.05 of a standard deviation satisfy baseline equivalence without adjustment. Differences <.25 of a standard deviation satisfy baseline equivalence with adjustment of difference-in-difference (2022).

Table A7

Dose Response: pre and post adjusted TCAP scale score means, across all students, 2021-2022 and 2022-2023*			
	Treatment and control adjusted baseline mean**	Treatment adjusted post***	Control adjusted post***
All Students			
25 to 29 lessons: math scale score	321.80	325.16	323.37
30 to 59 lessons: math scale score	324.85	328.76	326.38
60 to 89 lessons: math scale score	330.12	333.76	330.65

*For the 2021-2022 school year, pre scores are from spring '21 and post scores are from spring '22. For the 2022-2023 school year, pre scores are from spring '22 and post scores are from spring '23.
**The adjusted baseline mean was calculated as the average of treatment and control students' pre score.
*** The adjusted post scores for treatment and control were calculated by adding the actual score change between pre and post for each respective group to the adjusted baseline mean. See Appendix A Table A6 for the unadjusted scores; see Appendix A Table A8 for the actual change in score.

Table A8

Dose Response: comparison of changes in scores between consistent Zearn Math users and non-users, across all students, 2021-2022 and 2022-2023

	Treatment change in mean	Control change in mean	Difference	Pooled SD	Cohen's d
All Students					
25 to 29 lessons: math scale score	3.36	1.57	1.79*	33.18	0.05
30 to 59 lessons: math scale score	3.91	1.53	2.38***	32.20	0.07
60 to 89 lessons: math scale score	3.64	0.53	3.11***	31.52	0.08

*p < .05 **p < .01 ***p < .001

Table A9

CEM one-year impact analysis: breakdown of sample matching characteristics, 2021-2022 and 2022-2023		
	Treatment	**Control**
Total N's	1,537	1,537
Pre-scores (Spring '21 or Spring '22 assessment scores)		
Math scale score	339.84	339.74
ELA scale score	345.03	344.98
Starting Performance Level (N's)		
Below	174	172
Approaching	562	565
Met	540	537
Exceeded	261	263
Grade Level (N's)		
Grade 4	559	559
Grade 5	347	347
Grade 6	454	454
Grade 7	105	105
Grade 8	72	72
Demographic & academic subgroups (N's)		
Male	777	774
Female	760	763
Black and/or Hispanic	166	162
Economically disadvantaged	294	291
English learners	*	*
Students with disabilities	80	76
Gifted	15	16
Chronically out of school	95	93

*Subgroup analyses were not conducted because there were fewer than 10 students.

Table A10

CEM one-year impact analysis: breakdown of sample matching characteristics, 2021-2022		
	Treatment	**Control**
Total N's	390	390
Pre-scores (Spring '21 assessment scores)		
Math scale score	330.14	330.12
ELA scale score	346.23	346.10
Starting Performance Level (N's)		
Below	65	65
Approaching	164	163
Met	121	118
Exceeded	40	44
Grade Level (N's)		
Grade 4	237	237
Grade 5	129	129
Grade 6	24	24
Grade 7	No students included in the analysis	No students included in the analysis
Grade 8	No students included in the analysis	No students included in the analysis
Demographic & academic subgroups (N's)		
Male	201	201
Female	189	189
Black and/or Hispanic	30	30
Economically disadvantaged	76	74
English learners	*	*
Students with disabilities	24	24
Gifted	*	*
Chronically out of school	12	13

*Subgroup analyses were not conducted because there were less than 10 students.

Table A11

CEM one-year impact analysis: breakdown of sample matching characteristics, 2022-2023		
	Treatment	**Control**
Total N's	1,147	1,147
Pre-scores (Spring '22 assessment scores)		
Math scale score	343.13	343.01
ELA scale score	344.62	344.61
Starting Performance Level (N's)		
Below	109	107
Approaching	398	402
Met	419	419
Exceeded	221	219
Grade Level (N's)		
Grade 4	322	322
Grade 5	218	218
Grade 6	430	430
Grade 7	105	105
Grade 8	72	72
Demographic & academic subgroups (N's)		
Male	576	573
Female	571	574
Black and/or Hispanic	136	132
Economically disadvantaged	218	217
English learners	*	*
Students with disabilities	56	52
Gifted	*	*
Chronically out of school	83	80

*Subgroup analyses were not conducted because there were fewer than 10 students.

Table A12

Tennessee Schools: comparison of CEM one-year impact sample and statewide school population			
	Sample-Treatment	**Sample-Control**	**State***
Subgroups			
Black	5%	6%	24%
Hispanic	5%	5%	13%
White	88%	89%	60%
Economically disadvantaged	19%	19%	30%
English learners	**	**	8%
Gifted	**	**	N/A
Students with disabilities	5%	5%	13%
Chronically out of school	6%	6%	20%

*While the matched sample represents students in grades 4-8, the state percentages are based off of K-12.
**Suppressed because the subgroup had less than 10 students.

Table A13

CEM one-year impact analysis: pre and post TCAP scale score means, across all students and by subgroup, 2021-2022 and 2022-2023*

	Treatment pre	Treatment post	Control pre	Control post	Starting mean difference	Pooled SD	Difference in SDs**
All Students							
Math scale score	339.84	346.57	339.74	339.23	0.10	34.30	0.00
Starting Proficiency							
Below-proficient ("Below" and "Approaching")	311.78	324.67	311.66	314.18	0.12	22.23	0.01
Proficient and above ("Met" and "Exceeded")	365.62	366.69	365.61	362.31	0.01	20.31	0.00
Starting Performance Level							
Below	280.14	301.43	279.95	287.42	0.20	19.73	0.01
Approaching	321.58	331.87	321.32	322.33	0.26	11.12	0.02
Met	354.24	359.28	354.38	353.42	-0.15	10.49	-0.01
Exceeded	389.16	382.03	388.53	380.45	0.63	15.20	0.04
Grade Level							
Grade 4	341.23	342.41	341.10	333.67	0.13	32.36	0.00
Grade 5	326.17	348.43	326.20	338.96	-0.03	36.25	0.00
Grade 6	345.46	347.59	345.37	342.68	0.09	34.02	0.00
Grade 7	351.50	357.69	350.99	351.70	0.50	23.77	0.02
Grade 8	342.42	347.22	342.51	343.79	-0.10	35.35	0.00

*For the 2021-2022 school year, pre scores are from spring '21 and post scores are from spring '22. For the 2022-2023 school year, pre scores are from spring '22 and post scores are from spring '23.
**According to WWC, baseline differences <.05 of a standard deviation satisfy baseline equivalence without adjustment. Differences <.25 of a standard deviation satisfy baseline equivalence with adjustment of difference-in-difference (2022).

Table A13 cont.

Cont. CEM one-year impact analysis: pre and post TCAP scale score means, across all students and by subgroup, 2021-2022 and 2022-2023*

	Treatment pre	Treatment post	Control pre	Control post	Starting mean difference	Pooled SD	Difference in SDs**
Demographic & academic subgroups							
Male	340.77	347.86	340.50	339.78	0.27	34.42	0.01
Female	338.89	345.25	338.97	338.67	-0.09	34.16	0.00
Black and/or Hispanic	327.65	336.90	327.62	324.52	0.03	29.08	0.00
Economically disadvantaged	327.48	334.28	327.56	325.33	-0.08	32.49	0.00
Students with disabilities	300.63	312.26	299.00	294.37	1.63	31.73	0.05
Gifted	383.93	400.20	387.88	399.94	-3.94	24.41	-0.16
Chronically out of school	333.69	338.85	331.71	326.03	1.99	30.81	0.06

*For the 2021-2022 school year, pre scores are from spring '21 and post scores are from spring '22. For the 2022-2023 school year, pre scores are from spring '22 and post scores are from spring '23.
**According to WWC, baseline differences <.05 of a standard deviation satisfy baseline equivalence without adjustment. Differences <.25 of a standard deviation satisfy baseline equivalence with adjustment of difference-in-difference (2022).

Table A14

CEM one-year impact analysis: pre and post TCAP scale score means, across all students and by subgroup, 2021-2022

	Treatment spring '21	Treatment spring '22	Control spring '21	Control spring '22	Starting mean difference	Pooled SD	Difference in SDs*
All Students							
Math scale score	330.14	338.70	330.12	330.83	0.03	33.99	0.00
Starting Proficiency							
Below-proficient	308.35	320.78	308.07	311.36	0.28	24.00	0.01
Proficient and above	361.14	364.20	361.15	358.23	-0.01	18.22	0.00
Starting Performance Level							
Below	278.03	300.91	277.62	282.12	0.42	21.17	0.02
Approaching	320.36	328.65	320.21	323.01	0.15	10.69	0.01
Met	352.88	359.63	352.31	350.61	0.58	10.75	0.05
Exceeded	386.13	378.03	384.86	378.68	1.26	12.13	0.10
Grade Level							
Grade 4	333.84	333.46	333.82	328.24	0.01	29.14	0.00
Grade 5	316.43	343.26	316.48	329.28	-0.05	35.63	0.00
Grade 6	367.38	366.00	366.79	364.67	0.58	33.18	0.02
Demographic & academic subgroups							
Male	330.71	341.70	330.81	332.79	-0.10	33.32	0.00
Female	329.53	335.51	329.38	328.75	0.16	34.72	0.00
Black and/or Hispanic	318.50	335.50	319.23	320.17	-0.73	23.25	-0.03
Economically disadvantaged	317.24	331.29	317.62	319.73	-0.38	36.76	-0.01
Students with disabilities	294.79	307.63	294.46	291.13	0.33	32.94	0.01
Chronically out of school	321.58	325.92	319.77	313.00	1.81	26.86	0.07

*According to WWC, baseline differences <.05 of a standard deviation satisfy baseline equivalence without adjustment. Differences <.25 of a standard deviation satisfy baseline equivalence with adjustment of difference-in-difference (2022).

Table A15

CEM one-year impact analysis: pre and post TCAP scale score means, across all students and by subgroup, 2022-2023

	Treatment spring '22	Treatment spring '23	Control spring '22	Control spring '23	Starting mean difference	Pooled SD	Difference in SDs*
All Students							
Math scale score	343.13	349.24	343.01	342.09	0.12	33.79	0.00
Starting Proficiency							
Below-proficient	313.33	326.43	313.27	315.45	0.06	21.20	0.00
Proficient and above	366.74	367.32	366.74	363.34	0.00	20.66	0.00
Starting Performance Level							
Below	281.40	301.73	281.36	290.64	0.04	18.73	0.00
Approaching	322.08	333.19	321.77	322.05	0.31	11.26	0.03
Met	354.63	359.17	354.97	354.21	-0.34	10.37	-0.03
Exceeded	389.71	382.75	389.26	380.81	0.45	15.65	0.03
Grade Level							
Grade 4	346.68	349.00	346.46	337.66	0.22	33.55	0.01
Grade 5	331.93	351.49	331.95	344.69	-0.02	35.41	0.00
Grade 6	344.23	346.56	344.18	341.45	0.06	33.68	0.00
Grade 7	351.50	357.69	350.99	351.70	0.50	23.77	0.02
Grade 8	342.42	347.22	342.51	343.79	-0.10	35.35	0.00
Demographic & academic subgroups							
Male	344.27	350.01	343.90	342.24	0.38	34.14	0.01
Female	341.98	348.47	342.13	341.94	-0.15	33.41	0.00
Black and/or Hispanic	329.67	337.21	329.53	325.52	0.14	29.92	0.00
Economically disadvantaged	331.05	335.33	330.95	327.24	0.10	30.15	0.00
Students with disabilities	303.13	314.25	301.10	295.87	2.03	31.05	0.07
Chronically out of school	335.45	340.72	333.65	328.15	1.80	31.04	0.06

*According to WWC, baseline differences <.05 of a standard deviation satisfy baseline equivalence without adjustment. Differences <.25 of a standard deviation satisfy baseline equivalence with adjustment of difference-in-difference (2022).

Table A16

CEM one-year impact analysis: pre and post adjusted TCAP scale score means, across all students and by subgroup, 2021-2022 and 2022-2023*

	Treatment and control adjusted baseline mean**	Treatment adjusted post ***	Control adjusted post***
All Students			
Math scale score	339.79	346.52	339.28
Starting Proficiency			
Below-proficient	311.72	324.61	314.24
Proficient and above	365.61	366.69	362.31
Starting Performance Level			
Below	280.05	301.33	287.52
Approaching	321.45	331.74	322.46
Met	354.31	359.35	353.35
Exceeded****	388.84	381.71	380.77
Grade Level			
Grade 4	341.17	342.35	333.74
Grade 5	326.18	348.44	338.95
Grade 6	345.42	347.55	342.72
Grade 7 ****	351.24	357.43	351.95
Grade 8****	342.47	347.27	343.74

*For the 2021-2022 school year, pre scores are from spring '21 and post scores are from spring '22. For the 2022-2023 school year, pre scores are from spring '22 and post scores are from spring '23.
**The adjusted baseline mean was calculated as the average of treatment and control students' pre score.
***The adjusted post scores for treatment and control were calculated by adding the actual score change between pre and post for each respective group to the adjusted baseline mean. See Appendix A Table A13 for the unadjusted scores; see Appendix A Tables A22 and A23 for the actual change in score.
****The difference in scale score change was not significant. Full results are available in Appendix A Tables A22 and A23.

Table A16 cont.

Cont. CEM one-year impact analysis: pre and post adjusted TCAP scale score means, across all students and by subgroup, 2021-2022 and 2022-2023*

	Treatment and control adjusted baseline mean**	Treatment adjusted post***	Control adjusted post***
Demographic & academic subgroups			
Male	340.63	347.72	339.92
Female	338.93	345.29	338.63
Black and/or Hispanic	327.64	336.89	324.54
Economically disadvantaged	327.52	334.32	325.29
Students with disabilities	299.81	311.45	295.18
Gifted****	385.90	402.17	397.97
Chronically out of school	332.70	337.86	327.02

*For the 2021-2022 school year, pre scores are from spring '21 and post scores are from spring '22. For the 2022-2023 school year, pre scores are from spring '22 and post scores are from spring '23.
**The adjusted baseline mean was calculated as the average of treatment and control students' pre score.
***The adjusted post scores for treatment and control were calculated by adding the actual score change between pre and post for each respective group to the adjusted baseline mean. See Appendix A Table A13 for the unadjusted scores; see Appendix A Tables A22 and A23 for the actual change in score.
****The difference in scale score change was not significant. Full results are available in Appendix A Tables A22 and A23.

Table A17

CEM one-year impact analysis: pre and post adjusted TCAP scale score means, across all students and by subgroup, 2021-2022

	Treatment and control adjusted spring '21 baseline mean*	Treatment adjusted spring '22**	Control adjusted spring '22**
All Students			
Math scale score	330.13	338.69	330.84
Starting Proficiency			
Below-proficient	308.21	320.64	311.49
Proficient and above	361.15	364.20	358.23
Starting Performance Level			
Below	277.82	300.70	282.33
Approaching	320.28	328.58	323.09
Met	352.59	359.34	350.90
Exceeded***	385.49	377.39	379.31
Grade Level			
Grade 4	333.83	333.45	328.25
Grade 5	316.45	343.28	329.25
Grade 6***	367.08	365.71	364.96
Demographic & academic subgroups			
Male	330.76	341.75	332.74
Female	329.46	335.43	328.83
Black and/or Hispanic	318.87	335.87	319.80
Economically disadvantaged	317.43	331.48	319.54
Students with disabilities***	294.63	307.46	291.29
Chronically out of school***	320.68	325.01	313.91

*The adjusted baseline mean was calculated as the average of treatment and control students' spring '21 score.
**The adjusted spring '22 scale scores for treatment and control were calculated by adding the actual score change between spring '21 and '22 for each respective group to the spring '21 adjusted baseline mean. See Appendix A Table A14 for the unadjusted scores; see Appendix A Tables A24 and A25 for the actual change in score.
***The difference in scale score change was not significant. Full results are available in Appendix A Tables A24 and A25.

Table A18

CEM one-year impact analysis: pre and post adjusted TCAP scale score means, across all students and by subgroup, 2022-2023

	Treatment and control adjusted spring '22 baseline mean*	Treatment adjusted spring '23**	Control adjusted spring '23**
All Students			
Math scale score	343.07	349.18	342.15
All Students			
Below-proficient	313.30	326.40	315.48
Proficient and above	366.74	367.32	363.34
Starting Performance Level			
Below	281.38	301.71	290.66
Approaching	321.92	333.04	322.21
Met	354.80	359.34	354.04
Exceeded***	389.49	382.53	381.03
Grade Level			
Grade 4	346.57	348.89	337.77
Grade 5	331.94	351.50	344.68
Grade 6	344.21	346.53	341.48
Grade 7***	351.24	357.43	351.95
Grade 8***	342.47	347.27	343.74

*The adjusted baseline mean was calculated as the average of treatment and control students' spring '22 score.
**The adjusted spring '23 scale scores for treatment and control were calculated by adding the actual score change between spring '22 and '23 for each respective group to the spring '22 adjusted baseline mean. See Appendix A Table A15 for the unadjusted scores; see Appendix A Tables A26 and A27 for the actual change in score.
***The difference in scale score change was not significant. Full results are available in Appendix A Tables A26 and A27.

Table A18 cont.

Cont. CEM one-year impact analysis: pre and post adjusted TCAP scale score means, across all students and by subgroup, 2022-2023

	Treatment and control adjusted spring '22 baseline mean*	Treatment adjusted spring '23**	Control adjusted spring '23**
Demographic & academic subgroups			
Male	344.08	349.82	342.43
Female	342.06	348.55	341.87
Black and/or Hispanic	329.60	337.14	325.58
Economically disadvantaged	331.00	335.28	327.29
Students with disabilities	302.11	313.24	296.88
Chronically out of school	334.55	339.83	329.05

*The adjusted baseline mean was calculated as the average of treatment and control students' spring '22 score.
**The adjusted spring '23 scale scores for treatment and control were calculated by adding the actual score change between spring '22 and '23 for each respective group to the spring '22 adjusted baseline mean. See Appendix A Table A15 for the unadjusted scores; see Appendix A Tables A26 and A27 for the actual change in score.
***The difference in scale score change was not significant. Full results are available in Appendix A Tables A26 and A27.

Table A19

CEM one-year impact analysis: percent proficient on the pre and post TCAP, across all students and by subgroup, 2021-2022 and 2022-2023*

	Treatment pre	Treatment post	Control pre	Control post	Treatment and control adjusted baseline mean**	Treatment adjusted post***	Control adjusted post***
All Students							
Math percent proficient	52.11%	63.89%	52.05%	55.69%	52.1%	63.9%	55.7%
Grade Level							
Grade 4	48.30%	64.76%	48.66%	55.10%	48.48%	64.94%	54.92%
Grade 5	45.24%	61.96%	45.24%	52.74%	45.24%	61.96%	52.74%
Grade 6****	57.49%	59.91%	56.61%	54.85%	57.05%	59.47%	55.29%
Grade 7****	68.57%	79.05%	68.57%	69.52%	68.57%	79.05%	69.52%
Grade 8****	56.94%	69.44%	58.33%	59.72%	57.64%	70.14%	59.03%
Demographic & academic subgroups							
Male	52.12%	66.02%	51.94%	56.85%	52.03%	65.93%	56.94%
Female	52.11%	61.71%	52.16%	54.52%	52.13%	61.74%	54.49%
Black and/or Hispanic	34.34%	53.01%	35.19%	38.27%	34.76%	53.44%	37.85%
Economically disadvantaged	37.07%	48.98%	37.11%	40.55%	37.09%	49.00%	40.53%
Students with disabilities****	10.00%	27.50%	6.58%	14.47%	8.29%	25.79%	16.18%
Gifted****	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Chronically out of school****	43.16%	53.68%	41.94%	45.16%	42.55%	53.07%	45.77%

*For the 2021-2022 school year, pre scores are from spring '21 and post scores are from spring '22. For the 2022-2023 school year, pre scores are from spring '22 and post scores are from spring '23.
**The adjusted baseline mean was calculated as the average percent of treatment and control students who met proficiency ("Met" or "Exceeded") on the pre assessment.
***The adjusted post values for treatment and control were calculated by adding the actual percentage point change between pre and post for each respective group to the adjusted baseline mean.
****The change in percent proficient was not significant. Full results are available in Appendix A Tables A22 and A23.

Table A20

CEM one-year impact analysis: percent proficient on the pre and post TCAP, across all students and by subgroup, 2021-2022

	Treatment spring '21	Treatment spring '22	Control spring '21	Control spring '22	Treatment and control adjusted spring '21 baseline mean*	Treatment adjusted spring '22**	Control adjusted spring '22**
All Students							
Math percent proficient	41.28%	56.67%	41.54%	47.95%	41.41%	56.79%	47.82%
Grade Level							
Grade 4***	40.51%	54.01%	40.08%	46.84%	40.30%	53.80%	47.05%
Grade 5	35.66%	57.36%	37.21%	44.19%	36.43%	58.14%	43.41%
Grade 6***	79.17%	79.17%	79.17%	79.17%	79.17%	79.17%	79.17%
Demographic & academic subgroups							
Male	41.79%	59.70%	42.79%	49.75%	42.29%	60.20%	49.25%
Female***	40.74%	53.44%	40.21%	46.03%	40.48%	53.17%	46.30%
Black and/or Hispanic***	13.33%	40.00%	16.67%	33.33%	15.00%	41.67%	31.67%
Economically disadvantaged***	31.58%	42.11%	35.14%	37.84%	33.36%	43.88%	36.06%
Students with disabilities	8.33%	20.83%	8.33%	4.17%	8.33%	20.83%	4.17%
Chronically out of school***	25.00%	50.00%	23.08%	30.77%	24.04%	49.04%	31.73%

*The adjusted baseline mean was calculated as the average percent of treatment and control students who met proficiency ("Met" or "Exceeded") on the spring '21 assessment.
**The adjusted post values for treatment and control were calculated by adding the actual percentage point change between spring '21 and '22 for each respective group to the spring '21 adjusted baseline mean.
***The change in percent proficient was not significant. Full results are available in Appendix A Tables A24 and A25.

Table A21

CEM one-year impact analysis: percent proficient on the pre and post TCAP, across all students and by subgroup, 2022-2023

	Treatment spring '22	Treatment spring '23	Control spring '22	Control spring '23	Treatment and control adjusted spring '22 baseline mean*	Treatment adjusted spring '23**	Control adjusted spring '23**
All Students							
Math percent proficient	55.80%	66.35%	55.62%	58.33%	55.71%	66.26%	58.41%
Grade Level							
Grade 4	54.04%	72.67%	54.97%	61.18%	54.50%	73.14%	60.71%
Grade 5***	50.92%	64.68%	50.00%	57.80%	50.46%	64.22%	58.26%
Grade 6***	56.28%	58.84%	55.35%	53.49%	55.81%	58.37%	53.95%
Grade 7***	68.57%	79.05%	68.57%	69.52%	68.57%	79.05%	69.52%
Grade 8***	56.94%	69.44%	58.33%	59.72%	57.64%	70.14%	59.03%
Demographic & academic subgroups							
Male	55.73%	68.23%	55.15%	59.34%	55.44%	67.94%	59.63%
Female	55.87%	64.45%	56.10%	57.32%	55.98%	64.56%	57.20%
Black and/or Hispanic	38.97%	55.88%	39.39%	39.39%	39.18%	56.09%	39.18%
Economically disadvantaged***	38.99%	51.38%	37.79%	41.47%	38.39%	50.77%	42.08%
Students with disabilities****	10.71%	30.36%	5.77%	19.23%	8.24%	27.88%	21.70%
Chronically out of school***	45.78%	54.22%	45.00%	47.50%	45.39%	53.83%	47.89%

*The adjusted baseline mean was calculated as the average percent of treatment and control students who met proficiency ("Met" or "Exceeded") on the spring '22 assessment.
**The adjusted post values for treatment and control were calculated by adding the actual percentage point change between spring '22 and '23 for each respective group to the spring '22 adjusted baseline mean.
***The change in percent proficient was not significant. Full results are available in Appendix A Tables A26 and A27.

Table A22

CEM one-year impact analysis: comparison of changes in scores and percent proficient between consistent Zearn Math users and non-users, across all students, 2021-2022 and 2022-2023

	Treatment change in mean	Control change in mean	Difference	Pooled SD	Cohen's d
All Students					
Math scale score (SS)	6.73	-0.51	7.24***	31.32	0.23
Math percent proficient	11.78%	3.64%	8.13%***	0.54	0.15
Starting Proficiency					
Below-proficient SS	12.89	2.52	10.37***	25.70	0.40
Proficient and above SS	1.07	-3.30	4.37***	24.25	0.18
Starting Performance Level					
Below SS	21.28	7.47	13.81***	32.85	0.42
Approaching SS	10.29	1.01	9.28***	22.68	0.41
Met SS	5.04	-0.96	6.00***	22.74	0.26
Exceeded SS	-7.13	-8.08	0.94	25.94	0.04
Grade Level					
Grade 4 SS	1.18	-7.43	8.61***	23.71	0.36
Grade 4 percent proficient	16.46%	6.44%	10.02%***	0.46	0.22
Grade 5 SS	22.26	12.76	9.50***	26.02	0.36
Grade 5 percent proficient	16.71%	7.49%	9.22%**	0.46	0.20
Grade 6 SS	2.13	-2.69	4.83**	23.15	0.21
Grade 6 percent proficient	2.42%	-1.76%	4.19%	0.43	0.10
Grade 7 SS	6.19	0.70	5.49	23.52	0.23
Grade 7 percent proficient	10.48%	0.95%	9.52%	0.47	0.20
Grade 8 SS	4.81	1.28	3.53	23.08	0.15
Grade 8 percent proficient	12.50%	1.39%	11.11%	0.39	0.29

*p < .05 **p < .01 ***p < .001

Table A23

CEM one-year impact analysis: comparison of changes in scores and percent proficient between consistent Zearn Math users and non-users, by subgroup, 2021-2022 and 2022-2023

	Treatment change in mean	Control change in mean	Difference	Pooled SD	Cohen's d
Demographic & academic subgroups					
Male SS	7.09	-0.72	7.81***	26.08	0.30
Male percent proficient	13.90%	4.91%	8.99%***	0.46	0.20
Female SS	6.36	-0.30	6.66***	24.57	0.27
Female percent proficient	9.61%	2.36%	7.25%**	0.44	0.16
Black and/or Hispanic SS	9.25	-3.10	12.35***	25.11	0.49
Black and/or Hispanic percent proficient	18.67%	3.09%	15.59%**	0.49	0.32
Economically disadvantaged SS	6.81	-2.23	9.03***	26.14	0.35
Economically disadvantaged percent proficient	11.90%	3.44%	8.47%*	0.47	0.18
Students with disabilities SS	11.64	-4.63	16.27**	31.50	0.52
Students with disabilities percent proficient	17.50%	7.89%	9.61%	0.35	0.27
Gifted SS	16.27	12.06	4.20	29.95	0.14
Gifted percent proficient	0.00%	0.00%	0.00%	0.00	0.00
Chronically out of school SS	5.16	-5.68	10.84**	24.37	0.44
Chronically out of school percent proficient	10.53%	3.23%	7.30%	0.46	0.16

*p < .05 **p < .01 ***p < .001

Table A24

CEM one-year impact analysis: comparison of changes in scores and percent proficient between consistent Zearn Math users and non-users, across all students, 2021-2022

	Treatment change in mean	Control change in mean	Difference	Pooled SD	Cohen's d
All Students					
Math scale score (SS)	8.56	0.71	7.85***	33.36	0.24
Math percent proficient	15.38%	6.41%	8.97%**	0.56	0.16
Starting Proficiency					
Below-proficient SS	12.43	3.29	9.14***	26.60	0.34
Proficient and above SS	3.06	-2.91	5.97*	25.56	0.23
Starting Performance Level					
Below SS	22.88	4.51	18.37**	36.08	0.51
Approaching SS	8.29	2.80	5.49*	21.35	0.26
Met SS	6.74	-1.69	8.44**	24.54	0.34
Exceeded SS	-8.10	-6.18	-1.92	27.16	-0.07
Grade Level					
Grade 4 SS	-0.38	-5.58	5.20*	22.51	0.23
Grade 4 percent proficient	13.50%	6.75%	6.75%	0.47	0.14
Grade 5 SS	26.83	12.80	14.03***	27.74	0.51
Grade 5 percent proficient	21.71%	6.98%	14.73%**	0.46	0.32
Grade 6 SS	-1.38	-2.13	0.75	19.77	0.04
Grade 6 percent proficient	0.00%	0.00%	0.00%	0.00	0.00

*p < .05 **p < .01 ***p < .001

Table A25

CEM one-year impact analysis: comparison of changes in scores and percent proficient between consistent Zearn Math users and non-users, by subgroup, 2021-2022

	Treatment change in mean	Control change in mean	Difference	Pooled SD	Cohen's d
Demographic & academic subgroups					
Male SS	10.99	1.98	9.01***	26.73	0.34
Male percent proficient	17.91%	6.97%	10.95%*	0.47	0.23
Female SS	5.98	-0.63	6.61*	26.02	0.25
Female percent proficient	12.70%	5.82%	6.88%	0.44	0.16
Black and/or Hispanic SS	17.00	0.93	16.07**	23.34	0.69
Black and/or Hispanic percent proficient	26.67%	16.67%	10.00%	0.45	0.22
Economically disadvantaged SS	14.05	2.11	11.94*	31.26	0.38
Economically disadvantaged percent proficient	10.53%	2.70%	7.82%	0.46	0.17
Students with disabilities SS	12.83	-3.33	16.17	30.35	0.53
Students with disabilities percent proficient	12.50%	-4.17%	16.67%*	0.29	0.58
Chronically out of school SS	4.33	-6.77	11.10	17.79	0.62
Chronically out of school percent proficient	25.00%	7.69%	17.31%	0.47	0.37

*p < .05 **p < .01 ***p < .001

Table A26

CEM one-year impact analysis: comparison of changes in scores and percent proficient between consistent Zearn Math users and non-users, across all students, 2022-2023

	Treatment change in mean	Control change in mean	Difference	Pooled SD	Cohen's d
All Students					
Math scale score (SS)	6.11	-0.93	7.03***	30.61	0.23
Math percent proficient	10.55%	2.70%	7.85%***	0.54	0.15
Starting Proficiency					
Below-proficient SS	13.10	2.17	10.92***	25.29	0.43
Proficient and above SS	0.58	-3.40	3.97**	23.91	0.17
Starting Performance Level					
Below SS	20.33	9.27	11.06**	30.83	0.36
Approaching SS	11.12	0.29	10.83***	23.21	0.47
Met SS	4.55	-0.75	5.30***	22.21	0.24
Exceeded SS	-6.96	-8.46	1.50	25.73	0.06
Grade Level					
Grade 4 SS	2.32	-8.80	11.12***	24.57	0.45
Grade 4 percent proficient	18.63%	6.21%	12.42%***	0.46	0.27
Grade 5 SS	19.56	12.74	6.81**	24.87	0.27
Grade 5 percent proficient	13.76%	7.80%	5.96%	0.45	0.13
Grade 6 SS	2.33	-2.73	5.05**	23.33	0.22
Grade 6 percent proficient	2.56%	-1.86%	4.42%	0.44	0.10
Grade 7 SS	6.19	0.70	5.49	23.52	0.23
Grade 7 percent proficient	10.48%	0.95%	9.52%	0.47	0.20
Grade 8 SS	4.81	1.28	3.53	23.08	0.15
Grade 8 percent proficient	12.50%	1.39%	11.11%	0.39	0.29

*p < .05 **p < .01 ***p < .001

Table A27

CEM one-year impact analysis: comparison of changes in scores and percent proficient between consistent Zearn Math users and non-users, by subgroup, 2022-2023

	Treatment change in mean	Control change in mean	Difference	Pooled SD	Cohen's d
Demographic & academic subgroups					
Male SS	5.73	-1.66	7.39***	25.76	0.29
Male percent proficient	12.50%	4.19%	8.31%**	0.46	0.18
Female SS	6.49	-0.19	6.68***	24.08	0.28
Female percent proficient	8.58%	1.22%	7.36%**	0.44	0.17
Black and/or Hispanic SS	7.54	-4.02	11.56***	25.35	0.46
Black and/or Hispanic percent proficient	16.91%	0.00%	16.91%**	0.50	0.34
Economically disadvantaged SS	4.28	-3.71	7.98***	23.83	0.34
Economically disadvantaged percent proficient	12.39%	3.69%	8.70%	0.48	0.18
Students with disabilities SS	11.13	-5.23	16.36**	32.13	0.51
Students with disabilities percent proficient	19.64%	13.46%	6.18%	0.37	0.17
Chronically out of school SS	5.28	-5.50	10.78**	25.26	0.43
Chronically out of school percent proficient	8.43%	2.50%	5.93%	0.46	0.13

*p < .05 **p < .01 ***p < .001

Table A28

CEM two-year retention of academic gains - 1 year fidelity and 1 year no usage: breakdown of sample matching characteristics in 2021-2022		
	Treatment	**Control**
Total N's	170	170
Pre-scores (Spring '21 assessment scores)		
Math scale score	336.74	337.06
ELA scale score	345.82	346.16
Starting Performance Level (N's)		
Below	22	22
Approaching	67	67
Met	55	54
Exceeded	26	27
Grade Level (N's)		
Grade 4	95	95
Grade 5	53	53
Grade 6**	*	*
Grade 7**	*	*
Grade 8	No students included in the analysis	No students included in the analysis
Demographic & academic subgroups (N's)		
Male	83	82
Female	87	88
Black and/or Hispanic	No students included in the analysis	No students included in the analysis
Economically disadvantaged	39	38
English learners	*	*
Students with disabilities	*	*
Gifted	*	*
Chronically out of school	*	*

*Subgroup analyses were not conducted because there were fewer than 10 students.
**Complementary suppression applied to prevent calculation of the group with primary suppression.

Table A29

CEM two-year retention of academic gains - 1 year fidelity and 1 year no usage: pre and post TCAP scale score means, across all students and by subgroup

	Treatment spring '21	Treatment spring '23	Control spring '21	Control spring '23	Starting mean difference	Pooled SD	Difference in SDs*
All Students							
Math scale score	336.74	352.28	337.06	342.44	-0.32	33.81	-0.01
Starting Proficiency							
Below-proficient	311.74	329.73	312.10	318.35	-0.36	22.42	-0.02
Proficient and above	364.20	377.06	364.48	368.91	-0.28	20.18	-0.01
Starting Performance Level							
Below	280.09	305.59	280.86	298.68	-0.77	19.23	-0.04
Approaching	322.13	337.66	322.36	324.81	-0.22	10.71	-0.02
Met	352.84	369.24	352.78	357.09	0.06	10.79	0.01
Exceeded	388.23	393.62	387.89	392.56	0.34	12.93	0.03
Grade Level							
Grade 4	338.56	351.82	338.80	341.47	-0.24	27.97	-0.01
Grade 5	320.43	344.96	320.89	332.09	-0.45	34.20	-0.01
Demographic & academic subgroups							
Male	334.99	352.28	335.09	338.90	-0.10	31.59	0.00
Female	338.40	352.29	338.90	345.74	-0.50	35.78	-0.01
Economically disadvantaged	322.62	335.41	323.55	323.84	-0.94	32.62	-0.03

*According to WWC, baseline differences <.05 of a standard deviation satisfy baseline equivalence without adjustment. Differences <.25 of a standard deviation satisfy baseline equivalence with adjustment of difference-in-difference (2022).

Table A30

CEM two-year retention of academic gains - 1 year fidelity and 1 year no usage: pre and post adjusted TCAP scale score means, across all students and by subgroup

	Treatment and control adjusted spring '21 baseline mean*	Treatment adjusted spring '23**	Control adjusted spring '23**
All Students			
Math scale score	336.90	352.44	342.28
All Students			
Below-proficient	311.92	329.91	318.17
Proficient and above	364.34	377.20	368.77
Starting Performance Level			
Below***	280.48	305.98	298.30
Approaching	322.25	337.77	324.69
Met	352.81	369.21	357.12
Exceeded***	388.06	393.44	392.73
Grade Level			
Grade 4	338.68	351.94	341.35
Grade 5	320.66	345.19	331.87
Demographic & academic subgroups			
Male	335.04	352.33	338.85
Female***	338.65	352.54	345.49
Economically disadvantaged	323.08	335.88	323.37

*The adjusted baseline mean was calculated as the average of treatment and control students' spring '21 score.
**The adjusted spring '23 scale scores for treatment and control were calculated by adding the actual score change between spring '21 and '23 for each respective group to the spring '21 adjusted baseline mean. See Appendix A Table A29 for the unadjusted scores; see Appendix A Table A31 for the actual change in score.
***The difference in scale score change was not significant. Full results are available in Appendix A Table A31.

Table A31

CEM two-year retention of academic gains - 1 year fidelity and 1 year no usage: comparison of changes in scores between consistent Zearn Math users and non-users

	Treatment change in mean	Control change in mean	Difference	Pooled SD	Cohen's d
All Students					
Math scale score (SS)	15.55	5.38	10.16***	33.25	0.31
Starting Proficiency					
Below-proficient SS	17.99	6.25	11.74**	28.78	0.41
Proficient and above SS	12.86	4.43	8.43*	24.25	0.35
Starting Performance Level					
Below SS	25.50	17.82	7.68	31.71	0.24
Approaching SS	15.52	2.45	13.07**	15.52	0.84
Met SS	16.40	4.31	12.09**	24.01	0.50
Exceeded SS	4.58	5.03	-0.46	24.58	-0.02
Grade Level					
Grade 4 SS	13.26	2.67	10.59**	27.43	0.39
Grade 5 SS	24.53	11.21	13.32**	25.39	0.52
Demographic & academic subgroups					
Male SS	17.29	3.82	13.47**	28.24	0.48
Female SS	13.89	6.84	7.04	25.32	0.28
Economically disadvantaged SS	12.79	0.29	12.51*	28.45	0.44

*p < .05 **p < .01 ***p < .001

Appendix B

This study was designed to meet the rigorous standards set by the What Works Clearinghouse (WWC) and qualify as an Every Student Succeeds Act (ESSA) evidence-based intervention. This Appendix provides more detail about these standards and how this impact study meets those standards.

What Works Clearinghouse provides ratings of randomized control trials (RCTs) and quasi-experimental designs (QEDs) against their Group Design standards. There are three possible ratings: Meets WWC Standards without Reservations, Meets WWC Standards with Reservations, or Does Not Meet WWC Standards. Because QED studies that establish baseline equivalence or use acceptable statistical adjustments "reduce, but likely do not eliminate, the potential bias associated with the group assignment procedures," Meets WWC Standards with Reservations is the highest possible rating for QEDs (What Works Clearinghouse, 2022).

This study uses quasi-experimental matching methods to create baseline equivalency between treatment and control groups along major confounding factors. Consistent Zearn Math users were matched with non-users, in the same grade, on starting math and English Language Arts (ELA) standardized test scores, along with nine student characteristics using a two-step Coarsened Exact Matching (CEM) method with optimal matching. CEM is a technique that simulates block sampling by matching students on covariates related both to a student's likelihood of using Zearn Math consistently and their academic performance (Blackwell et al., 2009; Iacus et al., 2011).

A QED study must satisfy several criteria to meet the WWC standard of "Meets WWC Standards with Reservations." The first is that the outcome measure "meets four standards: (1) face validity, (2) reliability, (3) not over aligned with the intervention, and (4) consistent data collection procedures" (What Works Clearinghouse, 2022). In this study, the primary outcome is math achievement on the Tennessee Comprehensive Assessment Program (TCAP). WWC considers standardized tests that are routinely administered in educational settings, like the TCAP, to meet these standards.

The next criteria is the elimination of confounding factors (What Works Clearinghouse, 2022). By matching fidelity users to non-users within five scale score points on their pre-score for both math and ELA on the TCAP, as well as at least six of nine other student characteristics: district, school, gender, race, economic disadvantage status, English learner status, special education status, gifted status, and chronically out of school status, the design of this study creates two groups that are academically and demographically similar on the most relevant and measurable confounding factors that would impact academic growth.

While CEM allows researchers to control for observed confounders, a possibility exists that there are unmeasured factors that differentiate the comparison groups of students who reach fidelity and those with no usage. For example, it is possible that an unmeasured characteristic allows fidelity users to

reach higher usage than would be possible for non-users. However, this type of unmeasurable attribute is what WWC refers to as “imperfect overlap in the characteristic between the conditions” which they term a selection mechanism, not a confounding factor (2020, p. 82).

This possibility of an unmeasured characteristic that could bias estimates is similar to an example provided by WWC of a program based on voluntary enrollment in which students who volunteer could differ from those who did not in hard-to-measure qualities like introversion vs. extroversion. It clarifies that “the WWC does not consider this to be a confounding factor, but the selection mechanism and potential difference in unmeasured characteristics are reasons that QEDs are limited to a rating of Meets WWC Group Design Standards with Reservations, if the baseline equivalence requirement is satisfied” (2020, p. 82).

The final criteria for a quasi-experimental study to meet WWC Standards with Reservations is illustrating baseline equivalence between treatment and control groups. This can be done with a pre-intervention measure that is the same as the outcome measure (2022). In this case, TCAP math scores are used as a pre-intervention measure of baseline equivalence and as the outcome measure of the study.

According to WWC, baseline differences < .05 of a standard deviation satisfy baseline equivalence without adjustment. Differences < .25 of a standard deviation satisfy baseline equivalent with statistical adjustment. Difference-in-difference is an acceptable statistical adjustment (2022). All groups across all analyses in this study meet the criteria for baseline equivalence either without or with adjustment.

The one-year CEM and dose response analyses have sample sizes that exceed 350. The long-term retention analysis did not (see Appendix B Table B1 for sample sizes for each analysis). In addition, the study must have been conducted in more than one school. This study spans 400 treatment schools, with an additional 779 schools of the 1,117 total control schools.

Table B1

Sample Size, by Subgroup, for All Quasi-Experimental Studies

Sample size of all Tennessee analyses			
	Treatment sample	**Control sample**	**Total sample**
One-year CEM	1,537	1,537	3,074
Dose Response 25-29 lessons	1,982	1,982	3,964
Dose Response 30-59 lessons	6,126	6,126	12,252
Dose Response 60-89 lessons	2,673	2,673	5,346
Long-term retention	170	170	340

Finally, findings must be statistically significant and there can be "no strong negative findings from experimental or quasi-experimental studies" (Regional Educational Laboratory at American Institutes for Research, 2019, p. 2). Results from this study show statistically significant positive impacts from the implementation of Zearn Math. There have been no strong negative findings from other experimental or quasi-experimental studies, while there have been statistically significant positive findings from other QED Zearn studies (see 2022a, 2022b, 2022c; Szatrowski, 2022a, 2022b, 2022c; Szatrowski et al., 2022; Rickel, 2023, Rickel et al., 2023).