BookNook in Texas: Implementation and Student Outcomes

Douglas Ready & Jennifer Sallman
Teachers College, Columbia University

December, 2022

For information about this study, contact the first author at ddr2111@tc.columbia.edu. We offer our sincere thanks to BookNook staff for their effort and partnership. All errors of either fact or interpretation are solely those of the authors.

Abstract

Texas partnered with BookNook, an online platform focused on early literacy development, to address students' unfinished learning during the 2021-22 school year. Students who qualified based on criteria established in HB 4545 were provided the opportunity to participate in remote, small-group tutoring with a school-provided or BookNook professional tutor for three, thirty-minute BookNook sessions per week. This report examined the link between BookNook participation and student literacy development from mid-February through late-March, 2022. Our findings suggest that fourth- and fifth-grade BookNook students with higher levels of usage gained literacy skills at a somewhat faster rate compared to their same-school, same-grade peers who engaged the platform less frequently. The study has a number of limitations, including unreliable pre- and post-scores, a short implementation window, considerable variation in BookNook usage, and the non-random and voluntary enrollment process.

Introduction

In 2021, the Texas Legislature passed House Bill (HB) 4545 in response to concerns about unfinished learning due to the COVID-19 pandemic and related school closures and instructional disruptions. The legislation required school districts to provide accelerated instruction during the 2021-2022 school year for students who did not meet grade-level performance expectations on the State of Texas Assessments of Academic Readiness (STAAR), or who did not take the STAAR and STAAR end-of-course assessments. Local educational agencies were provided funding and programmatic support under Texas COVID Learning Acceleration Supports (TCLAS) to design and implement high-quality afterschool programs aimed at addressing unfinished learning in the wake of the pandemic.

The BookNook program was included as one instructional option as part of HB 4545's initiative to deliver high-impact literacy tutoring for Texas K-8 students. BookNook is a web-based tutoring platform focused on literacy development in small-group sessions during the school day. BookNook students are encouraged to participate in three sessions per week. However, participation is voluntary, which produces considerable variability in usage, as we discuss in more detail below. Students are placed into small groups based on beginning-of-year reading levels and generally have one tutor throughout their experience. Schools and districts determine how and when BookNook is implemented—some leverage BookNook professional tutors while others provide their own tutors. In addition to the platform's instructional materials, BookNook provides staff with simple, easy-to-read reports that indicate the skills each student has mastered and where they need more practice. The platform gathers data in real time and provides progress monitoring tools and actionable insights to adults.

In this study, we explore the implementation of BookNook in Texas as part of this State initiative. Our focus is on a single six-week implementation period that ran from mid-February through late-March, 2022. The period captures seven weeks, but one week was the Texas spring break, when usage rates were very low. As such, we refer to a six-week implementation period throughout the report. In the sections below, we begin by describing our data and methods. We then examine the degree to which students engaged with the platform with the expected degree of fidelity. In the following section we provide the results of our central analyses, which estimate the link between student BookNook usage and literacy development. We conclude with a brief discussion and possible directions for future research.

Data and Methods

Literacy Assessments

Neither formative assessment outcomes nor state assessment results were available for these analyses. Instead, BookNook staff created literacy assessment scores using the approximately four assessment items that participating students completed at the end of each BookNook session during this implementation. The baseline (pre-test) assessment scores are based on student responses from sessions completed between early November and early February. The outcome (post-test) assessment incorporated assessment items from BookNook sessions in April through the first week of May. We conceptualize the treatment phase as the seven-week intervening period. One limitation of this approach is that both the baseline and outcomes assessment scores were gathered while the treatment was occurring.

To estimate reliable individual student-level test scores, psychometricians ideally incorporate dozens of individual test items into a single assessment score. However, given the relatively low usage rates among BookNook students in Texas during the pre and post phases when the test items were administered, less than ten percent of the participating students had a sufficient number of test items to create reliable assessment scores. Recognizing this limitation, BookNook staff created IRT scale scores for all students with at least one completed BookNook session, and provided information on the number of items incorporated into each student's pre and post assessment. The analyses below leverage data from three student samples, based on the number of items included in each student's assessment scores. Our base analytic sample includes 2,624 second- through eighth-grade students who attended one of 141 schools in 70 Texas school districts or charter management organizations. This sample, which includes all students with at least one session during each assessment period, provides the most statistical power, but the assessment scores are also the least reliable, given that many students have scores created using a small number of items. The second sample is restricted to students who completed at least two sessions (5 or more items) during both the pre and post periods (n=1,559), while the third sample includes students with test items from at least three sessions (9 or more items; n=994). The unstable nature of the assessments at the student level is evident in the weak correlation between

the pre and post assessments (r=.28), lower than one might expect.[1] The assessments using larger numbers of items appear to be somewhat more reliable, with correlations between pre and post assessments from the medium sample of 0.32, and a correlation of 0.37 using the smallest sample (with the most items per score). BookNook staff are well aware of these limitations, and created the most reliable assessment scores possible given the data available.

Analytic Approach

We explore the link between the number of BookNook sessions students completed and their literacy development through a series of OLS regression models that included the baseline literacy assessment score as a covariate. These analysis of covariance (ANCOVA) or lagged-score models took the form:

$$Y_{ij} = b_0 + b_1(\text{BookNook}) + + \delta_k + e_{ij} \quad (1)$$

where Y_{ij} is the end-of-period literacy test score for student i in school j. The models employ two separate indicators of BookNook usage: a continuous indicator of the number of sessions completed and a series of dichotomous measures comparing literacy development among medium- (6-11 sessions) and high-usage students (12+ sessions) to low-usage students (1-5 sessions). We tested for a non-linear relationship between the continuous usage measures and student literacy development, but the quadratic (squared term) was non-significant. δ_k represents school fixed effects, which allow us to compare students with varying levels of BookNook usage to other students within the same school, and e_{ij} is the error term for student i in school j. Unfortunately, we were not provided student social or academic background information, and thus cannot estimate the extent to which either usage rates or the links between BookNook usage and student literacy development varied as a function of student characteristics. We use standardized (z-scored) versions of the both pre and post assessments.

[1] For example, the beginning- and end-of-year kindergarten literacy assessments from the Early Childhood Longitudinal Study, Kindergarten Cohort (ECLS-K), administered roughly six months apart, have correlations of approximately 0.79.

Results

Descriptive Results

Table 1 provides student information organized by the degree to which they engaged BookNook. The average student completed roughly 6.5 BookNook sessions in total during the six-week period, or just over one session per week. Students in the medium-usage category completed almost five more sessions in total than did their low-usage peers ($p<.001$), while students in the high-usage category completed over 11 more sessions ($p<.001$). It is important to note that even high-usage students, on average, did not complete the recommended dosage of three sessions per week (or a total of 18 sessions during the six-week period). As such, readers should bear in mind that our estimates of the link between BookNook usage and student literacy development may be conservative and biased downward, given the relative lack of student engagement with the platform.

In terms of student grade levels, although the sample includes students in second through eighth grade, well over half of the students are in either fourth or fifth grade. Importantly, participation rates vary across grade levels ($p<.001$), but the patterns are not consistent. Fourth graders are under-represented among medium-usage students (and over-represented among low- and high-usage students), while fifth graders are under-represented in the low-usage category.

Finally, we find that low-usage students started the period with slightly higher baseline literacy assessment scores compared to their medium-usage peers ($ES = 0.114$; $p<.01$). The gap with high-usage students is comparable, but non-significant due to the smaller subgroup sample size. However, during the six-week period, medium-usage students closed the gap with low-usage students, and now have statistically comparable assessment scores. High-usage students, who gained somewhat less during the period, continue to score marginally below their low-usage peers ($ES = 0.125$; $p<.10$). Given that BookNook usage is often related to student characteristics, these counterintuitive results are likely driven by the differences in the students in each usage category.

Table 1. Student Characteristics by BookNook Usage Categories (*n*=2,624)

	Low Usage 1-5 sessions (*n*=1,315)	**Medium Usage** 6-11 sessions (*n*=906)	**High Usage** 12+ sessions (*n*=403)	TOTAL
Number of Sessions	3.14	8.01***	14.37***	6.55
Grade***				
Second	3.9	4.6	7.4	4.7
Third	16.4	16.3	15.6	16.3
Fourth	35.1	27.2	32.0	31.9
Fifth	19.7	32.0	32.5	25.9
Sixth	12.1	11.9	9.4	11.6
Seventh	9.7	4.0	1.2	6.4
Eighth	3.0	4.0	1.7	3.2
Literacy Pre-Score				
Scale Score	518.2	508.1*	509.1	513.3
Scale Score (z)	0.055	-0.059*	-0.046	0.000
Literacy Post-Score				
Scale Score	531.8	537.4	519.8~	531.9
Scale Score (z)	-0.001	0.057	-0.126~	0.000

*p<.05; **p<.01; ***p<.001. Low-usage category is the comparison group for statistical tests involving continuous outcomes (all but grade).

Analytic Results

Table 2 presents the results of OLS regressions estimating the link between BookNook usage and student literacy development. As noted above, we present estimates from three different samples, based on the number of items incorporated into each assessment scale score. The BookNook estimates are adjusted for grade level and baseline literacy assessment score. All models included school fixed effects, meaning that the estimates compare literacy development among students attending the same school, thus removing potential differences across schools in their effectiveness at promoting literacy development.

Model 1, which employs data from the largest sample (with the overall least reliable assessments), indicates that each additional session completed was associated with a marginally significant 0.011 *SD* increase in literacy learning (p<.10). Assuming a linear relationship, this suggests that students who completed all 18 sessions during the treatment period would have experienced a 0.198 *SD* advantage, or a movement from the 50th to almost the 58th percentile. With the categorical usage variables in Model 2, we find some evidence of a linear relationship between usage and literacy learning—a positive medium-usage estimate (compared to low-usage

students), and a slightly larger high-usage estimate. The estimates, however, are non-significant (p>.10), due in part to the relatively small number of students in each category and the modest size of the estimates. Although the continuous usage estimates in Models 3 and 5 are virtually identical to that reported in Model 1, the smaller sample sizes render them non-significant (p>.10). The same is true for the high-dosage categorical usage estimate.

Table 2. BookNook Usage and Student Literacy Development in Grades 2-8

	Large Sample (n=2,624)		Medium Sample (n=1,559)		Small Sample (n=994)	
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
# Sessions	**0.011~**	--	**0.011**	--	**0.010**	--
6-11 Sessions	--	**0.070**	--	**0.018**	--	**-0.004**
12+ Sessions	--	**0.110**	--	**0.131**	--	**0.092**
Baseline Test Score	0.240***	0.241***	0.281***	0.281***	0.307***	0.306***
Constant	-0.240***	-0.209***	-0.261**	-0.215**	-0.227*	-0.171*

*p<.05; **p<.01; ***p<.001. Outcome is spring literacy assessment score; outcome and baseline assessment scores are standardized (z-scored). All models include school fixed effects and controls for grade (with fourth grade as the omitted comparison group).

Table 3 presents results from models that used the larger sample, with separate analyses for grades three, four, and five. These students represent almost three-quarters of the sample. Among third graders, we find no significant associations between usage and literacy growth, even at the p<.10 level. With the fourth-grade sample, each additional BookNook session was associated with a 0.02 *SD* increase in literacy learning (p<.10). Neither categorial indicator was significant. We find the most substantively and statistically meaningful estimates with the fifth-grade sample, where each additional session was associated with a 0.037 *SD* advantage (p<.01). For the small number of students who completed all 18 recommended sessions, this represents a 0.67 *SD* developmental advantage, or a move from the 50th to the 75th percentile. The categorical usage estimates suggest that high-usage students gained substantially more than their low-usage peers (*ES* = 0.459; p<.01). For these high-usage fifth-graders, that gain represents a move from the 50th to the 68th percentile.

Table 3. BookNook Usage and Student Literacy Development: Grades 3, 4 and 5

	Grade 3 (*n*=427)		Grade 4 (*n*=837)		Grade 5 (*n*=680)	
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
# Sessions	**-0.027**	--	**0.020~**	--	**0.037****	--
6-11 Sessions	--	**-0.084**	--	**0.000**	--	**0.169**
12+ Sessions	--	**-0.290**	--	**0.051**	--	**0.459****
Baseline Test Score	0.208***	0.210***	0.238***	0.238***	0.207***	0.208***
Constant	0.170**	0.072	-0.127	-0.008	-0.276*	-0.161*

*p<.05; **p<.01; ***p<.001. Large sample used for all grades. Outcome is spring literacy assessment score; outcome and baseline assessment scores are standardized (z-scored). All models include school fixed effects.

Conclusion and Discussion

This report explored the implementation of BookNook in Texas during the 2021-22 school year. We found that usage rates among BookNook students were lower than expected, with less than one percent experiencing the full recommended dosage. However, fifth graders who experienced relatively higher usage levels experienced stronger literacy development, compared to their same-school, same-grade low-usage peers. We also found positive (but less robust) associations with the fourth-grade sample. It is perhaps not surprising that we were able to detect significant associations between usage and literacy development among fourth and fifth graders. They constitute almost one-third and over one-quarter of the sample, respectively, and the pre- and post-assessment scores were calibrated primarily on third- through fifth-grade lessons, as not enough students completed sessions at the K-2 or 6-8 levels to calibrate scores for those grade bands. This suggests that second graders were completing higher grade-level lessons, and conversely, sixth through eighth graders were completing sessions well-below their grade level to be included in our sample. In fact, most sixth through eighth graders were reading at a fourth-grade level).

Our findings should be interpreted in light of several limitations with the literacy assessment scores as well as with the implementation. Foremost is the fact that the literacy assessment scores were created using the treatment (BookNook lessons). Fourteen weeks of the treatment were used to create the pre-scores and five weeks of the treatment were used to create the post-scale scores. Using outcomes from BookNook lessons to estimate student baseline

literacy scores is questionable, given that students were actively participating in the treatment to generate those baseline scores. Additionally, BookNook typically requires 28 items or more to estimate reliable scores for an individual; however, less than ten percent of students completed 28 or more items for both their pre-scores and post-scores.

There are also limitations associated with the implementation. First, student participation was voluntary. Since we do not have student demographic information for BookNook or non-BookNook students, we do not know if and how BookNook students differed from their non-BookNook peers (or between high and low users). Although we used school fixed effects to compare students within the same school, where students may be more similar to one another, one concern is that other unmeasured differences may bias our estimates in unknown directions. A second consideration is the very short, six-week implementation period. Even under the best of circumstances, measuring academic growth during such a short period is a challenge. Third, session completion rates varied dramatically among BookNook users, and very few participants experienced the three-times-weekly recommended dosage. One might not expect dramatic changes to student literacy ability given both the limited implementation window and the variable levels of engagement among participants. It is unclear what the platform's impact might have been had students engaged BookNook for the full academic year and to the extent expected. Fourth, the form and quality of instruction provided by the tutors certainly varied, but we have no information or data in this regard.

Directions for Future Research

Additional research on BookNook should seek to establish the causal impact of the platform on student outcomes using implementation approaches that differ from those employed in Texas. First and foremost, student assignment to BookNook (the treatment) should be random or based on clear metrics such as student baseline literacy test scores. Second, the implementation might be longer to allow the development of more skills and also include efforts to increase the likelihood that students will participate to the degree expected. Third, we strongly urge the use of a district-administered external assessment such as MAP as the primary indicator of student literacy achievement. BookNook has many compelling and promising elements. A stronger implementation and evaluation strategy would increase the likelihood of identifying its full potential.