ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 24-942

The Effects of In-School Virtual Tutoring on Student Reading Development: Evidence from a Short-Cycle Randomized Controlled Trial

Douglas D. Ready
Teachers College,
Columbia University

Sierra G. McCormick
Teachers College,
Columbia University

Rebecca J. Shmoys
Teachers College,
Columbia University

This paper describes a 12-week cluster randomized controlled trial that examined the efficacy of BookNook, a virtual tutoring platform focused on reading. Cohorts of first- through fourth-grade students attending six Rocketship public charter schools in Northern California were randomly assigned within grades to receive BookNook. Intent-to-Treat models indicate that students in cohorts assigned to BookNook outperformed their control-group peers by roughly 0.05 SDs. Given the substantial variability in usage rates among students enrolled in BookNook cohorts, we also leveraged Treatment-on-the-Treated approaches. These models suggest that students who completed 10 or more BookNook sessions experienced a reading advantage of 0.08 SDs, while those who completed 20 or more sessions—the recommended dosage—experienced a 0.26 SD developmental advantage.

VERSION: April 2024

Suggested citation: Ready, Douglas D., Sierra G. McCormick, and Rebecca J. Shmoys. (2024). The Effects of In-School Virtual Tutoring on Student Reading Development: Evidence from a Short-Cycle Randomized Controlled Trial. (EdWorkingPaper: 24-942). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/569p-wz78

The Effects of In-School Virtual Tutoring on Student Reading Development: Evidence from a Short-Cycle Randomized Controlled Trial

Douglas D. Ready

Sierra G. McCormick

Rebecca J. Shmoys

Teachers College, Columbia University

April 3, 2024

Corresponding Author:
Douglas D. Ready
Professor of Education and Public Policy
Teachers College, Columbia University
525 W. 120th St., Box 11
New York, NY 10027
ddr2111@columbia.edu

Abstract

This paper describes a 12-week cluster randomized controlled trial that examined the efficacy of BookNook, a virtual tutoring platform focused on reading. Cohorts of first- through fourth-grade students attending six Rocketship public charter schools in Northern California were randomly assigned within grades to receive BookNook. Intent-to-Treat models indicate that students in cohorts assigned to BookNook outperformed their control-group peers by roughly 0.05 SDs. Given the substantial variability in usage rates among students enrolled in BookNook cohorts, we also leveraged Treatment-on-the-Treated approaches. These models suggest that students who completed 10 or more BookNook sessions experienced a reading advantage of 0.08 SDs, while those who completed 20 or more sessions—the recommended dosage—experienced a 0.26 SD developmental advantage.

Keywords: tutoring, literacy, supplemental interventions

The Effects of In-School Virtual Tutoring on Student Reading Development: Evidence from a Short-Cycle Randomized Controlled Trial

Over the past several years, empirical studies have reported positive impacts of one-to-one and small-group tutoring on student academic development (Kraft & Falken, 2021). This research has garnered considerable attention from policymakers and practitioners, given the dramatic declines in student learning associated with the COVID-19 pandemic (Fahle et. al., 2023; Guryan & Ludwig, 2023; Peters et al., 2023). One takeaway from this literature, however, is the tremendous variability in effectiveness across tutoring approaches and implementations (Robinson & Loeb, 2021). BookNook, a Tier 2 supplemental intervention focused on struggling readers, was designed to address many of the limitations of previous tutoring models and platforms. In this paper, we describe the results of a short-cycle cluster randomized controlled trial that examined the impact of BookNook on student reading outcomes. Cohorts of first- through fourth-grade students attending six Rocketship public charter schools in Northern California were randomly assigned to receive tutoring via the BookNook platform. Control cohorts of students within the same grade and school continued to receive the reading supports and activities that Rocketship normally provides.

In the sections below we begin by reviewing the extant literature on the links between tutoring and student academic outcomes. We then provide information about BookNook and its implementation in these six schools. We follow with a description of our data and analytic approaches. Our results section includes information on baseline equivalency between treatment and control groups, insights into fidelity of implementation, and the results of our primary models estimating the causal impact of BookNook on student learning. We also conduct a series of robustness checks and explore treatment effect heterogeneity across student academic and

demographic subgroups. We close with a summary of our findings and discuss potential implications for future research on BookNook specifically and supplemental tutoring programs more broadly.

Background

Since Bloom's (1984) early reports on tutoring effects, the body of literature estimating the links between tutoring programs and student learning has grown dramatically, with high-dosage tutoring increasingly considered among the most effective instructional tools. One meta-analysis of academic interventions intended for low socioeconomic-status elementary and middle school students reported that tutoring had the largest average effect (0.36 SD) among 14 other interventions (Dietrichson et al., 2017). Other meta-analyses corroborate these benefits, reporting consistent and substantial positive impacts of tutoring on student learning (Fryer, 2016; Nickow et al., 2020). These links between tutoring and academic performance are generally stronger for students who struggle academically. A randomized control trial involving nearly 1,000 first graders in the bottom tertile of math achievement found that small-group, high-dosage tutoring led to a considerable increase in mathematics performance (0.34 SD; Gersten et al., 2015).

The efficacy of tutoring also tends to vary by subject and grade level. Reading interventions typically yield better results in the earlier grades, while math tutoring is generally more effective in later grades (Nickow et al., 2020). Indeed, a study on early elementary literacy tutoring by trained AmeriCorps members indicated positive effects across all grade levels (K-3), with the most substantial impact reported in kindergarten (1.06 SD; Markovitz et al., 2014). Similarly, kindergarten students who received short bursts of one-on-one reading instruction from part-time tutors were two times more likely to reach target reading levels by the end of kindergarten (Cortes et al., 2024).

Despite this body of promising research, the effectiveness of tutoring depends on program and implementation characteristics. Dosage stands out as a particularly critical feature, with high-dosage tutoring reportedly 20 times more effective than low-dosage tutoring in math, and 15 times more effective in reading (Fryer, 2016; Robinson et al., 2021). Although there is no uniform definition of high-dosage tutoring, considerations typically include session frequency and duration, program length, and student-to-tutor ratio. Fryer (2016) characterized high-dosage tutoring as involving small groups (six or fewer students) meeting more than three times per week, amounting to 50 or more hours over a 36-week period.

Beyond dosage, tutoring interventions differ in terms of tutor expertise, curriculum content, and delivery methods. Tutoring conducted by teachers has the largest impact on student learning, followed by programs led by paraprofessionals (Nickow et al., 2020), perhaps due to the importance of sustained and meaningful tutor-student relationships (Robinson & Loeb, 2021). Successful programs also tend to involve structured, sequenced approaches, with manuals and materials to guide instruction (Neitzel et al., 2022; Robinson & Loeb, 2021). Formative assessment, which provides tutors timely feedback on student progress, is also seen as a valuable tool to enable personalized instruction and increase tutor effectiveness (Jacob et al., 2016). Tutoring programs that are both high-dosage and delivered using effective strategies are often referred to as “high-impact” (Cortes et al., 2024; Robinson & Loeb, 2021).

Although the promise of high-impact tutoring is clear, widespread and sustained implementation has proven difficult. Schools and districts face challenges associated with the higher costs of individualized instruction compared to conventional classroom teaching, as well as difficulty hiring competent tutors (Makori et al., 2024). Although most schools implement some form of tutoring, roughly one-third report that not all students who would benefit from

high-dosage tutoring receive it (NCES, 2023). This dilemma highlights the need for effective strategies that support tutoring but within the constraints of educational budgets (Nickow et al., 2020). One recommendation is to leverage paraprofessionals, who produce positive tutoring impacts, but are more cost effective compared to certified teachers (Guryan et al., 2021). Another solution is to establish partnerships with colleges and universities or re-staff retired district teachers to maintain a well-trained tutor supply (Makori et al., 2024).

Beyond challenges with cost and hiring, student uptake of available tutoring services is also inconsistent, with few eligible students signing up for services and sporadic attendance among those who do (McCormick et al., 2023). Given the challenges associated with optional programs, providing tutoring during school hours might improve attendance and participation rates (Kraft & Falken, 2021). Tutoring programs held during the school day are also more effective, nearly doubling impacts compared to after-school options (Nickow et al., 2020). However, the timing of school-day sessions is important, as the displacement of other activities, particularly core academic subjects, can have unintended negative effects (Robinson & Loeb, 2021).

Virtual tutoring may also address resource limitations, as it widens the pool of potential tutors (Kraft et al., 2022; Kraft & Falken, 2021). Whether virtual tutoring is as effective as in-person tutoring remains an open question. One frequently cited randomized controlled trial of online tutoring for middle schoolers in Italy showed a substantial effect (0.26 SD; Carlana & Ferrara, 2021). However, this study was conducted during pandemic lockdown, suggesting that the findings should be interpreted with caution. Outside of a lockdown context, the results of virtual tutoring are mixed (Kraft et al., 2022; Loeb et al., 2023), and even in lockdown contexts not all virtual tutoring interventions are effective (Schueler & Rodriguez-Segura, 2023). One

concern with virtual tutoring is the challenge of ensuring consistent attendance, given that student usage rates tend to be lower. However, this research has only explored tutoring programs held outside of school hours (Burch et al., 2016); the question remains whether participation rates of in-school virtual tutoring match those of in-school, in-person tutoring approaches. Our current work is among the few causal studies to estimate the impact of virtual tutoring structured into the school day.

The BookNook Intervention

BookNook is a Tier 2 intervention designed for struggling readers. It uses a synchronous teaching platform and curriculum grounded in the science of reading to deliver high-dosage tutoring services. The model features lessons that support students in building phonological awareness, phonics, fluency, vocabulary and comprehension skills. The phonological awareness/phonics lessons focus on foundational reading skills. Lessons begin with a skill introduction. Tutors present the skill and relevant examples for students to both see and hear. For example, students may be introduced to consonant-vowel-consonant words, such as “cat” or “big,” and will be prompted to break the words apart to hear the different sounds. Students then move to skill practice, where they apply the skill or standard through two to five activities depending on the lesson. During the foundational text reading section, students practice the focal skill with a text that incorporates a high number of words aligned to that skill. The texts in these lessons integrate the components of phonics, fluency, and text reading comprehension skills through authentic practice reading. After finishing the text, students discuss what they have read. Tutors are encouraged to ask questions that aim to engage students in the text and form a deeper understanding of what they have just read. Lastly, each phonics lesson ends with a formative assessment, which provides data on student progress towards mastery of the skill through four to

five aligned questions. These formative data are incorporated into the program's algorithm for determining student advancement to a new set of lessons.

During BookNook's fluency lessons, students engage in activities that support practice in oral reading fluency. During the introduction, students review the three components of fluency: accuracy, pace, and expression. Tutors then model or play a recording of the fluent reading of a passage. Students also have the opportunity to read the same passage aloud and practice the same techniques they heard in the modeled reading. Following the modeled reading, students engage in a fluency activity called "What's Wrong," during which they listen to a tutor or audio reading and evaluate the components of fluency. Finally, students read a passage, while the tutor notes any errors. The BookNook algorithm uses these notes to calculate and help the tutor understand the student's accuracy and words-per-minute score.

Comprehension lessons begin with vocabulary instruction, given that lack of vocabulary understanding can impede text comprehension. Vocabulary that are essential to the comprehension of a lesson text are pre-taught in a scaffolded exposure that includes words, definitions, audio, images, and typing exercises. Following direct vocabulary instruction, students engage in an interactive matching activity that provides additional exposure to the lesson's vocabulary words and their meaning. At the conclusion of the vocabulary section of the lesson, students engage in a check-for-understanding activity that allows them to see and place vocabulary words in context in various sentences. As students transition from vocabulary to comprehension, they engage in a pre-read strategy session to help build engagement and motivation around the upcoming text. These discussions are designed to prompt activation of pre-reading strategies such as activating prior knowledge and making predictions.

Tutors then lead a readthrough with group discussion questions. These questions aim to support students in building the skills aligned to the lesson standard and to develop reading comprehension skills more generally. The discussion questions are designed to prompt students to think critically, make inferences, engage in vocabulary work, and focus on the lesson-aligned standard. Immediately following the readthrough, students discuss and synthesize what they have just read. Students then engage in a comprehension activity called "Feed the Animals" that focuses on identifying main ideas, themes, summarizing, retelling and/or sequencing to build comprehension of the text. Lastly, students engage in individual analysis of the text through text-dependent questions aligned to the lesson standard. This formative assessment provides data on their progress towards mastery of the skill or standard.

The Implementation

Our current study examined the effects of BookNook's virtual tutoring delivery format on student reading development. Most of the participating virtual tutors had previously taught K-12 academic subjects, with over three years of tutoring or teaching experience on average. The implementation involved first- through fourth-grade students enrolled in six Rocketship public charter schools in Northern California. As part of its regular instructional programming, Rocketship organizes students into same-grade cohorts containing roughly 20-30 students each, usually resulting in three to four cohorts per grade, depending on enrollments. Rocketship students experience four content blocks each day: Humanities, STEM, Enrichment, and Learning Lab. Within each school and grade, we randomly assigned all cohorts to treatment or control groups. Students enrolled in treatment cohorts were to receive BookNook tutoring during their Learning Lab period two to three times per week for 30 minutes per session. Students in cohorts assigned to the control condition would continue with the regular reading supports provided

during Learning Lab. The roughly 12-week implementation began in late January, 2023 and concluded in early May, 2023. During this period the Rocketship academic calendar included two week-long vacations. As such, the actual intervention period was 10 weeks, with full treatment exposure calculated as 20-30 BookNook sessions.

Research Questions

To address gaps in the extant literature and better understand the potential efficacy of in-school synchronous virtual tutoring, our study was designed to investigate the following research questions:

1. What is the impact of BookNook on student reading growth?
2. How does usage of BookNook vary among students, and how does usage relate to the relative impact of BookNook on student reading growth?
3. Do measured BookNook effects vary as a function of student academic and demographic background characteristics?

Data and Methods

Our data include student-level MAP test scores, student academic and demographic measures, variables that link students to grade-level cohorts and schools, and indicators of BookNook usage. All cohorts completed the study in their original treatment and control states, and the study experienced zero assignment-level attrition. Seven students declined to participate in the study prior to implementation; no students withdrew from the study during implementation. The initial sample included 1,900 first- through fourth-grade Rocketship students. Our analytic sample only includes students with full demographic and assessment data. No students were missing demographic data. There were, however, student-level missing data associated with the baseline and follow-up MAP assessments, with 6.5% of students missing

data on one or both assessments. Missingness rates were virtually identical across students assigned to treatment and control cohorts. This baseline and follow-up assessment restriction necessarily excludes students who enrolled in either treatment or control groups during the implementation (i.e., no joiners are included in the sample). No students with full data (and thus included in these analyses) switched treatment/control cohorts during the implementation.

Our final analytic sample includes 77 student cohorts (*n*=42 treatment, 35 control) containing 1,777 first- through fourth-grade students (*n*=959 treatment, 818 control), of whom 79% are Hispanic, 9% are Black, 8.6% are Asian, 2.4% are white and 1.1% are American Indian/Alaskan Native or Native Hawaiian/Pacific Islander. ELL students represent just over half of the sample, and 9.5% of students receive special education services. Roughly 48% of students are identified as female.

Measures

Outcome. MAP is a computer-adaptive assessment that measures student academic growth, producing scores that are vertically equated using the Rasch unit (RIT) scale. In first grade, MAP measures foundational skills (e.g., phonics and phonological awareness), language and writing, literature and informational text, and vocabulary use and functions. In second through fourth grade, MAP captures vocabulary acquisition and use, understanding and integrating key ideas and details for literature and informational text, and understanding and interpreting craft and structure for literature and informational text (NWEA, 2019). These multiple skill areas assessed by MAP overlap nicely with the content covered by BookNook. In norming studies, MAP test–retest reliabilities ranged from .73 to .89, and concurrent validity with elementary-level state reading tests ranged from .58 to .83 (NWEA, 2019).

Rocketship Schools administers the MAP assessments three times each year—Fall, Winter, and Spring—as part of its regular assessment program. We use reading results from the Winter administration as our baseline measure and reading scores from the Spring administration as the follow-up (post-implementation) outcome. The Winter MAP administration occurred in December prior to the BookNook implementation, and the Spring administration took place in mid-May, after the conclusion of the study. Scores at each timepoint were standardized (z-scored) within grade.

Covariates. Because of the random assignment process, OLS estimation will provide unbiased treatment estimates and it is not necessary to control for other student characteristics. However, including pre-random assignment covariates that are correlated with the outcome in our models can improve impact estimate precision. As covariates, our models include dummy indicators of student race/ethnicity (Asian, Black, white and other race/ethnicity) with Hispanic students as the uncoded comparison group. Unfortunately, confidentiality concerns related to small sample sizes required us to organize American Indian/Alaskan Native and Native Hawaiian/Pacific Islander students into a single "other race/ethnicity" category. Our analyses also leverage data on student sex (female = 1, male = 0) and special education (IEP) and English language learner (ELL) status (yes = 1, no = 0).

Cohort Baseline Equivalency

To establish baseline equivalence across treatment and control cohorts for the analytic sample of students, we constructed a series of nine separate OLS regression models in which the cohort-average baseline MAP assessment score and aggregate means of the eight student demographic variables served as outcomes. These models, which parallel the impact models discussed below, can be described as,

$$Y_{cg} = b_0 + b_1(BookNook) + \eta + e_i$$

where Y_{cg} represents the average standardized baseline MAP reading assessment score or demographic variable for cohort *c* in grade *g*. *BookNook* is an indicator of whether the cohort was randomly assigned to the treatment condition. School-by-grade fixed effects are indicated by η, while e_i indicates the cohort-level error term.

Analytic Approach

We employed two primary analytic techniques with these data to measure the impact of BookNook on student reading growth. The first approach provides the average causal effect of being assigned to the treatment group, often referred to as the "Intent-to-Treat" (ITT) estimate. This approach is thought of as producing the most policy-relevant indicator of program impact given the typical constraints faced by social interventions implemented in the field (Glennerster & Takavarasha, 2013). Individuals or groups assigned to a treatment may not comply—hence the phrase, "*intent* to treat." To estimate the average effect of being randomly assigned to a BookNook cohort, relative to the outcomes of students assigned to control cohorts, we estimated a two-level model with fixed block effects and a fixed treatment effect of the following form:

$$Y_{icg} = b_0 + b_1(BookNook) + X_i + \eta + e_i$$

where Y_{icg} represents the standardized MAP follow-up reading assessment score for student *i*, in cohort *c,* in grade *g*. *BookNook* is an indicator of whether the student's cohort was randomly assigned to participate in BookNook. X_i represents a vector of student-level covariates, including the baseline MAP reading assessment score, race/ethnicity, gender, and IEP and ELL status.

School-by-grade fixed effects are indicated by η, while e_i indicates the student-level error term. In all models robust standard errors are clustered at the cohort level.

Our second analytic approach entailed two-stage least squares instrumental variable models that explored whether increased BookNook usage among students in treatment cohorts was associated with increased learning. Recall that students assigned to treatment cohorts were to complete two to three, 30-minute sessions per week. However, as we discuss in more detail below, student usage rates were generally below what was expected. Because student cohorts were randomly assigned to BookNook, we can conceptualize the treatment of being assigned to a BookNook cohort as an "instrument" for participation in the program. Instrumental variable analysis is feasible in this case because we have met the "exclusion restriction," in which random assignment to the treatment group can only affect student test scores through actual participation in BookNook, or compliance with the prescribed treatment (Angrist & Pischke, 2009). This type of analysis is considered the "Treatment-on-the-Treated" (TOT) approach, revealing the complier average causal effect of BookNook. We are confident that students assigned to control cohorts were not provided BookNook accounts or logins during the implementation. Further, we know that random assignment at the cohort level was the only mechanism inducing student participation in the treatment, as again, control cohorts were not provided access to the platform.

With this approach, the first-stage model took the form,

$$BookNook\ Usage_{icg} = b_0 + b_1(Treatment\ Status)_{cg} + X_i + \eta + e_i \qquad \text{(First Stage)}$$

where $Treatment\ Status_{cg}$ is an instrument for $BookNook\ Usage_{icg}$. The second-stage model can then be expressed as,

$$Y_{icg} = b_0 + b_1(\widehat{BookNook\ Usage}) + X_i + \eta + e_i \qquad \text{(Second Stage)}$$

where Y_{icg} is the standardized MAP follow-up reading assessment score for student *i*, in cohort *c,* in grade *g*. This model uses the *BookNook Usage* estimates from the first-stage model. We estimate two separate parameters based on treated students' BookNook usage: 1) a binary indicator of students who completed 10 or more sessions during the implementation period; and 2) a binary indicator of students who completed 20 or more sessions, which is the minimum recommended treatment dosage based on two sessions per week for the 10 weeks of actual instruction. X_i represents the vector of student-level background covariates described above, as well as the standardized baseline MAP scores. School-by-grade fixed effects are indicated by η, while e_i indicates the student-level error term. In all models robust standard errors are again clustered at the cohort level. It is important to stress that the "complier average causal effect" resulting from these models are relevant only for the types of students who would use BookNook at these higher rates given the opportunity to do so. These effects would not necessarily result if all treatment students had engaged at these levels.

Robustness checks. We conducted two sets of robustness checks. With the first, we reconstructed the Intent-to-Treat model described above as a three-level random effects model with random block effects and a random treatment effect. This model—which analytically nested students within cohorts, nested within grade-by-school clusters—can be described as,

$$\textit{Level 1 (Students):}\ Y_{icg} = \pi_{0cg} + \pi_{1cg}(Baseline\ MAP_{icg}) + \pi_{2cg}(Asian_{icg}) + \pi_{3cg}(Black_{icg}) + \pi_{4cg}(White_{icg}) + \pi_{5cg}(Other_{icg}) + \pi_{6cg}(Female_{icg}) + \pi_{7cg}(IEP_{icg}) + \pi_{8cg}(ELL_{cg}) + e_{icg}$$

$$\textit{Level 2 (Cohort): } \pi_{0cg} = \beta_{00g} + \beta_{01g}(BookNook_{cg}) + r_{0cg}$$

$$\textit{Level 3 (Grade-by-School): } \beta_{00g} = \gamma_{000} + u_{00g}$$

where, at the student level (Level 1), Y_{icg} is the standardized MAP follow-up reading assessment score for student i in cohort c in grade-by-school cluster g; π_{0cg} is the intercept for cohort c in grade-by-school cluster g; $\pi_{1cg} \ldots \pi_{8cg}$ represent the coefficients for the student-level covariates, which include the standardized baseline MAP reading score, a series of race/ethnicity indicators (Asian, Black, white, and other race/ethnicity, with Hispanic students as the uncoded comparison group), female, and IEP and ELL status (1=yes, 0=no). The Level-1 error term for student i in cohort c in grade-by-school cluster g is represented by e_{icg}. At the cohort (treatment) level, β_{00g} is the intercept for grade-by-school cluster g, and *BookNook* indicates that cohort c was randomly assigned to the treatment. The Level-2 error term is represented by r_{0cg}. The model is unconditional at level 3, with γ_{000} indicating the grand mean and the Level-3 error term indicated by $u_{00g.}$

In a second robustness check, we modeled reading growth during the fall semester—prior to the implementation—as a function of assignment to a BookNook treatment cohort the following winter. A significant BookNook effect on student reading growth pre-implementation would suggest that treatment cohorts enrolled students who were different in some unmeasured ways compared to their control cohort peers. With these models, Winter MAP scores served as the outcome, while the Fall MAP score (baseline) was included as a covariate. All other model covariates and structures were identical. We ran both the two-level fixed-effects model and the three-level random-effects model described above. A small proportion of students (3.2%) from the analytic sample was missing the Fall MAP reading assessment score. Missingness was

distributed evenly across students who would subsequently be assigned to treatment and control cohorts.

Results

We begin with results from the models establishing pre-treatment equivalency between treatment/control cohorts (see Table 1). Fortunately, we find no statistically significant or substantively meaningful differences in terms of baseline student academic and socio-demographic characteristics. This increases our confidence that the impact estimates we discuss below stem from engagement with BookNook tutoring and not pre-existing differences between students who did and did not experience BookNook.

Implementation Fidelity

Although the extant literature suggests that in-school implementations of virtual tutoring might produce higher usage rates, we found considerable variability in BookNook usage among students in cohorts assigned to the treatment condition (see Table 2). Of the 959 students in treatment cohorts, 196 (20.4%) met the lower-bound threshold of recommended BookNook engagement, calculated as two completed sessions per week, for a total of 20 sessions. A plurality of students (45.2%) completed between 10 and 19 tutoring sessions in total, and 34.4% completed fewer than 10 sessions. Overall, treated students completed an average of thirteen total sessions during the implementation period. We explored the extent to which these usage rates were associated with other baseline student background characteristics. Low-usage students began the study with baseline MAP scores roughly 0.19 SDs below those of their moderate- ($p<.05$) and high-usage peers ($p<.10$). In other words, initially higher-achieving students engaged BookNook to somewhat higher degrees. However, we found no associations between BookNook usage rates and student race/ethnicity, sex, or IEP and ELL status.

One important question is the extent to which this variability in student usage flowed from the motivations and interests of individual students or from school staff. With the current implementation, the relevant adults were those staffing the Learning Labs, where treatment cohorts were to have received BookNook tutoring. One way to explore this question is to partition variance in usage into its within-Learning Lab and between-Learning Lab components. We found that almost half (45.5%) of the variability in usage rates exists across Learning Labs, with the remainder (54.5%) occurring within Learning Labs. This suggests that efforts to increase participation rates should target both students and staff. Clearly, some staff did not have appropriate expectations for student participation. However, even within the same Learning Labs, student participation rates varied substantially.

Impact Results

Table 3 provides estimates of the impact of BookNook on student reading growth. The Intent-to-Treat (ITT) estimates, displayed in the far left column, indicate that students assigned to treatment cohorts modestly outperformed their same-grade, same-school peers enrolled in control cohorts (ES = 0.052; $p<.05$). Recall that the ITT analytic approach does not account for actual BookNook usage, but instead considers only whether students were offered the treatment, in this case via membership in a cohort that was randomly assigned to the treatment condition. The Treatment-on-the-Treated (TOT) approach, however, allows us to explore the extent to which BookNook efficacy is associated with increased usage. The TOT results presented in the middle column indicate that students who completed 10 or more sessions also gained somewhat more reading skills compared to their control group peers (ES = 0.080; $p<.05$). Note that this estimate is slightly larger than the ITT estimate, though the two estimates are not significantly different from one another. However, we find a substantially larger effect for students who

complied with the recommended BookNook dosage of at least two completed sessions per week, for a total of 20 or more sessions. As displayed in the far-right column, these high-usage students outperformed their peers assigned to control cohorts by over one-quarter standard deviation (ES = 0.257; $p<.05$).

Treatment Heterogeneity and Robustness Checks

For our first robustness check, we reconstructed the Intent-to-Treat model discussed above as a three-level random effects model with random block effects and a random treatment effect (see Table 4). We find that the BookNook treatment estimate is virtually identical to that produced by the two-level fixed effect model from Table 3. For our second robustness check, we modeled reading development during the fall semester, the period immediately prior to BookNook implementation. Mirroring the Intent-to-Treat models of reading growth during the treatment period, we constructed both a two-level fixed-effects model and a three-level random-effects model in which winter MAP scores served as the outcome, and fall MAP scores were included as a covariate. As indicated in Table 5, the BookNook estimates are non-significant and quite close to zero. This suggests that treatment and control groups did not have differential effects on reading growth prior to the implementation, bolstering our confidence in the comparability of treatment and control cohorts and the reported treatment effects.

Finally, we explored whether the treatment effects reported above varied across student background characteristics. We leveraged the same two-level fixed-effects models and three-level random-effects models, but incorporated BookNook by student-level covariate interaction terms. As displayed in Table 6, we find no evidence that student characteristics moderated the BookNook impact. Rather, BookNook appeared to be equally effective across student academic and demographic subgroups.

Conclusion and Discussion

This study examined the implementation and efficacy of BookNook with first- through fourth-graders in six Rocketship schools in Northern California. We found evidence that BookNook tutoring supported student reading growth. Our ITT models suggest that students in cohorts assigned to receive BookNook virtual tutoring outperformed their control-group peers by roughly 0.05 SDs. Our TOT analyses indicate even larger positive effects among students with higher usage rates. Treatment students who completed 10 or more BookNook sessions experienced a reading skills advantage of 0.08 SDs, while those who completed 20 or more sessions—the recommended dosage—experienced a 0.26 SD developmental advantage. It is important to bear in mind the caveats associated with the TOT estimates. Namely, these effects are relevant for the types of students who completed more sessions. We cannot claim that providing all students the same levels of BookNook tutoring would have produced similar levels of reading performance.

The presence of these significant effects is somewhat surprising given the relatively short ten-week implementation period. Students and staff in these schools did not have prior experience working with the BookNook platform. We typically assume that new interventions take time for both students and staff to become comfortable with a given approach and its procedures. Bear in mind, however, that the virtual tutors did have considerable prior experience with the platform and that teacher involvement in actual instruction was minimal. This suggests that perhaps the BookNook platform is structured such that the start-up and launch costs we associate with many interventions are reduced. Importantly, the BookNook intervention incorporates many of the best practices found in the literature on high-impact tutoring: trained

tutors using structured materials with embedded formative assessment to monitor student progress and support tutor efficacy.

The recommended usage for BookNook is in line with definitions of high-dosage tutoring as students are expected to receive individualized tutoring for at least two sessions per week. In practice, many supplemental ed-tech implementations experience weak usage among students assigned to treatment conditions. This is indeed what we found with the current study, where few students received BookNook tutoring at the expected levels. Only 20% of students enrolled in treatment cohorts completed 20 or more tutoring sessions, the recommended dosage. This is particularly striking given that one of the best practices for increasing tutoring uptake is to conduct tutoring sessions during regular school hours (Nickow et al., 2020), a strategy already implemented by BookNook.

If virtual tutoring during the school day can match the uptake of in-person interventions, more work is needed to understand what factors lead to higher dosage. Schools that have already moved their tutoring programs to take place during the school day, but continue to struggle with dosage, are experimenting with appointing one person to manage tutoring and setting clearer expectations for tutoring implementation from the beginning of the year (White et al., 2021). BookNook and other developers might also consider deeper conversations with school staff, more meaningful professional development activities, and consistent and ongoing communications throughout the implementation period. With non-core instructional strategies such as BookNook, school staff will likely need to be convinced of the potential benefits for their students. Having a school staff member who is a “champion” of the program might help overcome some barriers to high-fidelity implementation (Makori et al., 2024). If these critical

issues of usage are not addressed, promising interventions such as BookNook are unlikely to fully achieve their aims of improving student academic outcomes.

Table 1. Cohort Baseline Equivalence (Cohort-Level Averages)

Characteristics	Treatment Cohorts (*n*=42)	Control Cohorts (*n*=35)	Difference (*SE*)
Baseline MAP	-0.034	0.039	-0.073 (0.050)
Asian	0.086	0.082	0.004 (0.009)
Black	0.099	0.086	0.013 (0.014)
Hispanic	0.783	0.796	-0.013 (0.018)
Other	0.008	0.016	-0.008 (0.006)
White	0.025	0.021	0.004 (0.008)
Female	0.483	0.479	0.004 (0.022)
IEP	0.086	0.106	-0.020 (0.017)
ELL	0.515	0.499	0.016 (0.024)

No differences significant at the $p<.10$ level. Baseline MAP scores are z-scored within grades.

Table 2. BookNook Usage Among Students in Treated Cohorts (*n*=959)

	Student-Level Usage Rates		
	Low: 0-9 Sessions (*n*=330)	Moderate: 10-19 sessions (*n*=433)	High: 20+ Sessions (*n*=196)
Sessions Completed	4.76	14.57***	22.89***
SD	(2.92)	(2.78)	(2.59)
Baseline MAP Score	-0.125	0.063*	0.070~
SD	(0.997)	(1.015)	(0.954)
Asian	8.5	9.7	13.3
Black	7.9	9.9	3.1
Hispanic	80.0	77.6	81.6
Other	0.3	0.5	0.4
White	3.3	2.3	1.5
Female	47.9	46.2	52.6
IEP	10.3	8.3	8.2
ELL	56.4	49.4	55.6

~ $p<0.10$; *$p<0.05$; **$p<.01$; ***$p<.001$. Sessions completed and MAP scores compared to low-usage category. Associations between usage and race/ethnicity, sex, IEP and ELL status are non-significant ($p>.05$).

Table 3. BookNook Effects on Student Reading Growth

	Intent-to-Treat	Treatment-on-the Treated	Treatment-on-the Treated
BookNook Cohort	0.052*	--	--
	(0.024)	--	--
BookNook: 10+ Sessions	--	0.080*	--
	--	(0.034)	--
BookNook: 20+ Sessions	--	--	0.257*
	--	--	(0.122)
Baseline MAP Score[1]	0.827***	0.826***	0.827***
	(0.170)	(0.017)	(0.016)
Asian[2]	0.111**	0.113**	0.115**
	(0.035)	(0.034)	(0.034)
Black	-0.020	-0.018	-0.009
	(0.045)	(0.045)	(0.045)
White	-0.130	-0.129	-0.127
	(0.098)	(0.095)	(0.097)
Other Race/Ethnicity	0.029	0.033	0.020
	(0.102)	(0.100)	(0.097)
Female	-0.025	-0.025	-0.029
	(0.025)	(0.025)	(0.026)
ELL	-0.148***	-0.146***	-0.148***
	(0.033)	(0.032)	(0.032)
IEP	-0.001	0.001	0.001
	(0.055)	(0.025)	(0.055)
Constant	-0.024	-0.036	-0.079
	(0.031)	(0.030)	(0.041)

*p<0.05; **p<.01; ***p<.001. Robust standard errors (indicated in parentheses) are clustered at the cohort level. All models include school-by-grade fixed effects.

[1] Outcome and baseline MAP scores are standardized (z-scored) within grades.

[2] All racial/ethnic groups compared to Hispanic students.

Table 4. Intent-to-Treat Estimates: Three-Level Random Effects Model

	Intent-to-Treat
BookNook Cohort	0.060*
	(0.028)
Baseline MAP Score[1]	0.828***
	(0.013)
Asian[2]	0.118
	(0.044)
Black	-0.030
	(0.048)
White	-0.121
	(0.079)
Other Race/Ethnicity	0.014
	(0.113)
Female	-0.021
	(0.024)
ELL	-0.148***
	(0.027)
IEP	0.002
	(0.042)
Constant	0.043
	(0.035)

*p<0.05; **p<.01; ***p<.001. Robust standard errors (indicated in parentheses) are clustered at the cohort level.

[1] Outcome and baseline MAP scores are standardized (z-scored) within grades.

[2] All racial/ethnic groups compared to Hispanic students.

Table 5. BookNook Effects on Pre-Implementation Reading Growth

	Two-Level Fixed Effects Model	Three-Level Random Effects Model
BookNook Cohort	-0.019	-0.012
	(0.027)	(0.033)
Baseline (Fall) MAP Score[1]	0.842***	0.849***
	(0.018)	(0.014)
Asian[2]	0.040	0.046
	(0.036)	(0.044)
Black	-0.080	-0.100
	(0.044)	(0.048)
White	0.101	0.099
	(0.089)	(0.081)
Other Race/Ethnicity	-0.212	-0.212
	(0.114)	(0.113)
Female	-0.021	-0.021
	(0.024)	(0.024)
ELL	-0.043	-0.039
	(0.032)	(0.028)
IEP	-0.107	-0.101
	(0.068)	(0.043)
Constant	0.014	0.050
	(0.029)	(0.030)

*p<0.05; **p<.01; ***p<.001. Robust standard errors (indicated in parentheses) are clustered at the cohort level.

[1] Outcome and baseline MAP scores are standardized (z-scored) within grades.

[2] All racial/ethnic groups compared to Hispanic students.

Table 6. Treatment by Student Background Characteristic Interactions

	Baseline MAP	Asian	Black	White	Other	Female	IEP	ELL
Fixed Effects Model								
Main BN Effect	0.052*	0.049~	0.052*	0.051*	0.051*	0.055	0.054*	0.072*
	(0.024)	(0.025)	(0.025)	(0.024)	(0.023)	(0.036)	(0.025)	(0.034)
Interaction Effect	0.003	0.034	-0.002	0.041	0.074	-0.007	-0.021	-0.039
	(0.029)	(0.071)	(0.070)	(0.183)	(0.227)	(0.050)	(0.109)	(0.056)
Random Effects Model								
Main BN Effect	0.060*	0.057*	0.059*	0.059*	0.059*	0.064~	0.060*	0.083*
	(0.028)	(0.029)	(0.029)	(0.028)	(0.028)	(0.036)	(0.029)	(0.037)
Interaction Effect	0.004	0.034	0.009	0.046	0.119	-0.009	-0.007	-0.045
	(0.024)	(0.086)	(0.086)	(0.157)	(0.235)	(0.047)	(0.081)	(0.048)

~p<.10; *p<0.05; **p<.01; ***p<.001. Model specifications are identical to those above. Robust standard errors (indicated in parentheses) are clustered at the cohort level. Outcome and baseline MAP scores are standardized (z-scored) within grades. All racial/ethnic groups compared to Hispanic students.

References

Angrist, J. D., & Pischke, J. S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher*, *13*(6), 4-16.

Burch, P., Good, A., & Heinrich, C. (2016). Improving access to, quality, and the effectiveness of digital tutoring in K–12 education. *Educational Evaluation and Policy Analysis*, *38*(1), 65-87.

Carlana, M., & La Ferrara, E. (2021). *Apart but connected: Online tutoring and student outcomes during the COVID-19 pandemic*. IZA Discussion Papers, No. 14094.

Cortes, K., Kortecamp, K., Loeb, S., & Robinson, C. (2024). *A Scalable Approach to High-Impact Tutoring for Young Readers: Results of a Randomized Controlled Trial* (No. w32039). National Bureau of Economic Research.

Dietrichson, J., Bøg, M., Filges, T., & Klint Jørgensen, A.-M. (2017). Academic interventions for elementary and middle school students with low socioeconomic status: A systematic review and meta-analysis. *Review of Educational Research, 87*(2), 243-282.

Fahle, E. M., Kane, T. J., Patterson, T., Reardon, S. F., Staiger, D. O., & Stuart, E. A. (2023). School district and community factors associated with learning loss during the COVID-19 pandemic. Center for Education Policy Research at Harvard University: Cambridge, MA.

Fryer Jr, R. G. (2016). The production of human capital in developed countries: Evidence from 196 randomized field experiments. In *Handbook of economic field experiments* (Vol. 2, pp. 95-322). North-Holland.

Gersten, R., Rolfhus, E., Clarke, B., Decker, L. E., Wilkins, C., & Dimino, J. (2015). Intervention for first graders with limited number knowledge: Large-scale replication of a randomized controlled trial. *American Educational Research Journal, 52*(3), 516-546.

Glennerster, R., & Takavarasha, K. (2013). *Running randomized evaluations: A practical guide*. Princeton University Press.

Guryan, J., & Ludwig, J. (2023). *Overcoming Pandemic-Induced Learning Loss*. Aspen Institute.

Guryan, J., Ludwig, J., Bhatt, M. P., Cook, P. J., Davis, J. M., Dodge, K., ... & Stoddard, G. (2023). Not too late: Improving academic outcomes among adolescents. *American Economic Review*, *113*(3), 738-765.

Jacob, R., Armstrong, C., Bowden, A. B., & Pan, Y. (2016). Leveraging volunteers: An experimental evaluation of a tutoring program for struggling readers. *Journal of Research on Educational Effectiveness, 9*(sup1), 67-92.

Kraft, M. A., & Falken, G. T. (2021). A blueprint for scaling tutoring and mentoring across public schools. *AERA Open*, *7*(1), 1-21.

Kraft, M. A., List, J. A., Livingston, J. A., & Sadoff, S. (2022). Online tutoring by college volunteers: Experimental evidence from a pilot program. *AEA Papers and Proceedings, 112*(May), 614-618.

Loeb, S., Novicoff, S., Pollard, C., Robinson, C., & White, S. (2023). *The effects of virtual tutoring on young readers: Results from a randomized controlled trial.* National Student Support Accelerator.

Makori, A., Burch, P., & Loeb, S. (2024). Scaling high-impact tutoring: School level perspectives on implementation challenges and strategies. (EdWorkingPaper: 24-923).

Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/h8z5-t461

Markovitz, C. E., Hernandez, M. W., Hedberg, E. C., & Silberglitt, B. (2014). *Impact Evaluation of the Minnesota Reading Corps K-3 Program*. Corporation for National and Community Service.

McCormick, R., Woo, J., Steiner, B., & Grossman, J. (2023). *How a pilot program targeting ninth-graders led to shifting sessions from weekends and evenings to regular school hour*s. MDRC.

NCES (2023, October). School pulse panel: Responses to the pandemic and efforts toward recovery. https://nces.ed.gov/surveys/spp/results.asp

NWEA. (2019). MAP Growth technical report. Portland, OR.

Neitzel, A. J., Lake, C., Pellegrini, M., & Slavin, R. E. (2022). A synthesis of quantitative research on programs for struggling readers in elementary schools. *Reading Research Quarterly*, *57*(1), 149-179.

Nickow, A., Oreopoulos, P. & Quan, V. (2020). *The impressive effects of tutoring on preK-12 learning: A systematic review and meta-analysis of the experimental evidence*. NBER Working Paper 27476

Peters, S. J., Langi, M., Kuhfeld, M., & Lewis, K. (2023). *Unequal learning loss: How the COVID-19 pandemic influenced the academic growth of learners at the tails of the achievement distribution.* Annenberg Ed. Working Paper No. 23-787.

Robinson, C. D., & Loeb, S. (2021). High-impact tutoring: State of the research and priorities for future learning. *National Student Support Accelerator*.

Robinson, C., Kraft, M., Loeb, S., & Schueler, B. (2021) *Design principles for accelerating*

student learning with high-impact tutoring. EdResearch for Action.

Schueler, B. E., & Rodriguez-Segura, D. (2023). A cautionary tale of tutoring hard-to-reach students in Kenya. *Journal of Research on Educational Effectiveness*, *16*(3), 442-472.

White, S., Carey, M., O'Donnell, A., & Loeb, S. (2021). Early Lessons from Implementing High-Impact Tutoring at Scale. *National Student Support Accelerator*.