Impact of Goalbook Toolkit

February 2024

This report leverages product usage data and school district data to explore the impact of Goalbook Toolkit usage on student outcomes.

Empirical Education®

Principal Findings

Empirical Education conducted a study to evaluate the impact of Goalbook Toolkit on special education students' reading achievement in a large suburban school district in the northeastern United States during the 2021–22 school year.

- Correlational analysis found evidence of a positive association between the frequency of Goalbook Toolkit use and student outcomes on the state assessment in reading among students of Goalbook Toolkit teachers. Weekly usage of Goalbook Toolkit during the school year is predicted to have a positive impact of a 9-percentile gain on the reading assessment.
- The association between Goalbook Toolkit use and student outcomes was the strongest for students eligible for free and reduced lunch: a 15-percentile gain on the state reading test. There were no significant differences in the Goalbook Toolkit effect among other student groups.
- A comparison group study found no significant differences in outcomes between students whose teachers used Goalbook Toolkit and students whose teachers did not use Goalbook Toolkit. However, we did find a significant positive impact of Goalbook Toolkit use in the sample of students of teachers who used it at least weekly when compared to those who used it less frequently.

BACKGROUND

This study of usage and effectiveness of Goalbook Toolkit is based on student data from a large suburban school district in the northeastern US and teacher-level application usage data from Goalbook Toolkit from the 2021–22 school year. Goalbook Toolkit enables teachers to personalize instructional supports for specialized student populations. In the district, special education teachers or resource teachers that provide direct services to students through pull-out, push-in, or co-teaching have access to Goalbook Toolkit. They use the application to support IEP development, progress monitoring, and specially-designed instruction in general. The main research question is whether use of Goalbook Toolkit has a positive impact on student achievement in reading.

STUDY DESIGN

This report presents the findings from two studies for an in-depth exploration of Goalbook Toolkit's effect on student outcomes. The comparison group and correlational analyses use different samples and provide different tiers of evidence of impact.

The comparison group portion of the study compares the outcomes of students in treatment and control groups. The student-level treatment condition was defined in accordance with early studies of Goalbook effectiveness and includes special education students whose teachers used Goalbook Toolkit more than a single day in the specified school year. The comparison group included special education students whose teachers did not use Goalbook Toolkit. Students who had a mix of user and non-user teachers were excluded from the analysis.

The correlational component of the study aimed to establish statistical associations between product usage metrics and student outcomes, among all students with at least one teacher who used Goalbook Toolkit. The correlational study uses a larger sample than the quasi-experimental analysis. It focuses entirely on product users and the differences in outcomes that can be attributed to the differences in usage, making appropriate adjustments for differences in users' individual and class characteristics and pretest scores.

The results of such an analysis are used to predict potential outcomes at some possible level of usage. In this study, that usage level is consistent weekly use throughout the school year (36 days of usage). We compare those potential outcomes to imputed outcomes for students of non-user teachers (zero days of usage). Positive results of this sort should be viewed as showing potential (promise of effectiveness) rather than proving effectiveness, because they are calculated for a hypothetical optimal level of usage that exceeds actual average usage in the sample.

In both the comparison group and correlational studies, there is a possibility of selection bias—i.e., the possibility that more effective teachers choose to use the application more actively or that more intensive usage is required for students who are in greater need of instructional support. If this is the case, the estimated effect may in fact reflect the differences among users. To address this issue, we establish baseline equivalence on the pretest between the treatment and comparison groups in the comparison group study. For the correlational study, we establish the absence of correlation between the pretest and usage metrics. We can therefore rule out the selection bias on observable characteristics within the available data.

DATA

Data collected for this study consisted of 4,378 individual student records. Records contained student demographics, school and teacher identifiers, pretest and posttest reading scores (the fall district benchmark and spring state assessment, respectively) for all special education students in the district, multiple metrics of Goalbook Toolkit usage by

teachers (user events such as page views, mouse clicks, etc.), and class rosters linking students to the teachers.

The set of teachers of interest for this study included 436 ELA and social studies teachers working exclusively with special education students and designated as resource teachers. They are the primary users of Goalbook Toolkit: one half of all designated resource teachers in the dataset used Goalbook Toolkit as opposed to 5% of other teachers that served special education students. Students assigned to at least one resource teacher were included in the analytic sample. For each student, Goalbook Toolkit usage metrics were averaged across all their resource teachers, as well as aggregated to obtain average usage days and total user events per student. Based on the available data, we see that resource teachers are typically present in middle schools in the district, but only some elementary schools have them. As a result, the analytic sample includes only middle school students. After removing records with missing pretest and/or posttest scores, the final analytic sample included 1,222 students in grades 6 through 8.

Parameters of the final analytic sample are presented in Tables 1-2.

TABLE 1. SAMPLE SIZES

Category	Number
Schools	36
Teachers	150
Students	1,222

TABLE 2. CHARACTERISTICS OF ANALYTIC SAMPLE

Category	% Total
FRPL	68.1
ELL	28.6
Black	59.8
Hispanic	35.8
White	2.5
GT	0.7

Note. FRPL stands for students who are eligible for free or reduced-price lunch. ELL stands for students classified as English language learners. GT stands for student enrolled in the gifted and talented program.

Table 3 shows the distribution of students, by grade, in the sample and average usage statistics for their teachers. The third column (Taught by Goalbook users, %) gives the percentage of students for whom all of their teachers were Goalbook Toolkit users (i.e. used for more than one day in the 2021–22 school year). The numbers in the fifth column, the average number of events per student, suggest that the intensity of Goalbook Toolkit usage tends to be lower among teachers of 8th grade students.

TABLE 3. GOALBOOK TOOLKIT USAGE BY GRADE

Grade	Students	Taught by Goalbook users, %	Average usage days	Events per student
Grade 6	394	56.0	9	17
Grade 7	582	58.6	10	14
Grade 8	576	43.9	9	10

ANALYSIS

The analysis was performed using a hierarchical linear regression model, whereby the product effect was estimated adjusting for student characteristics and pretest, and taking into account the clustering of students in schools. The Spring 2022 scores on the state reading assessment were the outcome variable in all analyses. In the comparison group study, Goalbook Toolkit usage was modeled by a single binary variable (treatment indicator). In the correlational analyses, Goalbook Toolkit usage was represented either by average usage days for the students' teachers or by total events per student (also averaged over the students' teachers).

The comparison group analysis yields an estimate of the average difference in student test outcomes. The correlational models produce estimates of the association between student test score gains and a 1-unit increase in Goalbook Toolkit usage (one active day or one event). Unit-effect estimates from the correlational analyses are used to project the differences in outcomes between non-users (zero active days) and users with a reasonable frequency of usage throughout the year. As described in Appendix A, we conducted exploratory analyses that suggested that 36 days per year, or weekly usage, is close to the minimal level that can have a substantial effect on student outcomes. The results are presented as percentile gains for a hypothetical student who would score at the 50th percentile on the test (among all students in the sample) if their teachers were not using Goalbook Toolkit.

Effects for student groups were estimated by including interaction terms in the model, allowing us to identify differences in the association between Goalbook Toolkit usage and student characteristics. Student group effects estimated from correlational analyses relate to the potential differences in outcomes between two 'average students' who only differ in one characteristic (e.g. if they are designated as an English language learner), but are otherwise "identical". Actual differences between two complementary groups (e.g. designated English language learners and non-English language learners) can be affected by characteristics other than their English language learner status.

RESULTS

Correlational Analysis

Correlational analysis yielded evidence that usage of Goalbook Toolkit, as measured by active days, is positively associated with student outcomes. The effect size for one unit of usage (one active day) is estimated to be 0.006, which translates into an effect size of .24 for weekly usage. This is equivalent to a 9-percentile gain for an average student in the sample. The level of confidence we have in these estimates is high (p = 0.02).

Moreover, the correlational analysis provides strong evidence that Goalbook Toolkit usage is positively associated with student outcomes across all student groups. However, a differential association between usage and outcomes is established only in one instance: the association is greater for students who are eligible for free or reduced-price lunch compared to those who are not eligible. There were no statistically significant differences in the association between Goalbook Toolkit usage and student outcomes, according to the students' race/ethnicity. These results may be due to the small size of most student groups. Significance of the differences between complementary student groups is indicated in the right column in Table 4. Detailed results are presented in Appendix B.

TABLE 4. STUDENT RESULTS OVERALL AND BY GROUP

Category	Predicted effect of weekly use (percentile)	Significant differential
All	9	
Female	16	No
Male	7	No
ELL	13	No
Non-ELL	9	No
FRPL	15	Yes
Non-FRPL	0	Yes
GT	0	No
Non-GT	10	No

Note. FRPL stands for students who are eligible for free or reduced-price lunch. ELL stands for students classified as English language learners. GT stands for student enrolled in the gifted and talented program.

Additional analyses of the association between various event-based usage metrics and student outcomes showed that the total Goalbook Toolkit user events (clicks, view, etc.) has a strong positive effect. Two specific event metrics were also positively associated with student outcomes: Clicked Develop Present Levels and Viewed Anchor Page. The model with specific event metrics was estimated iteratively with the elimination of the least

significant terms, until only a few significant metrics remained in the final model. The remaining metrics are shown in Table 5.

TABLE 5. ASSOCIATION OF EVENT METRICS WITH STUDENT OUTCOMES

Metric	Estimate	*p* value
Total events, log	0.116	0.03
Viewed Anchor Page	0.139	0.03
Clicked Develop Present Levels	0.128	0.13

These results should be interpreted with caution, because it is impossible to single out the effect of one type of event from the others given the strong correlations among all event metrics. In addition, each specific event metric was normalized using Box-Cox transformation so that their scales are not comparable to that of total events. This was included in the model as a decimal logarithm.

It is noteworthy that the predictive power of the model containing only the two specific event metrics—Clicked Develop Present Levels and Viewed Anchor Page—is almost the same (R^2 = 0.17) as for the model including the sum of all events. This suggests that these two metrics are key characteristics of Goalbook Toolkit user behavior.

Comparison Group Analysis

No statistically significant differences between treatment and comparison students were identified in the comparison group study (*p* =.37). This result contrasts the strong positive result from the correlational analysis. One possible explanation of this may be that the adopted threshold of 2 days of usage per year is too low to have an impact on student outcomes. We explored alternative approaches to defining users and found that the greatest difference in outcomes is observed between the students of teachers who used it at least weekly and those who used it less frequently. Details of these exploratory analyses are presented in the Appendix A.

CONCLUSION

Results of this study present strong evidence of promise that Goalbook Toolkit can improve special education student outcomes across student groups and middle school grade levels. Generalizability of these results is somewhat limited by the demographics of participating students, most of which belong to one racial/ethnic group. Whereas the study sample is large enough to obtain statistically significant average effects, it lacks adequate power for

smaller student groups. In the interpretation of the results, it is important to remember that this is a non-experimental study, with no pre-defined treatment and control groups, and that the reported usage effects are projections based on correlational results.

Appendix A. Comparison Group Results Using Alternative Designs

For a product like Goalbook Toolkit that is meant to be used repeatedly throughout the year, two active days per year is the absolute minimum that can qualify a teacher as a user. However, it is possible that a noticeable impact on student outcomes can be achieved only when user activities are more frequent. To test this hypothesis, we performed an exploratory iterative comparison groups analysis whereby the usage threshold for the inclusion of students in the treatment group was increased by one day in each iteration. With each subsequent iteration, the control group was increasing and the treatment group was decreasing, but the total sample size remained constant, and the average usage was going up in both groups.

This analysis showed that including users with low usage produced no positive results. A significant treatment effect is first observed when the dividing line is set at 19 days and it is maximized at 29 days. In the second case, the average usage days in the treatment group is 38. In the interim—when the threshold is set between 20–28 days—the effect fluctuates around 0.2 and the average usage days for the group equals 36. These results suggest that about weekly usage (taken conventionally at 36 days per school year) is the usage level that makes the difference.

TABLE 6. TREATMENT EFFECT ESTIMATES UNDER ALTERNATIVE DESIGNS

Design	Treatment group average usage days	Treatment effect	*p* value
Users (>1 day) vs. non-users	14	-0.063	0.37
High and moderate users vs. low users (threshold = 19 days)	27	0.137	0.04
High users vs. low and moderate users (threshold = 29 days)	38	0.251	0.01

Appendix B. Detailed Results of Moderator Analysis

The table below present detailed results of moderator analysis: analysis of variation in the association between Goalbook Toolkit usage (in days) and student outcomes across student groups. In each case, the results are reported as the additional effect (per one day of usage) for a given group, compared to the base group. For binary variables such as English language learner status, the base group is all the rest (non-English language learners in this case). For racial/ethnic groups, the results are reported in relation to Black students—the prevalent racial/ethnic group in the sample. Sixth grade is the base group for grade-level moderators. The effects are on the effect size scale since the outcomes are normalized.

TABLE 7. MODERATOR ANALYSIS RESULTS

Category	Differential effect (per usage day)	*p* value
Female	0.0059	.24
ELL	0.0027	.61
FRPL	0.0096	.04
GT	-0.0080	.66
Hispanic	0.0088	.72
White	0.0136	.08
Other racial/ethnic groups	0.0031	.85
Grade 7	0.0111	.05
Grade 8	0.0047	.42

Note. FRPL stands for students who are eligible for free or reduced-price lunch. ELL stands for students classified as English language learners. GT stands for student enrolled in the gifted and talented program.

Reference this report: Empirical Education Inc. (2024). *Impact of Goalbook Toolkit*. (Empirical Education report number: Empirical_Goalbook-6095-EVI-2024-O.3).