Effectiveness of Learning A-Z's *Raz-Plus* in Milwaukee Public Schools

October 2019

Learning A-Z contracted with Empirical Education to study the effectiveness of *Raz-Plus* in Milwaukee Public Schools during the 2016-17 school year. *Raz-Plus* is a literacy program that includes leveled books, skills practice, and digital activities and assessments.

This study focused on 3rd, 4th, and 5th grade students and examined the impact of *Raz-Plus* usage on student performance, as measured by the STAR Reading (STAR) assessment. We investigated the following questions.

(1) Do students in classes of teachers who actively use *Raz-Plus* perform better on the STAR test than comparable students in classes whose teachers did not use *Raz-Plus?*

(2) Is the impact of *Raz-Plus* different for students with different characteristics?

(3) Are differences in *Raz-Plus* usage associated with differences in student performance?

Results

(1) Do students in classes of teachers who actively use *Raz-Plus* perform better on the STAR test than comparable students whose teachers did not use *Raz-Plus*?

Students in classes of teachers who actively used Raz-Plus performed better on the STAR test than comparison students in classes where teachers did not use Raz-Plus.

We find that reading scores for students in classes of teachers who actively used *Raz-Plus* are better than for comparison students. The result corresponds to a 3-percentile point gain on the STAR test, adjusting for differences in student demographics and pretest between *Raz-Plus* and comparison students.

As shown in Figure 1 (and Table 2), the effect size of the overall impact of *Raz-Plus* on the standardized STAR score is 0.083, and we have strong confidence in this result ($p < .01$).

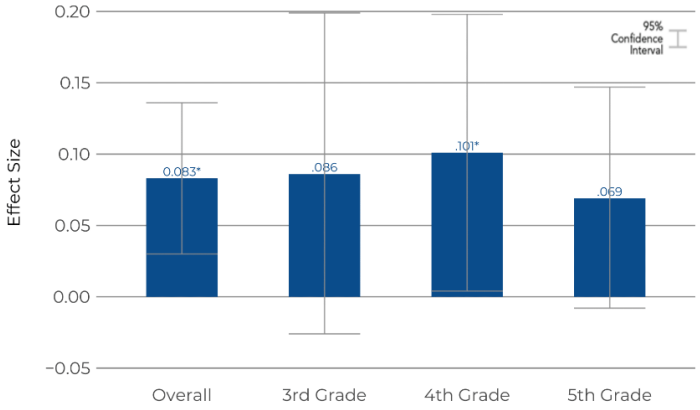

FIGURE 1. AVERAGE EFFECTS BY GRADE LEVEL

Note. 95% confidence intervals convey that we have strong confidence of the results falling within that area.

(2) Is the impact of *Raz-Plus* different for students with different characteristics?

There is a positive impact of Raz-Plus for several student subgroups, including 4th grade students, non-white students, economically disadvantaged students, and English Language Learners.

As shown in Figure 1, we found a positive impact of *Raz-Plus* for students in 4th grade: 0.101 (4 percentiles of test score distribution; $p < .05$). The impacts in grades 3 and 5 are smaller and we have only moderate confidence in their effect: 0.086 (3-percentile point gain; $p = .13$) for 3rd grade, and 0.069 (3-percentile point gain; $p = .08$) for 5th grade.

We found significant impacts on several other student subgroups. There was a positive effect of *Raz-Plus* for Asian, African American, and Hispanic students with effect sizes of 0.235, 0.077, and 0.101 respectively, all with $p < .05$.

The impact of *Raz-Plus* was also positive for economically disadvantaged students and English Language Learners (ELLs), with effect sizes of 0.086 ($p < .01$) and 0.132 ($p = .06$) respectively. Subgroup impacts are shown in detail in Table 3.

(3) Are differences in *Raz-Plus* usage associated with differences in student performance?

There is a positive association between several Raz-Plus usage metrics and student performance.

We estimated the effects of selected usage metrics on student STAR test scores for all *Raz-Plus* users included in the impact study. We found that outcomes on the STAR are significantly positively affected by the number of quizzes assigned in *Raz-Plus* (0.007 percentile improvement per unit, $p < .01$). Quizzes assigned refers to the number of electronic quizzes about an electronic text which teachers can digitally assign to students for them to complete independently online.

To put these results in context, the result of bi-monthly quizzes over the school year (21 quizzes assigned) would increase STAR scores by 1 percentile point. While several other usage metrics were correlated with student outcomes with high levels of statistical significance, the estimates of the amount of usage necessary to achieve measurable differences in outcomes are higher than actual student usage; all of these results are shown in Table 4 and detailed descriptions of the activities in Table 5.

It should be emphasized that since students and/or teachers have freedom to choose the level of usage and the type of activities, none of these behavioral relationships can be considered causal. It is still a possibility that usage is affected by unmeasured student abilities or interest in using computer-based learning tools.

LEVELS OF CONFIDENCE IN OUR RESULTS

Results are reported based on statistical calculations that give a measure of confidence expressed as a probability or *p* value. A low *p* value indicates a low probability that we would detect a difference like the one found in the study if no difference actually existed. A *p* value less than .05 gives us strong confidence in the result (a level conventionally called statistically significant). A *p* value between .05 and .15 gives us moderate confidence, while a p value between .15 and .20 gives us limited confidence. A *p* value greater than .20 gives no confidence.

Study Description

Study Design

The study compared achievement for students in two groups: one group where teachers of students in grades 3, 4, and 5 actively implemented *Raz-Plus*, and the other group who did not, adjusting for the differences in student characteristics. Program developers identified the importance of teacher logins for high-quality implementation: the treatment group was restricted to students in classes where teachers had at least two logins over the school year. For this analysis, we also restricted the sample by excluding schools where a large percentage of students could not be matched between *Raz-Plus* records and Milwaukee Public Schools records, classes with fewer than eight students, and students with missing pre-test values.

Participants

The study took place in Milwaukee Public Schools during the 2016-17 school year. The district provided student data including unique student ID; school, teacher, and course data; student demographics; and STAR pretest and outcome data. These data were combined with data from the 2016-2017 *Learning A-Z* logs that included the total time students spent on the program, the number of books students listened to and/or read independently, the number of resources (books, quizzes) assigned to students by their teachers, and the number of times teachers logged in to the portal to view resources, check on student progress, or assign resources to students.

Conclusion

We found a positive impact of *Raz-Plus* on the STAR test, and these impacts were significant for several student subgroups, including non-white students, economically disadvantaged students, and ELL students. Moreover, several *Raz-Plus* usage metrics were found to have positive associations with student outcomes on the STAR assessment.

CAUTIONS FOR INTERPRETING THESE RESULTS

Results shown in the figures and tables are not actual differences in test outcomes but estimates that adjust for the differences between users of *RAZ-PLUS* and a comparison group, and they should be interpreted as the hypothetical improvement in outcome for the average comparison student if they were in an *RAZ-PLUS* classroom. The actual outcomes for actual *RAZ-PLUS* students may vary depending on their characteristics.

Results reported as no difference do not imply that no real differences exist, but that a large study is needed to estimate them accurately.

This case study was conducted on behalf of Milwaukee Public Schools with the technical assistance of Empirical Education. In conducting or supporting the agency's conduct of the study, Empirical does not intend to generate evidence valid beyond the agency in which the case study was conducted.

Technical Details

Data Preparation

The district provided student data for the 2016-2017 school year for all students in grades 3 through 5.

Learning A-Z provided student log data for the 2016-2017 school year from the *Raz-Plus* system. All log data was for the time period beginning July 1, 2016 and ending June 31, 2017. *Learning A-Z* provided log data for all students in the district, which was then merged with demographic and achievement data from Milwaukee Public Schools.

Analytic Sample

The analytic sample for the comparison study, shown in Table 1, consisted of 3rd, 4th, and 5th grade students with both fall and spring STAR test scores. *Raz-Plus* classes were matched one-to-one to comparison classes using the "nearest-neighbor" method, including the following covariates at the class level: gender, ethnicity, ELL status, eligibility for free or reduced-price lunch (FRPL), disability, and STAR pretest score. In the group of *Raz-Plus* users there were 3,461 students, and 3,622 in the comparison group. The sample for the usage analysis was consistent with the impact study. All covariates had differences of less than .25 standard deviations between the comparison and *Raz-Plus* groups, as shown in Table 6.

TABLE 1. ANALYTIC SAMPLE

	Raz-Plus		Comparison	
	Classes	Students	Classes	Students
3rd	41	728	41	747
4th	54	1,075	54	1,047
5th	82	1,658	82	1,828

Analysis

We used a hierarchical linear mixed effects regression model, which accounts for the clustering of students within classes and within schools, adjusting for student demographics and pretest scores to compare performance for *Raz-Plus* and comparison students. Table 2 below displays the results. STAR scale scores were standardized across grades.

The usage analysis was also performed using hierarchical linear mixed effects regression with the STAR test as the outcome variable, and student characteristics, pretest, and usage metrics as covariates.

The results of the usage analysis were obtained using an iterative procedure aimed at establishing the strongest association between student-level usage metrics and STAR outcomes. The procedure starts from estimating a model with all covariates and usage metrics. Quality is assessed using Akaike Information Criteria (AIC), and the least significant usage metrics are removed from the model, leaving only those with the highest predictive power. This does not imply a causal link, because teachers may choose to assign different tasks to different students depending on their ability.

TABLE 2. DETAILED RESULTS

Value	All users	3rd grade	4th grade	5th grade
***Raz-Plus* effect size, standardized STAR score**	0.083	0.086	0.101	0.069
***p* value**	<.01	.13	.04	.08
Percentile gain	3	3	4	3

TABLE 3. RESULTS OF SUBGROUP ANALYSIS

Subgroup	Effect size	*p* value
African American	0.077	.02
Hispanic	0.101	.03
Asian/Pacific Islander	0.235	<.01
White	0.045	.39
Economically disadvantaged	0.086	<.01
Special education	-0.001	1.0
English Language Learner	0.132	.06

TABLE 4. ASSOCIATIONS BETWEEN RAZ-PLUS USAGE AND STAR POSTTEST SCORES

Usage metric	Estimate (per unit)	Std error	*p*	Mean	Max	SD	Usage required for 1 percentile point difference
Total Minutes	0.0001	0.00000	<.01	384.3	5543	573.6	167,227
Level-Up Reads	0.0036	0.0017	.03	16.7	398	30.9	213
Reading Room Listens	-0.0015	0.0004	<.01	26.1	741	50	841
Reading Room Quizzes	0.0011	0.0004	<.01	15.9	1023	43.5	1,032
My Assignment Listens	-0.0114	0.0047	.02	0.8	47	3.5	8
My Assignment Quizzes	0.0071	0.0026	<.01	1	185	6	21
Total Resources Assigned	-0.0011	0.0004	.01	63	1612	122.4	2,725

TABLE 5. DESCRIPTION OF USAGE METRICS

Usage Metric	Description
Total Minutes	Amount of time the student spent actively completing *Raz-Plus* activities online
Level-Up Reads	Number of books read by the student at his or her independent reading level (as defined by the teacher)
Reading Room Listens	Number of audio books heard by the student in a free-reading area where students choose from a wide variety of books at different levels
Reading Room Quizzes	Number of quizzes completed by the student in in a free-reading area where students choose from a wide variety of books at different levels
My Assignment Listens	Number of audio books heard by the student that were specifically assigned by the teacher
My Assignment Quizzes	Number of quizzes completed by the student that were specifically assigned by the teacher
Total Resources Assigned	Total number of books, audio books, quizzes, and passages assigned to the student by the teacher

Note. All usage metrics refer to the 2016-17 school year.

TABLE 6. BASELINE EQUIVALENCE

	Pretest	% Male	% FRPL	% ELL	% Disability	% White	% African American	% Hispanic	% Asian
3rd grade									
Comparison mean	272.0	51.4	86.5	2.1	15.4	12	63.9	17.1	3.1
***Raz-Plus* mean**	266.6	50.6	84.5	2.9	17.6	14.6	63.7	16.5	0.8
Pooled SD	80.4	0.500	0.352	0.156	0.371	0.340	0.481	0.374	0.017
Diff % SD	0.068	0.017	0.057	0.048	0.059	0.074	0.002	0.017	0.162
4th grade									
Comparison mean	383.2	52.4	83.1	6.9	16.5	11.1	64.7	16.3	4.2
***Raz-Plus* mean**	381.2	52.4	84.5	6.4	17.9	11.0	64.0	16.5	5.0
Pooled SD	86.9	0.500	0.369	0.249	0.378	0.313	0.479	0.370	0.210
Diff % SD	0.023	0.001	0.037	0.018	0.035	0.003	0.014	0.004	0.039
5th grade									
Comparison mean	488.3	50.8	82.1	11.8	15.6	15.5	51.1	27.3	2.8
***Raz-Plus* mean**	485.4	47.4	85.8	11.9	15.4	11.4	50.7	29.4	2.8
Pooled SD	90.0	0.500	0.368	0.323	0.362	0.343	0.500	0.451	0.196
Diff % SD	0.032	0.067	0.099	0.002	0.004	0.121	0.009	0.047	0.129

Reference for this report: Empirical Education Inc. (2019). *Effectiveness of Learning A-Z's Raz-Plus in Milwaukee Public Schools.* (Empirical Education Rep. No. Empirical_LAZ-6057-DDS-R-2019-O.1). San Mateo, CA: Empirical Education Inc. Retrievable from https://www.empiricaleducation.com/learning-a-z/