Efficacy Analysis of Zearn Math in Nebraska

September 2022

Alisa Szatrowski, Ph.D.
Zearn

Abstract

Analysis of consistent Zearn Math users and a comparable group of low- or non-users, across 6 districts in Nebraska, shows that students who used Zearn Math consistently had higher levels of academic growth than similar students with low- or no-usage. For this analysis, 784 students who completed an average of 3+ Zearn Math lessons per week, during the 2021-2022 school year, were matched to similar students who completed fewer than one lesson per week using the quasi-experimental method Coarsened Exact Match (CEM). Consistent users were matched to low- or non-users on pre-Nebraska Student Centered Assessment System (NSCAS) scores in English and Math, grade, and seven academic or demographic factors. In comparison to matched low- or non-users, consistent Zearn Math users gained an additional 25 scale score points on the NSCAS (effect size=0.40) and were more likely to increase their proficiency level. Results were consistent across historically marginalized subgroups.

Table of Contents

List of Figures

List of 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 across the state of Nebraska. 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 across the six districts with the highest Zearn Math usage in the state. Nebraska has 324,176 students, of whom 46% are free/reduced lunch-eligible, 7% are English learners, 16% are students with disabilities, 13% are students in gifted programs and 20% are Black and Latino students (NEP, 2021b).

In grades 4-6, 872 students consistently used Zearn Math. Consistent users are those students who completed three or more digital lessons per week, i.e., 90 or more digital lessons per year, and could be matched to assessment data from the 2020-2021 and 2021-2022 school years. Appendix A Table A2 contains a comparison of the sample composition and state student population.

This study was designed to meet the What Works Clearinghouse (WWC) Meets WWC Group Design Standards with Reservations and to meet an Every Student Succeeds Act (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.)

Matching Methodology

Quasi-experimental matching techniques were used to isolate the impact of Zearn Math on student achievement. Consistent Zearn Math users were matched with non-users on starting math and English Language Arts (ELA) achievement scores, along with seven 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 school year. The outcome under investigation was the average treatment effect 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 little to no Zearn Math usage, operationalized as an average of less than one digital lesson per week; fewer than 30 digital lessons per year.[1]

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 little to 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 potentially confounding factors.[2]

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.[3]

Using CEM, treatment students were put into matching strata with control students that were in the

[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 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 little or no usage. The implications of Zearn's approach are discussed further in the limitations section.

[2] That students who reach fidelity and those with little to no usage may have unmeasurable differences is not considered a confounding factor by WWC but what WWC terms, "imperfect overlap in the characteristic between the conditions'. WWC provides the example 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. They clarify 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).

[3] 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).

same grade and within five scale score points on math and ten points on ELA on the Nebraska Student-Centered Assessment System (NSCAS) spring 2021 assessment.[4] Then, within strata, treatment students were matched to control students with whom they shared at least four of seven other student characteristics: district, gender, race, ethnicity, special education status, English-learner status and free/reduced lunch eligibility.

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 (1981; Rosenbaum, 2020).[5]

If a treatment student had no match within their grade and score strata with whom they shared at least four characteristics, they were excluded from the treatment group. 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 Nebraska's 872 consistent Zearn Math users, all but 88 were matched. Treatment and control populations differed by an average of 1.38 demographic factors, 2.17 points in math starting score and 4.72 points in ELA starting score. Mean pretest math scores between treatment and control students differed by .24 scale score points on the NSCAS. 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).

The 88 consistent Zearn Math users excluded from the study, due to lack of match, did not concentrate in any demographic category that would bias the sample on observable characteristics (See Appendix A Table A1 for a breakdown of sample demographics).

[4] WWC guidelines require that baseline equivalence be established using "a pretest in the same domain as the outcome" (What Works Clearinghouse, 2022). In this study both pre-scores and post-intervention outcome scores are measured by the same assessment, the NSCAS.

[5] 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.

Analysis

Once consistent Zearn Math users were matched to a similar group of low- or 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 proficiency and demographic factors.

Academic growth was measured as the change in NSCAS scores between the spring 2021 and spring 2022 assessment administration. NSCAS has three achievement levels: Developing, On Track and CCR Benchmark. Students scoring On Track and above are considered proficient. Outcomes are reported in terms of change in scale score, change in achievement level and change in percent proficient.

Difference in means t-tests were run on the average academic gains of the treatment group vs. the average academic gains of the control group to determine if the impact of treatment 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:

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

On average, consistent Zearn Math users in Nebraska gained 42 scale score points whereas matched low- or non-users gained 17 points between spring 2021 and spring 2022, a difference of 25 scale score points (effect size=0.40). Gains were highest among consistent users who started the year below proficient (“developing”). These students gained 51.7 scale score points while low- or non-users gained only 20.8, a difference of 30.8 points (effect size=0.66) (See Results Table 1). The outsized impact of Zearn Math usage among students starting below proficient has been previously reported by Zearn (2022a; 2022b) (See Appendix A Table A4 for findings from the difference in means analysis).

RESULTS TABLE 1

Growth in Scale Score for Consistent Zearn Users vs. Low- or Non-Users

Nebraska statewide growth in scale scores for consistent Zearn Math users *(treatment)* vs. Low- or Non-users *(control)*, by starting achievement

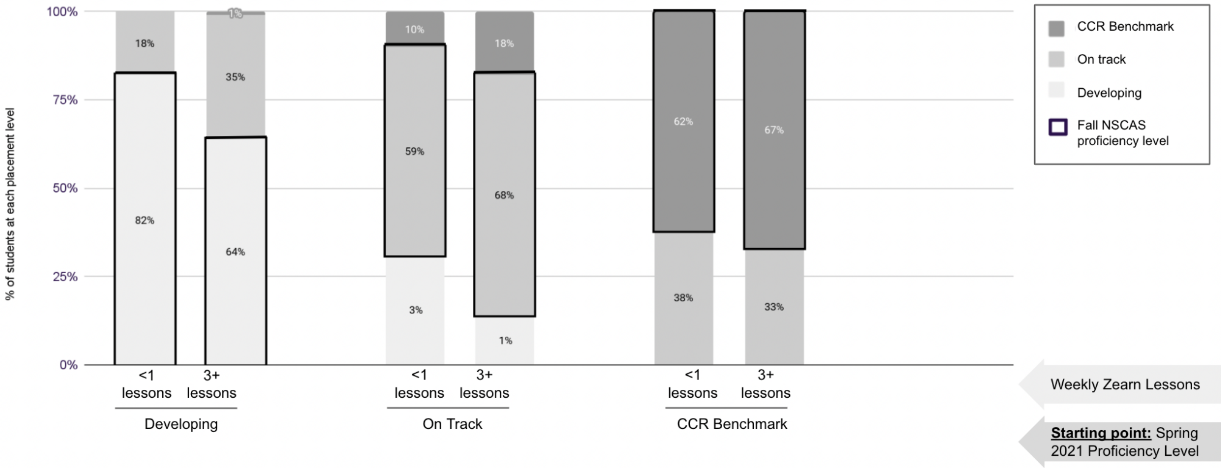
	Developing	On Track	CCR Benchmark
Treatment growth in percentile points	51.7	35.2	17.8
Control growth in percentile points	20.8	13.5	13.5
Growth difference in percentile points	30.8	21.7	4.3

In addition to absolute growth, mobility models compared the change in achievement level for treatment and control students based on starting achievement level. Across all achievement levels, consistent Zearn Math users maintained or increased their achievement levels at higher rates than non- or low-users.

Notably, among Nebraska students who started below proficient, two times as many students who used Zearn Math consistently improved their achievement level, relative to students with little to no Zearn Math usage. Results Table 2 illustrates the mobility between achievement levels, for consistent users and low to non-users.

Figure 1

NSCAS spring 2022 Proficiency Level, by Spring 2021 Proficiency and Zearn Usage



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 low- or non-users were allowed to mismatch on up to three demographic characteristics, subgroups did not always align on starting proficiency. Therefore differences in proficiency by demographic subgroup were reported as the change in percent proficient, i.e., difference-in-difference,[6] rather than as raw scores (See Descriptive Table 3 for a breakdown of starting and ending means, by subgroup).

Across all subgroups, consistent Zearn Math users saw gains in percent proficient while non- or low-users saw drops in percent proficient in the 2021-2022 school year. On average, consistent users across Nebraska saw a 9% increase in percent proficient, while low- and non-users saw a 7% decrease. These rates of increase were consistent or larger across subgroups of students including: female students, Black and Latino students and students with free/reduced lunch eligibility (See Results Tables 4 and 5 for more details).[7]

RESULTS TABLE 2

Percent of Students Meeting Proficiency

	Consistent Users *(Treatment)*	Low- or Non-Users *(Control)*
All Students	9%	-7%
Female	9%	-6%
Male	8%	-8%
Black & Latino	11%	-7%
Special education	++	++
Economic disadvantage	11%	-5%
English Language Learners	++	++

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

[6] All students, male students, and Black and/or Latino students had baseline differences <.05 of a standard deviation which satisfies baseline equivalence without adjustment, according to WWC. Female students and students with economic disadvantage had differences <.25 of a standard deviation, satisfying baseline equivalency with a difference-in-difference adjustment (2022). Students in special education and ELL did not satisfy baseline equivalence in this analysis. (See Appendix A Table A3 for full details on baseline equivalence.)

[7] For each subgroup in treatment and control, percent proficient in spring 2021 was subtracted from percent proficient in spring 2022. This change is depicted in Results Table 3. If the percent proficient within a subgroup was the same in the two periods, the change listed in Results Table 3 would be 0.

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 proficient saw even larger gains than the average student. The finding that Zearn Math usage impacts all students positively, but is associated with even more growth among those starting below proficiency, further substantiates findings from efficacy analyses of Zearn Math's impact in other districts (Zearn 2022a & 2022b).

By matching students closely on starting 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 correlational analyses. 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, some limitations are present. While quasi-experimental methods allow researchers to control for observed confounders, a possibility exists that unobserved confounders mediate the relationship between Zearn Math usage and academic performance. Even with robust quasi-experimental methods, accuracy of estimates is limited by the ability to model all variables relevant to selection into treatment and control.

This analysis examines the impact of fidelity usage of Zearn Math rather than utilizing an intention-to-treat analytic framework that would define the treatment group as all students to whom Zearn Math was available (McCoy, 2017). The focus on fidelity usage better aligned with the interests of this partner, for whom the results can help to encourage more universal fidelity usage of Zearn Math. However, utilizing fidelity usage as the benchmark for treatment means that estimates may be biased as this usage represents the best version of implementation which may exceed "typical use".

Finally, this study was conducted in six districts across one state. It is possible that the impact of Zearn Math in other locations, or across a larger number of students within Nebraska, might show a different effect size, whether larger or smaller. It is also possible that there are features specific to Nebraska public schools that facilitate large gains with Zearn Math usage that may not be present in other districts and states. The geographic specificity of this study may limit the generalizability to a more nationally representative population.

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.

Works Cited

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

NEP. (2021a). Nebraska Public Schools Data Download. *Nebraska Dept of Education*. Retrieved August 17, 2022, from https://nep.education.ne.gov/Links

NEP. (2021b). Nebraska Public Schools State Snapshot. *Nebraska Dept of Education*. Retrieved August 17, 2022, from https://nep.education.ne.gov/statedata.html#00-0000-000/snapshot/20202021

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

Thum, Y. M., & Kuhfeld, M. (2020a). NWEA 2020 MAP Growth Achievement Status and Growth Norms for Students and Schools. *NWEA Research Report*. Portland, OR: NWEA

Thum, Y. M., & Kuhfeld, M. (2020b). NWEA 2020 MAP Growth Achievement Status and Growth Norms Tables for Students and Schools. *NWEA Research Report*. Portland, OR: NWEA.

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

Appendix A

Table A1

Breakdown of sample matching characteristics		
	Treatment	**Control**
Total N's	784	784
Pre-scores (Spring 2020-2021 assessment scores)		
Math scale score	1212.8	1212.5
ELA scale score	2498.3	2498.3
Grades		
Grade 4	401	401
Grade 5	278	278
Grade 6	105	105
Demographic & academic subgroups		
Female	367	362
Male	417	422
Black and/or Latino	309	322
Students with disabilities	57	46
Free/reduced lunch eligible	411	406
English learners	76	42

Table A2

Comparison of sample and statewide school population			
	Treatment	**Control**	**State**
Demographic & academic subgroups			
Black and/or Latino	39%	41%	20%
Students with disabilities	7%	6%	16%
Free/reduced lunch eligible	52%	52%	46%
English learners	10%	5%	7%

Table A3

Spring 2021 and Spring 2022 performance means by subgroup

	Treatment Spring 2021	Treatment Spring 2022	Control Spring 2021	Control Spring 2022	Fall mean difference	Pooled SD	Difference in SDs*
All Students							
Math scale score	1212.8	1254.3	1212.6	1229.4	0.24	58.73	0.00
Starting proficiency							
Below Average	1163.5	1215.1	1163.8	1184.6	-0.33	38.65	-0.01
Average	1243.8	1279.0	1243.8	1257.3	0.03	30.31	0.00
Above Average	1325.8	1343.6	1324.6	1338.1	1.17	33.20	0.04
Grades							
Grade 4	1200.3	1250.3	1200.1	1224.1	0.22	57.18	0.00
Grade 5	1223.3	1259.0	1223.0	1235.3	0.27	59.86	0.00
Grade 6	1232.8	1257.1	1232.5	1233.7	0.33	50.21	0.01
Demographic & academic subgroups							
Female	1213.0	1253.2	1209.2	1228.2	3.81	56.44	0.07
Male	1212.6	1255.3	1215.4	1230.3	-2.82	60.64	-0.05
Black and/or Latino students	1192.1	1236.1	1188.4	1202.7	3.67	55.97	0.07
Students with disabilities	1186.6	1221.4	1210.4	1226.5	-23.81	63.89	-0.37
Free/reduced lunch eligible	1193.5	1233.6	1188.6	1204.3	4.84	55.07	0.09
English learners	1149.5	1206.1	1130.9	1146.0	18.56	41.67	0.45

*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).

Table A4

Comparison of changes in scores and proficiency between consistent Zearn users and low- or non-users

	Treatment change in mean	Control change in mean	Difference	Pooled SD	Cohen's d
All Students					
Math scale score (SS)	41.527	16.797	24.730***	61.675	0.401
Math percent proficient	8.79%	-7.26%	16.05%***	0.560	0.287
Starting proficiency					
Developing SS	51.663	20.830	30.833***	46.471	0.664
On track SS	35.210	13.526	21.684***	47.964	0.452
CCR benchmark SS	17.767	13.467	4.301	42.153	0.102
Grade					
Grade 4 SS	50.06484	24.06234	26.003***	47.6	0.546
Grade 4 percent proficient	7.73%	-10.47%	18.20%***	0.4906735	0.371
Grade 5 SS	35.74194	12.20072	23.541***	45.70826	0.515
Grade 5 percent proficient	11.47%	-1.79%	13.26%***	0.4557503	0.291
Grade 6 SS	24.29524	1.266667	23.029***	45.27544	0.509
Grade 6 percent proficient	5.71%	-9.52%	15.24%*	0.5172795	0.295

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

Table A5

Comparison of changes in scores and proficiency between consistent Zearn users and low- or non-users, by subgroup

	Treatment change in mean	Control change in mean	Difference	Pooled SD	Cohen's d
Subgroup					
Female SS	40.18529	19.02486	21.160***	45.768	0.462
Female percent proficient	9.26%	-6.08%	15.34%***	0.466	0.330
Male SS	42.70574	14.89125	27.814***	48.833	0.570
Male percent proficient	8.37%	-8.27%	16.65%***	0.498	0.335
Black and/or Latino SS	43.99676	14.24768	29.749***	45.72033	0.651
Black and/or Latino percent proficient	11.33%	-6.50%	17.80%***	0.4837942	0.369
Free/reduced lunch eligible SS	40.12621	15.64128	24.485***	47.34244	0.517
Free/reduced lunch eligible percent proficient	11.65%	-5.16%	16.81%***	0.5068117	0.332
English learner SS	++	++	++	++	++
English learner percent proficient	++	++	++	++	++
Special education SS	++	++	++	++	++
Special education percent proficient	++	++	++	++	++

* 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 low- or 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 of growth in math is the i-Ready Diagnostic. WWC considers standardized tests that are routinely administered in educational settings, like i-Ready Diagnostic, to meet these standards.

The next criteria is the elimination of confounding factors (What Works Clearinghouse, 2022). By matching fidelity users to low- or non-users within five scale score points on their math and ELA NSCAS pre-scores, as well as at least four of seven other student characteristics: district, gender, race, ethnicity, special education status, English-learner status and free/reduced lunch eligibility, 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 little to no usage. For example, it is possible that an unmeasured characteristic allows fidelity users to reach higher usage than would be possible for low- or 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, i-Ready Diagnostic 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 and fixed effects are both acceptable statistical adjustments (2022). All groups in this study meet the criteria for baseline equivalence either without or with adjustment, with the exception of students in special education in the two year impact (see Appendix B Table B1). Results for that subgroup are not reported in the two-year impact results as they do not qualify as baseline equivalent even with statistical adjustment.

Table B1

Study qualification for WWC baseline equivalence standards, by analysis and subgroup*

All students	Meets
Grades	Meets
Starting proficiency	Meets
Female	Meets w/adjustment
Male	Meets
Black and/or Latino	Meets
FRL	Meets w/adjustment
MLL	Does not meet
Special education	Does not meet

*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 Tables A3 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. The sample size for this study includes 784 treatment students and 784 control students (see Appendix B, Table B2). In addition, the study must have been conducted in more than one school or district. This study spans 6 districts and over 100 schools.

Table B2

Sample size of Nebraska analysis

Treatment sample	Control sample	Total sample
784	784	1568

*In order to qualify for ESSA Tier 2, a study must include at least 350 participants.

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 (2022a, 2022b).