Efficacy analysis of Zearn Math:
Findings from implementation in a large Southern district

January 2023

Jessica Rickel, M.A.
Zearn

Abstract

Analysis of consistent Zearn Math users and a comparable group of non-users in a large Southern district shows that students who used Zearn Math had higher levels of academic growth than similar students with no usage. This analysis used the quasi-experimental method, Coarsened Exact Matching (CEM), to examine the impact of Zearn Math across 1,111 students who completed an average of 3+ Zearn Math lessons per week, during the 2021-2022 school year and similarly matched peers who completed less than 1 lesson per week. Findings showed that consistent Zearn Math users grew an additional 14.6 scale score points on the state's standardized assessment; the equivalent of 1.3 years of learning as compared to non-users. Consistent Zearn Math users who started below standards grew an additional 20.9 scale points, in comparison to non-users who lost 4.4 scale score points; the equivalent of an additional 2.1 years of learning. Consistent Zearn Math users were also more likely to improve their Achievement Level and had a larger increase in the percent of students meeting standards. Results were consistent across historically marginalized subgroups.

Table of Contents

List of Figures and Tables

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 a large district in the South. 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 a district with over 80,000 students in grades PK-12, over 18,000 of whom are in grades four to six. The student body is 51% eligible for free or reduced lunch, 15% students in special education, 5% Multilingual learners (MLLs), 5% gifted and talented and 33% Black and/or Latino.

In grades 4-6, there were 1,134 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 2020-2021 and 2021-2022 school years. (See Appendix A Table A1 for a breakdown of sample demographics.)

This study was designed to meet the What Works Clearinghouse (WWC) Meets WWC Group Design Standards with Reservations and to meet an ESSA Tier 2 (Moderate) rating on the ESSA guidelines for evidence-based interventions. 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.)

Methodology

Quasi-experimental matching techniques were used to isolate the impact of Zearn Math on student achievement. Consistent Zearn Math users were matched to non-users on starting math and English Language Arts (ELA) assessment scores, along with 7 academic and demographic characteristics. The goal of matching was to create 1:1 pairings between similar students, differing primarily on Zearn Math usage during the 2021-2022 school year. 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 group was composed of students who consistently used Zearn Math during the 2021-2022 school year, 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 the district with no Zearn Math usage, operationalized as an average of fewer than one digital lesson per week; fewer than 30 digital lessons per year.[1]

[1] 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

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 was 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.[2]

Using CEM, treatment students were put into matching strata with control students that were in the same grade and within five scale score points on both the math and ELA spring 2021 assessments. Then, within strata, treatment students were matched to control students with whom they shared at least four of seven other demographic and academic characteristics: school, gender, race, free/reduced lunch eligibility, Multilingual learner status, special education status, and gifted status.

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,

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. with low to no usage. The implications of Zearn's approach are discussed further in the limitations section.

[2] 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).

2020).[3]

If a treatment student had no match within their grade and score strata with whom they shared at least 4 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***.

Out of the district's 1,134 consistent Zearn Math users all but 23 were matched. Treatment and control groups differed by an average of 1.27 demographic factors, 2.22 scale score points in starting math score and 2.23 scale score points in starting ELA score.[4] The 23 consistently using 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 A1 for breakdown of sample demographics).

Analysis

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 math Achievement Level and demographic factors.

Academic growth was measured as the change in scale score on the state's standardized assessment between the spring 2021 and spring 2022 assessment administrations. The assessment has five Achievement Levels: Inadequate (1), Below Satisfactory (2), Satisfactory (3), Proficient (4), and Mastery (5). Students scoring at Level 3, "Satisfactory", and above are considered passing and having met the standard. Outcomes are reported in terms of change in scale score, change in Achievement Level, and change in percent meeting standards.

[3] In other words, if Control Student A was the best match for Treatment Student 1 and Treatment Student 2, sharing 6 out of 7 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.

[4] Mean pretest math scores differed by .7 points. This is less than .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).

In addition to growth in scale score points, growth among consistent Zearn Math users and non-users was translated into years of learning. The average growth between grades 3-5 and 4-6 is 11.2 scale score points per year. A description of this calculation is discussed in the "Yearly Growth" section.

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:

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

Yearly Growth

Yearly growth expectations were calculated for each student based on the change in scale score points they were required to grow on the assessment to maintain the same Achievement Level in subsequent grades. On average, students were expected to grow 11.2 scale score points between 2021-2022.[5] Yearly growth expectations were also calculated for each academic and demographic subgroup and differ slightly from the overall sample.

On average, consistent Zearn Math users gained 14.7 scale score points whereas matched non-users gained .2 of a scale score point between the spring 2021 and spring 2022 assessment administrations, a difference of 14.6 points (effect size=0.85)[6]. (See Appendix A Table A3 for findings from the difference in means analysis.) This translates to an additional 1.3 years of growth for consistent Zearn Math users relative to non-users (see Results Table 1).

[5] Each student's expected growth is unique to their grade and starting Achievement Level.For example, for a third grader starting at Level 1, their expected growth would be the average of the difference between the score range for third grade Level 1 and fourth grade Level 1. Expected growth was averaged across all students to get an overall expected growth average, as well as by subgroup.

[6] Due to rounding, totals may not correspond to the difference of the separate figures.

Gains were highest among consistent Zearn Math users who started the year below standards. These students gained 16.6 scale score points, while non-users who started the year below standards lost 4.4 points, a difference of 20.9 points (effect size=1.1). (see Appendix A Table A3 for findings from the difference in means analysis.) This translated to an additional 2.1 years of growth for consistent Zearn Math users relative to matched non-users (see Results Table 1). The outsized impact of Zearn Math usage among students starting below standards has been a consistent finding across Zearn efficacy research (2022a, 2022b; Szatrowski, 2022a, 2022b, 2022c; Szatrowski et al., 2022).

RESULTS TABLE 1

Growth Across Achievement Levels

Growth in scale score for Consistent Zearn Math users *(Treatment)* vs. Non-users *(Control)*, by starting Achievement Level*

	All Students	Below Standards *(Levels 1 & 2)*	Meeting Standards *(Level 3)*	Above Standards *(Level 4 & 5)*
Treatment growth in scale score	14.7	16.6	16.5	12.8
Treatment years of growth	1.3	1.7	1.4	1.1
Control growth in scale score	0.2	-4.4	0.0	2.8
Control years of growth	0.0	-0.4	0.0	0.2
Growth difference in scale score	14.6	20.9	16.5	10.0
Years of growth difference	1.3	2.1	1.4	0.8

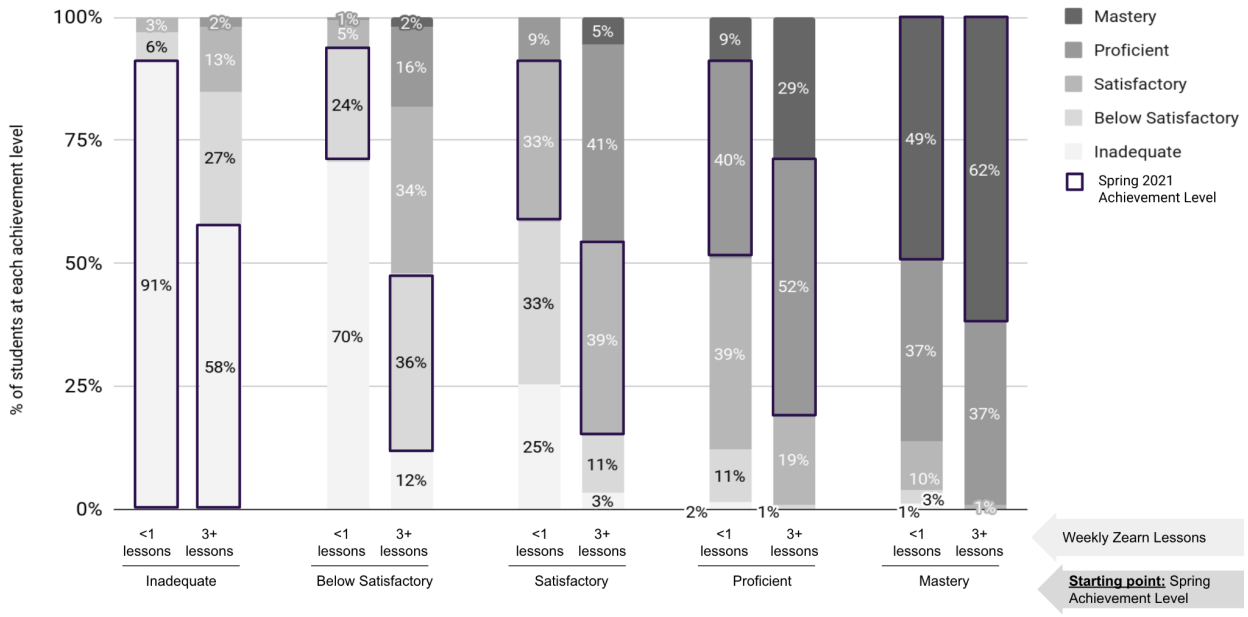
*On average, students in the sample were expected to grow 11.2 scale score points per year on the assessment in 3rd, 4th, and 5th grade. Disaggregated by starting Achievement Level, those who started below standards were expected to grow 9.8 scale score points per year, those who started at standard were expected to grow 11.4 points per year, and those who started above standards were expected to grow 11.8 points per year.

Mobility models compared the change in Achievement Level for treatment and control students based on their starting Achievement Level in spring 2021. Across all Achievement Levels, consistent Zearn Math users maintained or increased their Achievement Levels at higher rates than non-users.

Notably, among students who started at Achievement Level 1, “Inadequate”, and used Zearn Math with fidelity, 42% improved their Achievement Level, compared to only 9% of matched students with no usage. Similarly, among students who started at Achievement Level 2, “Below Satisfactory”, and used Zearn Math with fidelity, 52% improved their Achievement Level, compared to only 6% of students with no Zearn Math usage. Figure 1 illustrates the mobility between Achievement Levels, for consistent Zearn Math users and non-users.

FIGURE 1

Change in Achievement Level



Subgroup Analysis

In addition to capturing changes in student achievement across all users, the analysis examined how Zearn Math usage impacted the performance of student subgroups. 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[7], rather than as raw scores. (See Appendix A Table A2 for a breakdown of starting and ending means, by subgroup.)

Across all subgroups, consistent Zearn Math users saw gains in the percent of students meeting standards while non-users saw drops in this percentage between spring 2021 and spring 2022. On average, consistent Zearn Math users saw a 4.1 percentage point increase in the percent of students meeting standards, while non-users saw a 19.4 percentage point decrease. These differences were even larger across traditionally disadvantaged subgroups of students including: females, Black and/or Latino students, economically disadvantaged students, and students in special education. Notably,

[7] Subgroups of female and male had baseline differences <.05 of a standard deviation which satisfies baseline equivalence without adjustment, according to WWC. All other subgroups, with the exception of MLLs, had baseline differences <.25 of a standard deviation, satisfying baseline equivalency with a difference-in-difference adjustment (2022). MLL did not qualify for baseline equivalence with adjustment, therefore their results are not reported. (See Appendix A Table A2 for full details on baseline equivalence.)

Black and/or Latino consistent Zearn Math users had the largest increase in percent meeting standards across all subgroups.[8] (See Results Table 2 and Appendix A Tables A3 and A4 for more details.)

RESULTS TABLE 2

Change in Students Meeting Standards (Achievement Level 3) or Above

Change in percent meeting standards among consistent Zearn Math users vs. Non-users, by subgroup

	Consistent Users *(Treatment)*	Non-Users (Control)
All Students	4.1%	-19.4%
Female	4.9%	-18.8%
Male	3.5%	-19.8%
Black and/or Latino	8.1%	-19.6%
Economically disadvantaged	6.6%	-23.6%
MLL	++	++
Special Education	5.1%	-12.0%
Gifted	**	**

++Subgroup does not satisfy baseline equivalence even with statistical adjustment.

**Excluded due to lack of statistical significance. Full results available in Appendix A Table A4.

Across all subgroups, consistent Zearn Math users exceeded yearly learning benchmarks[9], on average, while non-users did not meet these benchmarks. Black and/or Latino and economically disadvantaged consistent Zearn Math users had the most growth at 1.4 years.[10] Notably, the largest difference in years of growth between consistent Zearn Math users and non-users was for economically disadvantaged students, among whom consistent Zearn Math users gained an additional 1.7 years of growth relative to economically disadvantaged students with no usage. (See Results Table 3.)

[8] For each subgroup in treatment and control, percent meeting standards in spring 2021 was subtracted from percent meeting standards in spring 2022. This change is depicted in Results Table 2. If the percent meeting standards within a subgroup was the same in spring 2021 and spring 2022 the change listed in Results Table 2 would be 0.

[9] Yearly growth expectations were calculated for each subgroup using the method described in footnote 5.

[10] Years of growth projections are based on each individual's starting Achievement Level and grade. Depending on the composition of students, it is possible for a subgroup to have smaller raw scale score point changes and larger yearly growth relative to another subgroup represented.

RESULTS TABLE 3

Growth Across Subgroups

Scale Score Gains Translated to Years of Growth for consistent Zearn Math users vs. Non-users, by subgroup*

	Consistent Zearn Math users		Non-users		Difference	
	Growth in Scale Score Points	Years of Growth	Growth in Scale Score Points	Years of Growth	Scale Score	Years
All Students	14.7	1.3	0.2	0.0	14.6	1.3
Female scale score growth	14.6	1.3	-0.2	0.0	14.8	1.3
Male scale score growth	14.8	1.3	0.5	0.0	14.3	1.3
Black and/or Latino scale score growth	15.0	1.4	0.2	0.0	14.8	1.4
Economically disadvantaged scale score growth	14.4	1.4	-3.8	-0.4	18.2	1.7
MLL scale score growth	++	++	++	++	++	++
Special education scale score growth	11.8	1.2	-2.7	-0.3	14.5	1.4
Gifted scale score growth	15.9	1.3	9.4	0.7	6.5	0.5

*On average, students were expected to grow 11.2 scale score points per year on the assessment. Disaggregated by subgroup, yearly growth was 11.1 for females, 11.2 for males, 10.9 for Black and/or Latino students, 10.5 for economically disadvantaged students, 10.1 for special education students, and 12.6 for gifted students.

++Subgroup does not satisfy baseline equivalence even with statistical adjustment.

Conclusion and Limitations

This analysis provides promising evidence of Zearn Math's positive impact on student achievement. In addition to positive changes in student performance overall, students who started below standards, Black and/or Latino students, economically disadvantaged students, and students in special education who consistently used Zearn Math saw even larger gains[11] 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 districts (2022a, 2022b, 2022c; Szatrowski, 2022a, 2022b, 2022c; Szatrowski et al., 2022).

By matching students closely on starting assessment scores in both math and ELA, grade, and seven 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

[11] Refers to either gains in scale score or change in the percent of students meeting standards.

correlational analyses and meets the criteria for Meets WWC Standards with Reservation and ESSA Tier 2 (see Appendix B for more details). For both students overall and 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 one district. It is possible that the impact of Zearn Math in other locations, or across a larger number of students, might show a different effect size, whether larger or smaller. This sample may not be completely representative of the district as a whole and there may be features specific to the district that facilitate large gains with Zearn Math usage that may not be present in other districts. The geographic specificity of this study may also 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). 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

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

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). *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

Zearn. (2022c). *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

Appendix A

Table A1

Breakdown of sample matching characteristics		
	Treatment	**Control**
Total N's	1,111	1,111
Pre-scores (Spring '21 assessment scores)		
Math scale score	316.41	315.73
ELA scale score	316.07	315.98
Starting Achievement Level (N's)		
Below Standards (Levels 1 & 2)	260	286
Meeting Standards (Level 3)	315	311
Above Standards (Levels 4 & 5)	536	514
Grade Level (N's)		
Grade 4	618	618
Grade 5	373	373
Grade 6	120	120
Demographic & academic subgroups (N's)		
Female	513	520
Male	598	591
Black and/or Latino	271	271
Economically disadvantaged	410	403
MLL	23	10
Special education	98	83
Gifted	188	107

Table A2

Spring 2021 and Spring 2022 performance means by subgroup							
	Treatment Spring 2021	**Treatment Spring 2022**	**Control Spring 2021**	**Control Spring 2022**	**Starting Mean Difference**	**Pooled SD**	**Difference in SDs***
All Students							
Math scale score	316.41	331.12	315.73	315.88	0.68	20.50	0.03
Starting Achievement Level							
Below Standards	292.15	308.71	292.68	288.31	-0.54	12.69	-0.04
Meeting Standards	310.21	326.69	310.39	310.38	-0.18	9.50	-0.02
Above Standards	331.82	344.60	331.79	334.55	0.03	13.93	0.00
Grade Level							
Grade 4	310.13	328.91	309.43	315.14	0.70	18.08	0.04
Grade 5	323.72	334.30	323.05	316.61	0.67	20.86	0.03
Grade 6	326.04	332.64	325.48	317.45	0.57	19.87	0.03
Demographic & academic subgroups							
Female	313.50	328.14	313.01	312.80	0.49	20.71	0.02
Male	318.90	333.68	318.13	318.59	0.78	20.01	0.04
Black and/or Latino	311.29	326.25	310.21	310.39	1.08	19.67	0.05
Economically disadvantaged	312.81	327.25	310.11	306.32	2.71	20.34	0.13
MLL	301.65	317.52	291.20	287.50	10.45	18.42	0.57+
Special education	303.96	315.77	301.88	299.23	2.08	19.70	0.11
Gifted	328.4	344.3	330.5	339.9	-2.2	18.2	-0.12

*According to WWC, baseline differences <.05 of a standard deviation satisfy baseline equivalence without adjustment, according to WWC. Differences <.25 of a standard deviation satisfy baseline equivalent with adjustment of difference-in-difference (2022).
++Subgroup does not satisfy baseline equivalence even with statistical adjustment.

Table A3

Comparison of changes in scores and percent meeting standards between consistent Zearn Math users and Non-users, across all students

	Treatment change in mean	Control change in mean	Difference	Pooled SD	Cohen's d
All Students					
Math scale score (SS)	14.71	0.15	14.56***	17.07	0.85
Math percent meeting standards	4.14%	-19.35%	23.49%***	0.49	0.48
Starting Achievement Level					
Below Standards SS	16.56	-4.37	20.94***	19.09	1.10
Meeting Standards SS	16.48	-0.01	16.49***	15.97	1.03
Above Standards SS	12.78	2.76	10.02***	14.21	0.70
Grade Level					
G4 SS	18.78	5.71	13.07***	14.60	0.90
G4 percent meeting standards	6.63%	-15.86%	22.49%***	0.40	0.56
G5 SS	10.58	-6.43	17.02***	15.65	1.09
G5 percent meeting standards	-0.27%	-24.40%	24.13%***	0.43	0.56
G6 SS	6.60	-8.03	14.63***	15.91	0.92
G6 percent meeting standards	5.00%	-21.67%	26.67%***	0.40	0.67

Table A4

Comparison of changes in scores and percent meeting standards between consistent Zearn Math users and Non-users, by subgroup

	Treatment change in mean	Control change in mean	Difference	Pooled SD	Cohen's d
Subgroup					
Female SS	14.64	-0.21	14.85***	15.70	0.95
Female percent meeting standards	4.87%	-18.85%	23.72%***	0.42	0.57
Male SS	14.78	0.47	14.31***	16.36	0.88
Male percent meeting standards	3.51%	-19.80%	23.31%***	0.41	0.57
Black and/or Latino SS	14.96	0.18	14.78***	15.49	0.95
Black and/or Latino percent meeting standards	8.12%	-19.56%	27.68%***	0.45	0.61
Economically disadvantaged SS	14.44	-3.79	18.23***	17.72	1.03
Economically disadvantaged percent meeting standards	6.59%	-23.6%	30.16%***	0.45	0.67
MLL SS	++	++	++	++	++
MLL percent meeting standards	++	++	++	++	++
Special education SS	11.81	-2.65	14.46***	17.85	0.81
Special education percent meeting standards	5.10%	-12.05%	17.15%**	0.40	0.43
Gifted SS	15.93	9.40	6.53***	14.67	0.45
Gifted percent meeting standards	-2.13%	-5.61%	3.48%	0.23	0.15

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

++Subgroup does not satisfy WWC standards for baseline equivalence even with statistical adjustment.

Appendix B

This study was designed to meet the What Works Clearinghouse (WWC) “Meets WWC Group Design Standards with Reservations'' rating and to meet an ESSA Tier 2 (Moderate) rating on the ESSA guidelines for evidence-based interventions. This Appendix provides more detail about the criteria for these designations and how this impact study meets those criteria.

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 seven 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 state’s standardized assessment. WWC considers standardized tests that are routinely administered in educational settings, like this, 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 assessments, as well as at least four of seven other student characteristics: school, gender, race/ethnicity, special education status, English learner status, free or reduced lunch status, and gifted 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, math assessment 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 in this study meet the criteria for baseline equivalence either without or with adjustment, with the exception of students who are MLLs (see Appendix B Table B1). Results for that subgroup are not reported in the results as they do not qualify as baseline equivalent even with statistical adjustment.

Table B1

Study qualification for WWC baseline equivalence standards	
All students	Meets
Grades	Meets
Achievement Levels	Meets
Female	Meets
Male	Meets
Black and/or Latino	Meets w/adjustment
Economically disadvantaged	Meets w/adjustment
MLL	Did not meet
Special education	Meets w/adjustment
Gifted	Meets w/adjustment

*Baseline differences <.05 of a standard deviation satisfy baseline equivalence without adjustment. Differences <.25 of a standard deviation satisfy baseline equivalent with statistical adjustment.
**See Appendix A Table A2 for baseline equivalence data.

WWC Essa Tier 2 designation requires a strong quasi-experimental research design that would qualify for Meets WWC Standards with Reservations. In addition, an ESSA Tier 2 rating requires a minimum of 350 students. This analysis has 1,111 students in each group (2,222 total), so the sample size exceeds 350. In addition, the study must have been conducted in more than one school. This study spans 53 treatment schools, with an additional 25 schools of the 77 total control schools.

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).