Published Citation:
Robinson, C. D., Pollard, C., Novicoff, S., White, S., & Loeb, S. (2024). The Effects of Virtual Tutoring on Young Readers: Results from a Randomized Controlled Trial. *Educational Evaluation and Policy Analysis*. doi: 10.3102/01623737241288845

The Effects of Virtual Tutoring on Young Readers:

Results from a Randomized Controlled Trial

Carly D. Robinson[1], Cynthia Pollard, Sarah Novicoff, Sara White, Susanna Loeb

Abstract

In-person tutoring has been shown to improve academic achievement. Fewer studies have examined the impact of virtual tutoring and have focused on older students. We present findings from the first randomized controlled trial of virtual tutoring for young children. Students in grades K-2 were assigned to 1:1 tutoring, 2:1 tutoring, or a control group. Virtual tutoring increased early literacy skills by 0.05-SD for all students and 0.08-SD for a sample excluding English learners and students with disabilities (i.e., students not eligible for additional support services). One-on-one tutoring tended to produce larger gains, especially for students initially scoring well below benchmark (0.15-SD). Effects are smaller than typically seen from in-person early literacy tutoring programs but still positive and statistically significant.

[1] Corresponding author: carly_robinson@stanford.edu

Acknowledgements: We are grateful to our partners in this research, On Your Mark and Uplift Education. We especially thank Nick Erber, Mindy Sjoblom, Ashley Chin Morefield, and Aaron Schlessman for their invaluable contribution to project implementation. We thank the Overdeck Family Foundation for their generous support of this research. We also thank the Smith Richardson Foundation and Arnold Ventures/Accelerate for their support of our full research program. We received insightful feedback and support from the National Student Support Accelerator team, in particular Kathy Bendheim, Kristine Gaffaney, Monica Lee, Nancy Waymack, and Helen Zhou.

Introduction

Learning to read presents a challenge for many students in the United States. In the 30 years that the National Assessment of Educational Progress has included a reading test for fourth graders, no more than 37 percent of students have ever scored proficient; in other words, for at least 30 years, more than 60 percent of U.S. students have struggled to learn to read (*NAEP Reading*, 2022). From 2019 to 2022 specifically, the percentage of students scoring proficient in fourth grade reading declined by two percentage points.

In addition to negatively affecting students' reading skills, the COVID-19 pandemic brought tutoring and virtual learning to the forefront of educational practice. Tutoring emerged as one of the most evidence-based and promising interventions for accelerating student learning (Robinson & Loeb, 2021), although implementing it effectively and at scale can be challenging for schools (Groom-Thomas et al., 2023). From a logistical standpoint, virtual tutoring may cost schools less money and reduce staffing burdens. Moreover, virtual tutoring programs were specifically mentioned as one of the allowable expenses under the Elementary and Secondary School Emergency Relief (ESSER) Fund (U.S. Department of Education, 2021). Thus, virtual tutoring has become an increasingly attractive option for district leaders and policymakers.

Despite the resulting influx of virtual tutoring offerings, few rigorous research studies have tested whether virtual tutoring can meaningfully improve student learning–particularly in the early grades. Providing tutoring in small groups can also be an appealing option when tutors are in short supply. However, to our knowledge, no large-scale research has directly tested the marginal impact of increasing tutoring group size on student reading achievement, which has important implications for how to efficiently direct resources. As tutoring for elementary students becomes an increasingly popular intervention with 48% of elementary schools in the

U.S. reporting in October 2023's School Pulse Survey that they have high-dosage tutoring programs (*School Pulse Survey*, 2023), it is critical to understand how to effectively meet demand and cut costs while maintaining efficacy.

In this paper, we present results from a randomized controlled trial evaluating OnYourMark, a virtual reading tutoring program for early elementary students. We randomly assigned 2,085 students to receive 1:1 tutoring, 2:1 tutoring, or a control group during the 2022-23 school year. We examine results from this experimental research design across two samples: a full sample of kindergarten through second grade students, and a restricted sample that excludes English learners and students with disabilities who were more likely to be withdrawn from tutoring due to receiving other support services and also experience a different counterfactual condition (i.e., for whom the control group also receives extensive support services in the acquisition of early literacy skills). We find positive and statistically significant effects of OnYourMark on students' end-of-year reading scores in both samples (0.05 SD and 0.08 SD respectively), with stronger effects from 1:1 tutoring (0.12 SD in our restricted sample). Additionally, we find that 1:1 tutoring most benefited first graders and students who performed below grade level on their beginning-of-year reading scores (0.20 and 0.15 SD, respectively).

Background

As the COVID-19 pandemic closed schools, widened achievement gaps, and led to the allocation of additional federal funding for academic recovery, tutoring expanded rapidly. Research studies provide evidence that tutoring is a promising practice for improving academic achievement generally (Dietrichson et al., 2017; Nickow et al., 2024) and in early literacy specifically (Elbaum et al., 2000; Gersten et al., 2020; Neitzel et al., 2022), with effect sizes

ranging from 0.24 to 0.41 SD. Several meta-analyses provide insights into the program features that may be more (or less) effective. For instance, teacher-led tutoring tends to produce the largest academic achievement gains (Nickow et al., 2024), though volunteers often still have positive effects (Ritter et al., 2009). The same pattern holds in the early literacy context (Gersten et al., 2020; Slavin et al., 2011). The meta-analyses also suggest that group size may play a role in program effectiveness. Specifically, programs that deliver 1:1 tutoring tend to produce larger achievement gains than those emphasizing working in small groups (Nickow et al., 2024). Similarly, in the early literacy context, 1:1 tutoring has generally been shown to be more effective than other interventions (Gersten et al., 2020; Neitzel et al., 2022), though researchers have not tested differences in group size explicitly and rigorously (e.g., in a preregistered randomized controlled trial). Conversely, some research suggests that the benefit of 1:1 instruction may not hold, in part because students may benefit from additional opportunities to hear and practice language among their peers (Miles et al., 2022; Richards-Tutor et al., 2016).

Among the programs that began or expanded during and after the COVID-19 pandemic, many tutoring programs were virtual. On one hand, these programs often offer the same affordances of in-person tutoring with a small-group format providing opportunities for immediate individualized attention and feedback addressing gaps in understanding. District leaders who want to implement a high-impact tutoring program to support student learning may face resource constraints and logistical challenges, and virtual tutoring often offers services at lower costs and in communities where staffing tutoring can be challenging. By delivering tutoring virtually, schools are not constrained by the tutor labor supply in their communities. A qualitative investigation found that early adopters of tutoring as a pandemic recovery strategy faced a shortage of high-quality tutors (White et al., 2021). Virtual tutoring platforms can recruit

tutors across the country, or even internationally, widening the pool of qualified tutors available to students (Kraft & Falken, 2021).

On the other hand, though the efficacy of in-person tutoring has been well-established, rigorous causal research on online tutoring programs is sparse and focuses on older students.[1] Hybrid programs, in which students work through an online material facilitated by a tutor in the room, offer some relevant comparison. In one program, children worked on the computer while a paraprofessional tutor walked around the small group (up to 6:1) and supported students when flagged to do so by the computer program; though the effect size was large for students who attended sessions regularly, the overall effect size was small and not significant (Wang et al., 2023). In another intervention, classroom teachers worked with a student in person and 1:1 while a literacy coach observed the sessions live on video and assisted; this program showed large effects on nonsense word fluency, letter-word identification, passage comprehension, and spelling (Amendum et al., 2011).

Fully virtual tutoring remains understudied, and many might wonder if the lack of in-person interaction would affect efficacy through diminished attention to a screen. The limited research base on online tutoring for secondary students shows consistent positive effects. A randomized control trial of online video call-based tutoring delivered by volunteer university students to students in Italy in grades 6-8 produced an effect size of 0.26 SD on multiple subjects (Carlana & Ferrara, 2021). A similar video call-based tutoring intervention with teachers in Spain produced an effect size of 0.26 SD in math (Gortazar et al., 2023). A recent randomized controlled trial in the United States, in which volunteers worked online with high school students, showed an effect size of 0.23 SD on math (Deacon & Chojnacki, 2023). Another

randomized control trial testing online tutoring for middle school students by volunteers found small positive, but not statistically significant, effects (Kraft et al., 2022).

Online tutoring for elementary students has been implemented and evaluated even less than those for older students and predominantly in math, whose effect sizes may not generalize to interventions in other subjects. We identified only two randomized controlled trials of online tutoring for elementary students in developed countries, and both targeted math (not language skills). In one online tutoring intervention by teachers of fifth graders in the United States and Canada, research showed a 0.41 SD effect size on math achievement, though the measure used was not a standardized test and covered only targeted content emphasized in the tutoring (Roschelle et al., 2020). In another study in which tutors from Sri Lanka and India virtually supported year 6 students (ages 10-11) in the United Kingdom in math, the intervention produced a not statistically significant 0.002 SD change in math test scores and a 0.01 SD change in English (Torgerson et al., 2016).

In this paper, we present the first rigorous causal evidence of the efficacy of online literacy tutoring for young students. Specifically, we use a randomized controlled trial of approximately 2,000 students in kindergarten through second grade to examine the effects of OnYourMark, a fully virtual reading tutoring program integrated into the school day. Our examination of an in-school, virtual early literacy tutoring program makes several contributions beyond the scope of the existing studies of online tutoring interventions.

First, we provide some of the first causal evidence that virtual tutoring can work with young students. Existing research has focused on older elementary students, whose ability to learn online may exceed that of early elementary students. Specifically, younger students often struggle with self-regulation and executive functioning (Howard & Vasseleu, 2020), which are

likely key components of online learning (Cho & Shen, 2013; Harel-Gadassi, 2022). Many of the previously studied interventions for virtual tutoring for younger students also focused on math, when research shows math and literacy interventions in the early grades can have different effects (Nickow et al., 2024). Moreover, none of the virtual tutoring studies had frequent enough sessions to be classified as an examination of high-impact tutoring. In Roschelle et al. (2020), students received two 20-25 minute sessions per week while in Torgerson et al (2016), students received one 45-minute session per week. High-impact tutoring typically includes three or more 30–45-minute sessions (*High-Impact Tutoring*, 2021).

Second, this study explores the impact of tutoring when it is embedded into the school day as part of the student learning experience. Much of the research to date has studied the impact of opt-in tutoring programs. However, these programs often have low and differential take-up which means that the results might not generalize to the broader population of students. In a study of a large charter system in the United States where tutoring was optional and took place mostly during out-of-school hours, only 19% of middle and high school students ever used the platform, and those students tended to be higher achieving (Robinson et al., 2022). In the six studies of online tutoring described above, the four with the largest effect sizes (Carlana & Ferrara, 2021; Deacon & Chojnacki, 2023; Gortazar et al., 2023; Roschelle et al., 2020) were programs that took place after or outside of the school day. Though the study designs rely on randomization, they randomize among a group of students who are motivated or able to take up the program during those outside-of-school hours and their effects may not generalize to programming embedded within the school day for a more representative group of students.

Finally, we leverage multiple treatment arms to show whether group size (i.e., 1:1 tutoring as compared to 2:1 tutoring) affects program efficacy. A meta-analysis of reading and

math small group interventions found only four studies that quasi-experimentally or experimentally varied group size, too few studies from which to draw conclusions about group size efficacy, and demonstrating another major contribution of our study into the literature (Dietrichson et al, 2021). One recent small study by Miles and colleagues (2022) did randomly vary whether students receive 1:1 or small group reading instruction and found no consistent differences on student outcomes. However, large meta-analyses suggest that 1:1 instruction is associated with larger effect sizes than small group instruction (Nietzel et al., 2022; Nickow et al., 2024). Compared to the existing studies in the literature, the present study has two distinct advantages. First, we randomly assign students to a control group, in addition to 1:1 and 2:1 tutoring. Second, we have a larger sample size and therefore are better powered to detect small, but meaningful effects. As school districts look for ways to support more students in the face of declining federal funding, understanding the marginal effect of increasing the student-tutor ratio has important implications for cost and scale.

Program Context

In 2021, OnYourMark Education began offering virtual early literacy tutoring grounded in the science of reading. OnYourMark partners with school districts to deliver tutoring to early elementary school students. In addition to recruiting and hiring tutors, the organization provides the initial tutoring training and ongoing professional development focused on content knowledge, building relationships, and effective delivery of the intervention. The curriculum includes a focus on phonics, phonological awareness, and fluency and is delivered in 1:1 or 2:1 sessions embedded into the school day. The program aims to promote positive tutor-student relationships through a small student-tutor ratio and by pairing students with a consistent tutor

for the duration of the program. In its initial launch in fall 2021, OnYourMark served 58 students in one school and, by spring 2022, had expanded to serve 180 students across two states in two schools.

During the 2022-23 school year, a charter management organization in Texas partnered with OnYourMark to provide early literacy tutoring to kindergarteners, first graders, and second graders in 12 of its elementary schools. Students met their tutor online for 20 minutes during the school day, four times per week. Tutoring rolled out in September and continued through May.

OnYourMark hires former classroom teachers, part-time teachers, and college students as tutors, among others. OnYourMark does not require that tutors have a postsecondary degree or prior education-related experience. Table S1 in Appendix A presents summary statistics for tutor demographic characteristics, though not all tutors answered these questions. Among OnYourMark tutors that responded to the survey, 47 percent were former teachers, while 70 percent had graduated college.

During the 2022-23 school year, OnYourMark charged schools about $1400 per student to provide tutoring services; this price includes salaries for tutors as well as indirect costs for OnYourMark to hire and train tutors and provide technical support during tutoring sessions but does not include the cost of the complete set of inputs needed to make tutoring work (like facilities and technology). OnYourMark lowers costs by relying on schools to utilize technology they already have, such as tablets and headphones, and using many non-college-educated tutors whose wage premium is lower. However, by paying their tutors at all, OnYourMark has a higher cost to the school than that of volunteer-led programming; for example, ReadingPartners charges schools $710 for their volunteer-led tutoring (Jacob et al, 2016).

Study Design

We conducted our randomization in partnership with OnYourMark and the 12 partner schools. In each of the 12 schools, building-level staff first selected approximately 10 students who would be guaranteed to receive tutoring to ensure that students with the highest need were offered the service and to mitigate school staff concerns about randomization. In total, 121 students in the 12 schools enrolled in tutoring outside of the randomized sample. We exclude them from the study sample and from all analyses.

After identifying and reserving tutoring spots for high-need students, school-site staff selected double the number of students as they had remaining seats to potentially receive tutoring in the randomized study. Seven schools had 96 tutoring slots, three schools had 112 tutoring slots, one school had 128 slots, and one school had 64 slots. In a school allotted 96 tutoring seats, staff first selected 10 students guaranteed to receive tutoring and then indicated eligibility and availability for another 172 students to fill the remaining 86 tutoring seats. School leaders were guided to use student beginning-of-year Dynamic Indicators of Basic Literacy Skills (DIBELS) performance to identify which students to select as being eligible for tutoring. Specifically, OYM guided leaders to target students who were performing below benchmark. However, the individuals at the school ultimately had leeway to select which students were eligible to receive OYM tutoring. Schools indicated the grade level and available free class periods for these students, and researchers used grade level and availability as blocks in a stratified randomization. Ultimately, our sample included 34 school x grade strata that ranged in size from 32 to 98 students.

OnYourMark also used students' beginning-of-year DIBELS performance to identify areas of need and group *all* eligible students into pairs. Students could be paired with one another

if they were enrolled in the same school, available during the same period, and had at least one overlapping "target area" (or low score on the same DIBELS sub-test) to work on based on the DIBELS beginning-of-year assessment. English learner status and disability status were not considered. We created pairs prior to randomization to ensure that students could be further randomized for the group size study to receive tutoring either 1:1 or 2:1 tutoring.

After receiving the information on pairs from the schools and OYM, we conducted a block-cluster randomization to assign eligible student pairs as clusters into either the treatment or control group. This randomization process ensured that all tutoring spots were filled by students who could attend in the relevant class period and that the treatment and control groups were balanced by grade level and initial performance. Once all tutoring spots were filled, researchers further used randomization to divide treatment students into either receiving 1:1 or 2:1 tutoring. Pairs assigned to receive 1:1 tutoring each received tutoring alone. Pairs assigned to receive 2:1 tutoring received tutoring together. Once randomization was complete, students in tutoring slots were randomly assigned to tutors. Thus, tutors were equally likely to instruct students in groups of 1:1 or 2:1 in each period. Figure 1 illustrates this process.

Our sample, as a result of this process, consists of 2,085 students, of which 510 students received 1:1 tutoring, 570 students received 2:1 tutoring, and the remaining 1,005 students continued business-as-usual (BAU). We placed students in the control group on a randomly ordered waitlist to replace any treatment students who attrited from the group; even if students moved off the waitlist and thus received tutoring, they remained in the analytic sample and are still considered control students in our intent-to-treat analyses. If a student was assigned to 2:1 tutoring and was withdrawn, a student on the waitlist would take their place in their tutoring

session. Thus, students assigned to 2:1 sessions should have maintained their group size throughout the duration of the study.

The school indicated at least two time periods in a school day during which *all* eligible students could potentially receive OnYourMark tutoring. These time periods could align with regular instruction in the classroom or an intervention block. Therefore, what constituted BAU for a given student depended on both the school and the time of the scheduled tutoring.

Federal law mandates that students with disabilities and English learners are entitled to more intensive support, and this requirement (as well as subsequent associated scheduling changes) often resulted in these students being withdrawn from tutoring due to a conflict in when those required services could be received. These withdrawals are non-random and present challenges to the internal validity of our study. English learners and students with disabilities in the Control group received additional services (and continued to not receive tutoring) whereas these students in the Treatment group also received those additional services and were more likely to be taken out of the tutoring program. As a result, these groups of students were more likely to receive the same services no matter their assigned condition, thus weakening the treatment-control contrast. In our preregistered analysis plan (https://osf.io/pq4g6), we outlined that "If attrition appears to be equal across conditions after tutoring started, we will exclude students who could no longer receive tutoring services at the school from our analysis." Attrition (i.e., missingness of the outcome variable) is even across Treatment and Control groups as shown in Table 3 and discussed in more depth below. Furthermore, our exclusion criteria section also noted that we would retain students in our primary sample if students assigned to the Treatment group exited tutoring services and "we cannot identify the analog students in the Control group." In this case, we *can* identify students with disabilities and English learners in

both the Treatment and Control group. Therefore, we present our main findings on the effects of tutoring and group size for both the full sample of students randomized, and then for our preferred sample excluding all 731 English learners and students with disabilities. This preferred sample construction allows us to account for this group's disproportionate lack of tutoring in the treatment group due to enrollment in other services and disproportionately large frequency to receive support services even if in the control group. We weigh the implications and limitations of this approach in our discussion.

Data

Using OnYourMark data and school administrative data, we developed a student-level dataset consisting of the following variables: grade, date of birth, their race/ethnicity, gender, whether they received free or reduced price lunch or were otherwise indicated as economically disadvantaged based on the receipt of other public assistance, whether they had an Individualized Education Plan or 504 Plan, whether they were designated as an English learner, and their availability for tutoring within the school day. For any covariates that had "missing" data, as specified in our pre-analysis plan, we created a vector variable that included a "missing" category so students were not dropped from the analysis, as suggested by Zhao and Ding (2024). As discussed above, we utilized information on scheduling availability to ensure that all tutoring spots were filled by students who were free at the requisite time. For students assigned to tutoring, OnYourMark provided administrative data including the name of the tutor, tutor demographics, and tutoring attendance.

Our primary outcome of interest is student composite scores on the end-of-year Dynamic Indicators of Basic Literacy Skills (DIBELS), 8th edition.[2] DIBELS is a widely-used and

extensively validated set of measures and procedures to assess the acquisition of literacy skills (Smolkowski & Cummings, 2016). The components of the DIBELS are closely aligned to the early literacy skills targeted by the OnYourMark intervention. Students were assessed by their classroom teachers on DIBELS at the beginning (BOY), middle (MOY) and end (EOY) of the academic year. Classroom teachers necessarily knew what interventions their students were receiving and, therefore, are not blind to treatment status. However, there is little reason to expect that classroom teachers would be systematically biased in their administration of the assessment. Specifically, an Institute of Education Sciences examination of DIBELS found an inter-rater reliability above 95 percent for all sub-tests except words read correctly, which had an inter-rater reliability of 65-98 percent (Chernoff et al., 2021).

The DIBELS composite score is made up of a series of subtests that are typically 60-second, individually administered assessments that measure specific literacy subskills (e.g., letter sounds, decoding, reading fluency). DIBELS sections are purposefully ordered to test specific language skills as they develop; as a result, subtests vary by grade level. This variation is described in more detail in Appendix B.

For the purpose of this research study, we examine the impact of treatment on standardized EOY DIBELS composite scores as well as raw subtest scores. The composite score was standardized within each grade using the mean and standard deviation of the control group prior to any sample exclusions. We only report the effect of treatment on subtest scores if said subtest was taken by all students in the grade in our analytic sample. We do not examine the impact of treatment on subtests taken only by students who scored below or above a cut score, as the sample could be affected by our treatment and bias our results. As a result, we report results on five kindergarten skills – letter naming, phoneme segmentation, identification of correct letter

sounds, decoding, and word reading – and five first-grade skills – identification of correct letter sounds, decoding, word reading, passage reading, and reading accuracy. In second grade, we report results on three skills – passage reading, reading accuracy, and comprehension.

DIBELS subtests have moderate to strong predictive validity compared to other assessments of reading ability with the predictive power increasing as the tests become more challenging over time. Specifically, later DIBELS skills like passage reading and reading accuracy are more predictive of performance on the Iowa Assessment than earlier tests like letter naming; the correlation coefficient for the passage reading skills (from the DIBELS Oral Reading Fluency subtest) is 0.82 in the spring of grade-1 (University of Oregon, 2018).

In addition to EOY DIBELS, we also consider MAP Reading scores as an exploratory outcome measure. The MAP Reading Fluency assessment is a 20-minute online adaptive assessment designed for universal screening and progress monitoring of literacy skills for students in grades PK-5, and was administered to our study sample at the end of the school year in addition to DIBELS. MAP Reading Fluency has a strong focus on reading fluency, comprehension, and foundational reading skills (NWEA, 2023). We did not receive data on specific literacy skills on MAP and only report data on the overall standardized score earned by students at the end of the year. Like our primary DIBELS outcome, the end-of-year score was standardized within each grade using the mean and standard deviation of the control group.

Methods

We preregistered our study design, hypotheses, and analytic plan on the Open Science Framework prior to conducting the primary analysis (https://osf.io/pq4g6). We use the following model to evaluate the impact of receiving tutoring on student outcomes:

$$Y_{ijk} = \alpha + \beta_1 Treatment_i + \gamma X_i + \omega_k + \epsilon_{ij}$$

where Y_{ijk} is the outcome of interest for student i with tutor j in school by grade k. $Treatment_i$ is the indicator for student assignment to receive OYM tutoring. X_i is a vector of student-level covariates, including demographics. X_i also includes standardized DIBELS BOY score, the associated squared term, and an indicator for whether students received the minimum score in their respective grades to reflect that minimum score students may disproportionately draw from students without prior formal educational experiences and thus may be different than their peers. ω_k is a fixed effect for strata (school x grade). ϵ_{ij} is a residual clustered at the pair level to account for the nesting of students within tutor groups. The p-values for our intent-to-treat (ITT) analysis noted in the manuscript are derived from Fisher Randomization Tests (Athey & Imbens, 2016).

We used a similar regression model to evaluate the impact of receiving 1:1 tutoring and 2:1 tutoring relative to the control group, where $Treatment_i$ is a categorical variable for assignment to BAU control, 1:1 tutoring, or 2:1 tutoring (0, 1, 2, respectively).

In addition to our ITT analysis, we conducted a Treatment-on-the-Treated (TOT) analysis in which we examined the effect of taking up the tutoring and address compliance challenges as students were removed from our waitlist. Specifically, we take an instrumental variables approach (2SLS) using the exogenous assignment to tutoring as an instrument for ever receiving tutoring. The analysis estimates the impact of ever receiving tutoring on student end-of-year DIBELS composite scores.

Finally, we conducted exploratory analyses that examined heterogeneity of the effect of OnYourMark tutoring for students by baseline reading performance on DIBELS and by grade

level. The by-grade analyses include examination of OnYourMark tutoring on individual specific subskills measured by DIBELS subtests.

Results

Descriptive Statistics

To confirm that the randomization process produced groups of students with similar characteristics, we test for differences in observable characteristics across the three conditions and report the P-values from two-way analysis of variance (ANOVA) tests in Table 1.[3]

Panel A of Table 1 shows that students were similar across study conditions in demographic characteristics and on their baseline DIBELS scores. Panel B shows that when English learners and students with disabilities are excluded from the analytic sample of students, the demographics and baseline scores of the students remain similar across conditions. We find no evidence of differences across groups at an alpha level of 0.05. Though the difference is not significant, we note that the Control group does have slightly lower initial DIBELS scores and we therefore control for this variable in all of our analyses.

Table 2 displays information on student withdrawal from program eligibility and being switched from the waitlist control group to treatment. The most frequent reasons for withdrawal included transferring out of the school (n=80) and accommodations for separate support needs for English learners and students with disabilities (n=61). Students were equally likely to be withdrawn in the 1:1 tutoring condition (11%) as compared to the 2:1 tutoring condition (13%), $\chi^2(1) = 0.676$, $p = .411$. All eligible students, not just those assigned to tutoring, were supposed to be withdrawn from tutoring eligibility if other circumstances arose that would impact their

ability to receive tutoring. However, in practice students in the Control group were much less likely to be officially withdrawn from tutoring eligibility (n=15).

The first three columns of Table 2 also show that treatment students (compared to control students) were more likely to be withdrawn from eligibility, as were English learners, students with disabilities, and, to a lesser extent, male students (compared to female students). The tutoring administrators were more likely to record withdrawals for students enrolled in tutoring than for students who were in the control group, at least in part because tutoring conflicted with the other services the withdrawn students were receiving while control students could receive those additional services as part of BAU. As discussed above, English learners and students with disabilities were often withdrawn as they were entitled to other academic supports, and our analysis leverages samples that include and exclude these students. Column 4 shows that, among students who were assigned to tutoring, the likelihood English learners and students with disabilities withdrew was even larger compared to native English speakers (6-pp) and students without disabilities (26-pp). Column 5 shows that students switched from the waitlisted control group into tutoring had slightly lower beginning-of-year DIBELS scores than students overall. These switched students remain in our control group for our intent-to-treat (ITT) analysis. Our ITT estimates likely therefore provide a lower bound of the treatment effect as some students in the control group did receive OnYourMark tutoring despite their initial assignment.

We detail attrition from the study in Table 3. We see that missingness rates of end-of-year DIBELS and MAP scores are similar across the treatment and control groups, though Asian American, Black and Latina/o/x students were slightly more likely to be missing EOY MAP scores.

Results on Overall Literacy

Table 4 presents our primary results, including our preregistered methodology. In the overall sample of students with end-of-year DIBELS scores, with no controls, students assigned to receive OnYourMark tutoring performed 0.10 SD (p=0.03) higher than students assigned to the BAU control group. We present several alternative models, all showing positive effects of OnYourMark. When controlling for baseline reading performance and student demographic characteristics, students assigned to receive OnYourMark tutoring performed 0.05 SD (p=.07) higher on their end-of-year DIBELS scores than students assigned to the BAU control group.

Within our preferred sample with covariates, excluding English learners and students with disabilities due to their differential rates of eligibility for tutoring as discussed above, we find that students assigned to OnYourMark tutoring performed 0.077 SD higher (p=.041) on the end-of-year DIBELS assessment including all prior controls. This estimate is slightly larger in magnitude relative to our initial sample containing these students. This estimated effect translates to OnYourMark students performing 1.64 percentile points higher than students assigned to the BAU control group (see Figure 2).

Table 5 displays results of analyses estimating the effects of OnYourMark tutoring on end-of-year MAP Reading Fluency scores, which is an exploratory outcome. Despite finding a positive effect on DIBELS scores, we find no effect on MAP scores regardless of sample definition. This difference may be because MAP and DIBELS test different sub-skills; for example, both test decoding but MAP also tests print concepts (e.g., point to the letter on your screen) while DIBELS does not. In addition, DIBELS tests a fixed set of skills based on the point in a child's development whereas MAP administered one K-2 adaptive test that tests whatever skills it thinks a student is ready for, regardless of grade level. Finally, DIBELS is administered

individually whereas MAP is typically taken by the whole class simultaneously. These differences in test format and content can lead to different effects.

Results by Group Size

Table 6 shows results for analyses estimating the effect of 1:1 tutoring and 2:1 tutoring relative to the control group on both EOY DIBELS and EOY MAP scores. Across the board, estimates of the effects of 1:1 tutoring are higher than estimates of the effects of 2:1 tutoring. In our preferred model controlling for baseline reading achievement and student demographics and excluding English learners and students with disabilities, we find that students assigned to 1:1 tutoring performed 0.12 SD (p=.017) higher than students assigned to the BAU control group. The estimated benefit of 2:1 tutoring relative to the BAU control group was smaller, with students who received 2:1 tutoring scoring only 0.04 SD higher, a difference that is not statistically different from zero (p=.382). Figure 3 illustrates these results using percentile scores.

Examining results on MAP Reading Fluency scores, we find that the estimated effect of tutoring relative to the control group was greater for students assigned to 1:1 than to 2:1 (0.061 SD vs. 0.03 SD), though neither of these estimated effects is statistically different from zero.

Treatment-on-the-Treated Analyses

Table 7 presents the results of treatment-on-the treated (TOT) analysis examining effects on students who were "ever tutored".[4] The first three columns show the TOT analysis for the full sample. Columns 4-6 of Table 7 show the TOT analysis for the sample excluding English learners and students with disabilities. Panel B shows the first stage, in which we show the probability of ever being tutored based on assignment to treatment. In the full sample, any

compliance issues are largely due to students being withdrawn from the treatment group because they were eligible for other services and students transferring from the control condition to the treatment group. In the sample excluding English learners and students with disabilities, this analysis almost exclusively captures students switching from the randomly ordered waitlist control to receive tutoring.

As expected, the effect of ever receiving tutoring is greater than the effect of intent to treat in both the full sample and our preferred sample (0.072 SD vs. 0.054 SD and 0.104 SD vs. 0.075 SD, respectively). In our preferred sample excluding English Learners and students with disabilities, when comparing tutored students assigned to 1:1 tutoring to non-tutored control students, the treatment-on-the-treated estimate of tutoring is a statistically significant effect of 0.16 SD (p=.017). When comparing tutored students assigned to 2:1 tutoring to non-tutored control students, the treatment-on-the-treated is smaller and not statistically significant with an effect size of 0.06 SD (p=0.303).

Heterogeneity Analysis

The effect of OnYourMark tutoring differed somewhat for students depending on their initial reading performance level, as shown in Table 8. We used the cut-off scores set by DIBELS developers to classify students into four groups based on their scores on beginning-of-year DIBELS tests: “Red, Intensive Support/Well Below Benchmark,” “Yellow, Strategic Support/Below Benchmark,” “Green, Core Support/At Benchmark,” and “Blue, Core Support/Above Benchmark.” We find positive and statistically significant effects (0.11 SD, p=0.036) for students who scored in the bottom performance level on their BOY DIBELS, especially for students assigned to 1:1 tutoring (0.15 SD, p=0.035). We also see positive but not

statistically significant effects for students who scored “Below Benchmark” and “Above Benchmark” but note that the small number of students who score “Above Benchmark” limits our power to detect effects.

Looking by grade level in Table 9, the effect of OnYourMark tutoring was strongest for first graders, followed by kindergarteners and second graders, although the positive effects seem to be largely driven by first graders receiving 1:1 tutoring (0.20 SD, p=017). For first graders, we see positive effects on MAP scores overall, suggesting that the MAP Reading Fluency assessment may be more sensitive to the OnYourMark intervention in first grade compared to in kindergarten and second grade.

DIBELS Subtests

Tables 10, 11, and 12 display results from analyses estimating the impact of assignment to OYM on DIBELS subtests for kindergartners, first, and second graders, respectively. DIBELS subtests measure discrete literacy skills and, as such, scores can provide more actionable information than composite scores, which have the potential to obscure gains and needs on specific literacy skills. Subtests are typically one-minute tests administered in order of skill development. As students age, the sub-tests that they are administered changes. We report results on all sub-tests taken by all students in the grade in our analytic sample; as mentioned, we do not report scores on sub-tests taken only by students who scored below or above a cut-score because the sample could be affected by our treatment and bias our results.

Among kindergarteners, OYM tutoring was most effective in improving kindergarteners’ letter sound identification (2.78, p =0.102). These changes in raw scores translate to an additional 2.8 letter sounds identified per minute. The gains on first graders’ subtests are largely driven by

students who received 1:1 tutoring, with positive effects from 1:1 tutoring on letter sound identification (9.25 additional letter sounds identified, p= 0.019), decoding (4.41 additional words decoded, p=.003), word reading (3.72 additional words read, p =0.043), and passage reading (6.06 words read correctly, p=0.044). We see generally positive, but not statistically significant, effects of tutoring on subtests among second graders in the sample.

These differences by grade level may reflect differences in the skills assessed at each grade level and the skills emphasized in the tutoring. For example, OnYourMark tutoring included significant instructional time devoted to word and passage reading (skills typically taught in and assessed in kindergarten and first grade) and did not emphasize comprehension (a second-grade skill) as much in favor of building up foundational skills first.

Discussion

We find positive and statistically significant effects of OnYourMark on students' end-of-year reading scores (0.05-0.08 SD) with stronger effects from 1:1 tutoring (0.12 SD). Additionally, we find that first graders and students who performed below grade level on their beginning-of-year reading scores benefitted the most from 1:1 tutoring (0.20 and 0.15 SD, respectively).

This study presents the first rigorous evaluation of early literacy tutoring delivered completely virtually. Findings are especially timely considering the rapid expansion of high-impact tutoring programs across the US (NSSA, 2023) and with many districts looking for alternatives to in-person models to streamline and sustain their programs (Stanford, 2022). The results from our research into OnYourMark's model are promising, especially given that the

organization is in just its second year of operation, and we studied the intervention as they expanded to serve more than seven times the number of students from the previous school year.

At the same time, we note that the positive effects produced by this virtual model are more modest than many early literacy tutoring programs delivered in person (e.g., Cortes et al., 2023). With limited comparisons of virtual tutoring models, it is difficult to determine whether some of the difference in effect sizes are due to differences in core programmatic features (e.g., online vs. in-person delivery) or to a host of other differences related to program implementation (e.g., degree of tutor training), or to differences in the early literacy test administered. We also note that features of the study design may lend themselves to more conservative estimates. Specifically, excluded from our analyses are 121 students selected by their schools to receive tutoring no matter what. These students have low literacy skills (i.e., 120 of 121 scored Well Below Benchmark on their BOY DIBELS assessment) and may possess other unobservable characteristics (like school staff selecting students whom they think would benefit most from virtual early literacy tutoring) that might promote more rapid reading acquisition. Because our analysis of effects by performance levels shows that students with the lowest BOY scores benefitted the most from tutoring, our estimates represent the lower bound of effects we might expect had the school-selected students been included.

Our findings examining the impact of tutoring group size on student achievement show that 1:1 instruction appears to drive the positive effects. These results are generally in line with the research base on the relative benefits of 1:1 instruction in small-group early literacy interventions (Gersten et al., 2020; Nickow et al., 2024). However, our exploratory analyses of effects by grade and on discrete early literacy skills suggest that the effect of group size may depend on the grade of students and the skills being targeted. We also need more research on the

impact of group size on tutoring effectiveness among different populations of students, as certain subpopulations of students may differentially benefit from peer interactions (e.g., Miles et al., 2022; Richards-Tutor et al., 2016).

Translated to additional learning, the positive effects on students' specific reading skills can be interpreted as the proportion of the control group's average gains from beginning to end of year. For example, kindergartners assigned to tutoring gained an additional 11.2% of the control group's average gains in letter sound identification from beginning to end of year, translating to an additional about 20 days of school based upon the charter system's 178-day school year. First graders assigned to 1:1 tutoring gained an additional 17.2% of the BAU control group's average gains in word reading from beginning to end of year, or an additional about 30 days of school.[5]

Our findings should be interpreted in light of the limitations of the study. Analyses excluding students identified as English learners (ELs) and with a disability limits the inferences that can be made about the effectiveness of the program for an increasingly large proportion of students in US schools from historically marginalized backgrounds (NCES, 2022). At the same time, estimates from analyses that include these students may be biased for two reasons. First, English learners and students with disabilities in the treatment group were disproportionately withdrawn from tutoring (i.e., the treatment) over the course of the year. Second, these students likely experienced a weakened treatment-control contrast: English learners and students with disabilities experienced a different counterfactual condition in which their control group counterparts also receive additional services. Our study therefore presents results from both study definitions, each of which has trade-offs between external and internal validity. Additional research should be conducted on the impact of virtual tutoring among English learners and

students with disabilities and consider from the outset how tutoring might complement the other services these students receive by law.

The OnYourMark program represents a model for high impact tutoring with the potential to address some of the challenges associated with implementing high-quality, relationship-based personalized instruction at scale (e.g., Groom-Thomas et al., 2023), especially for contexts where the supply of in-person tutors is particularly constrained. Virtual tutoring can lead to improved early literacy outcomes, and understanding what models work for whom will allow us to better understand how to increase equity in access to the most effective high-quality tutoring.

Notes

[1.] We restrict our discussions of online tutoring to studies that involve an adult providing instructional assistance to a student through chat, voice, or video. For research into computer-assisted learning without human supervision, see the review in Jamshidifarsani et al. (2019).

[2.] We pre-registered DIBELS end-of-year composite scores as our primary outcome in advance of receiving end-of-year data from OnYourMark and from schools.

[3.] Some demographic information provided by schools and OnYourMark were missing for some students. Specifically, about seven percent of students in the sample were missing data on identification as an English Learner or possession of an IEP plan, and about 11 percent of students were missing data on identification as economically disadvantaged.

[4.] Notably, a simple correlational analysis among students assigned to tutoring finds a *negative* correlation between the number of tutoring sessions a student attends and their end-of-year assessment scores (B=-0.004, SE=0.002, p=.012). One potential reason for this correlation could be that schools often "graduate out" students who appear to have made sufficient progress while keeping struggling students in the program for longer periods of time to ensure they receive additional help.

[5.] To calculate additional days of learning, we divide the effect of tutoring overall on kindergarteners' end-of-year letter sound identification score (2.78) by the difference in the control group's mean end- and beginning-of-year letter sound scores (31.49 - 6.69 = 24.8). This is .112 or 11.2% of the control group's average gains during the year. We then find 11.2% of the 178 days of the charter system's school year, which is 19.94 days. We use the same approach to calculate the additional days equivalent from 1:1 tutoring on first graders' word reading score.

References

Amendum, S. J., Vernon-Feagans, L., & Ginsberg, M. C. (2011). The Effectiveness of a Technologically Facilitated Classroom-Based Early Reading Intervention: The Targeted Reading Intervention. *The Elementary School Journal*, *112*(1), 107–131. https://doi.org/10.1086/660684

Athey, S., & Imbens, G. W. (2017). The econometrics of randomized experiments. In *Handbook of economic field experiments* (Vol. 1, pp. 73-140). North-Holland.

Carlana, M., & Ferrara, E. L. (2021). Apart but Connected: Online Tutoring and Student Outcomes during the COVID-19 Pandemic. In *EdWorkingPapers.com*. Annenberg Institute at Brown University. https://www.edworkingpapers.com/ai21-350

Chernoff, J. J., Jiang, J., Lentz, A., & Farmer, M. (2021). Guide to Using a Research-Based Process to Review and Select Early Literacy Assessments. Regional Educational Laboratory (REL) Midwest, Institute of Education Sciences. https://ies.ed.gov/ncee/rel/regions/midwest/pdf/training-and-coaching/ResearchBasedProcess-508.pdf

Cho, M.-H., & Shen, D. (2013). Self-regulation in online learning. *Distance Education*, *34*(3), 290–301. https://doi.org/10.1080/01587919.2013.835770

Cortes, K., Kortecamp, K., Loeb, S., & Robinson, C. D. (2023). *A Scalable Approach to High-Impact Tutoring for Young Readers: Results of a Randomized Controlled Trial.* https://studentsupportaccelerator.org/sites/default/files/Scalable%20Approach%20to%20High-Impact%20Tutoring.pdf

Deacon, G., & Chojnacki, G. (2023). Impacts of UPchieve On-Demand Tutoring on Students' Math Knowledge and Perceptions. Middle Years Math Grantee Report Series. In *Mathematica*. Mathematica. https://eric.ed.gov/?id=ED628646

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. https://doi.org/10.3102/0034654316687036

Dietrichson, J., Filges, T., Seerup, J. K., Klokker, R. H., Viinholt, B. C. A., Bøg, M., & Eiberg, M. (2021). Targeted school-based interventions for improving reading and mathematics for students with or at risk of academic difficulties in Grades K-6: A systematic review. *Campbell Systematic Reviews*, *17*(2), e1152. https://doi.org/10.1002/cl2.1152

Elbaum, B., Vaughn, S., Tejero Hughes, M., & Watson Moody, S. (2000). How effective are one-to-one tutoring programs in reading for elementary students at risk for reading failure? A meta-analysis of the intervention research. *Journal of Educational Psychology*, *92*(4), 605–619. https://doi.org/10.1037/0022-0663.92.4.605

Gersten, R., Haymond, K., Newman-Gonchar, R., Dimino, J., & Jayanthi, M. (2020). Meta-Analysis of the Impact of Reading Interventions for Students in the Primary Grades. *Journal of Research on Educational Effectiveness*, *13*(2), 401–427. https://doi.org/10.1080/19345747.2019.1689591

Gortazar, L., Hupkau, C., & Roldan, A. (2023). Online tutoring works: Experimental evidence from a program with vulnerable children. In *EdWorkingPapers.com*. Annenberg Institute at Brown University. https://www.edworkingpapers.com/ai23-743

Groom-Thomas, L., Leung, C., Loeb, S., Pollard, C., Waymack, N., & White, S. (2023). Challenges and Solutions: Scaling Tutoring Programs. *IDB Publications*. https://doi.org/10.18235/0005070

Harel-Gadassi, A. (2022). For whom is distance learning suitable? The relationship between distance learning, executive functions, and academic achievements. *Universal Journal of Education Research*, *10*(2), 129–136.

High-Impact Tutoring: An Equitable, Proven Approach to Address Pandemic Learning Loss and Accelerate Learning. (2021). National Student Support Accelerator. https://studentsupportaccelerator.org/sites/default/files/Presentation%20-%20What%20is%20High-Impact%20Tutoring.pdf

Howard, S. J., & Vasseleu, E. (2020). Self-Regulation and Executive Function Longitudinally Predict Advanced Learning in Preschool. *Frontiers in Psychology*, *11*. https://www.frontiersin.org/articles/10.3389/fpsyg.2020.00049

Jamshidifarsani, H., Garbaya, S., Lim, T., Blazevic, P., & Ritchie, J. M. (2019). Technology-based reading intervention programs for elementary grades: An analytical review. *Computers & Education*, *128*, 427–451. https://doi.org/10.1016/j.compedu.2018.10.003

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*, 614–618. https://doi.org/10.1257/pandp.20221038

Miles, K. P., McFadden, K. E., Colenbrander, D., & Ehri, L. C. (2022). Maximising access to reading intervention: comparing small group and one-to-one protocols of Reading Rescue. *Journal of Research in Reading*, *45*(3), 299-323.

National Center for Education Statistics. (2022). *Digest of education statistics.* https://nces.ed.gov/programs/digest/d22/tables/dt22_204.20.asp & https://nces.ed.gov/programs/digest/d22/tables/dt22_204.70.asp

National Student Support Accelerator (2023). A snapshot of state tutoring policies. Retrieved from https://studentsupportaccelerator.com/briefs/snapshot-state-tutoring-policies

NAEP Reading, Grade 4. (2022). National Assessment of Education Progress. https://www.nationsreportcard.gov/ndecore/shareredirect?su=NDE&sb=RED&gr=4&fr=2&yr=2022R3-2019R3-2017R3-2015R3-2013R3-2011R3-2009R3-2007R3-2005R3-2003R3-2002R3-2000R3-2000R2-1998R3-1998R2-1994R2-1992R2&sc=RRPCM&ju=NT&vr=TOTAL-false&st=MN-MN--ALC-BB-AB-AP-AD&sht=REPORT&urls=xplore&mi=false&svt=true&nd=0&vl=SHORT&yo=DESC&inc=NONE&up=true&rrl=SAMPLE%7CSAMPLE%7C1--JURISDICTION%7CJURISDICTION%7C2--TOTAL%7CVARIABLE%7C3&rtl=&sm=false

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. https://doi.org/10.1002/rrq.379

Nickow, A., Oreopoulos, P., & Quan, V. (2024). The Promise of Tutoring for PreK–12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence. American Educational Research Journal, 0(0). https://doi.org/10.3102/00028312231208687

NWEA (2023, October). *MAP Reading Fluency*. Retrieved October, 2023 from https://www.nwea.org/map-reading-fluency/

Richards-Tutor, C., Baker, D. L., Gersten, R., Baker, S. K., & Smith, J. M. (2016). The Effectiveness of Reading Interventions for English Learners: A Research Synthesis. *Exceptional Children*, *82*(2), 144–169. https://doi.org/10.1177/0014402915585483

Ritter, G. W., Barnett, J. H., Denny, G. S., & Albin, G. R. (2009). The Effectiveness of Volunteer Tutoring Programs for Elementary and Middle School Students: A Meta-Analysis. *Review of Educational Research*, *79*(1), 3–38. https://doi.org/10.3102/0034654308325690

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

Robinson, C. D., Bisht, B., & Loeb, S. (2022). The inequity of opt-in educational resources and an intervention to increase equitable access. In *EdWorkingPapers.com*. Annenberg Institute at Brown University. https://www.edworkingpapers.com/ai22-654

Roschelle, J., Cheng, B. H., Hodkowski, N., Neisler, J., & Haldar, L. (2020). *Evaluation of an online tutoring program in elementary mathematics*. Digital Promise. https://files.eric.ed.gov/fulltext/ED604743.pdf

School Pulse Survey. (2023). National Center for Education Statistics. https://nces.ed.gov/surveys/spp/results.asp

Slavin, R. E., Lake, C., Davis, S., & Madden, N. A. (2011). Effective programs for struggling readers: A best-evidence synthesis. *Educational Research Review*, *6*(1), 1–26. https://doi.org/10.1016/j.edurev.2010.07.002

Smolkowski, K., & Cummings, K. D. (2016). Evaluation of the DIBELS diagnostic system for the selection of native and proficient English speakers at risk of reading difficulties. Journal of Psychoeducational Assessment, 34(2), 103-118.

Stanford, L. (2022). Schools are spending big bucks on online tutoring. Here's what they've learned. EdWeek. Retrieved from https://www.edweek.org/technology/schools-are-spending-big-bucks-on-online-tutoring-heres-what-theyve-learned/2022/12

Torgerson, C., Ainsworth, H., Buckley, H., Hampden-Thompson, G., Hewitt, C., Humphry, D., Jefferson, L., Mitchell, N., & Torgerson, D. (2016). *Affordable online maths tuition: Evaluation report and executive summary*. Education Endowment Foundation. https://durham-repository.worktribe.com/output/1606609

University of Oregon, Center on Teaching and Learning (2018). Understanding the research behind DIBELS® 8th Edition (Technical Report 1801). Eugene, OR: Author.

University of Oregon (2023). 8th Edition of Dynamic Indicators of Basic Early Literacy Skills (DIBELS®): Administration and Scoring Guide, 2023 Edition. Eugene, OR: University of Oregon. Available: https:// dibels.uoregon.edu

U.S. Office of the Press Secretary (2021). Frequently Asked Questions: Elementary and Secondary School Emergency Relief Governor's Emergency Education Program Relief Programs, Washington, D.C. https://oese.ed.gov/files/2020/05/ESSER-Fund-Frequently-Asked-Questions.pdf

Wang, X., Neitzel, A. J., & Madden, N. (2023). *Lightning Squad: Assessing the Dosage Effect of Computer-Assisted Tutoring with Cooperative Learning for Struggling Readers*. EdArXiv. https://doi.org/10.35542/osf.io/qzfd4

Zhao, A., & Ding, P. (2024). To adjust or not to adjust? Estimating the average treatment effect in randomized experiments with missing covariates. Journal of the American Statistical Association, 119(545), 450-460. https://doi.org/10.1080/01621459.2022.2123814

Figures

Figure 1. Randomization Details

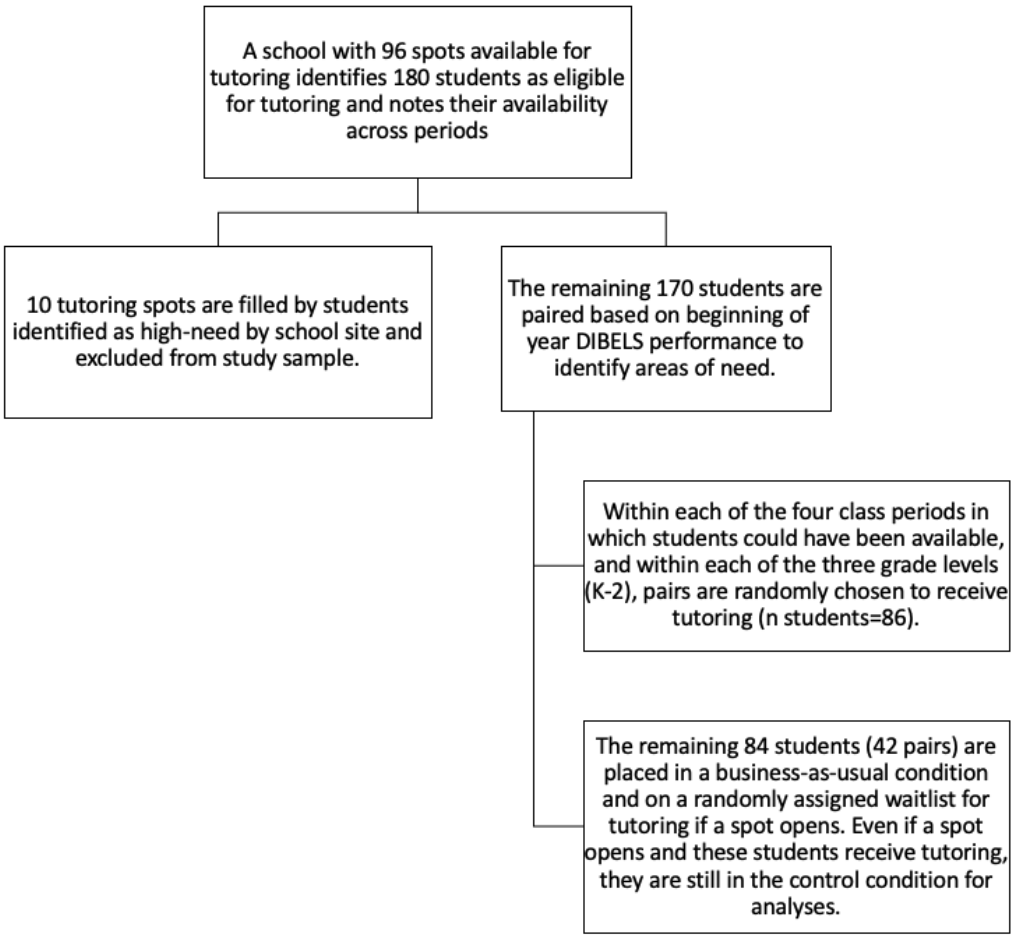


Figure 2. Average DIBELS Composite Percentile by Assignment to Tutoring

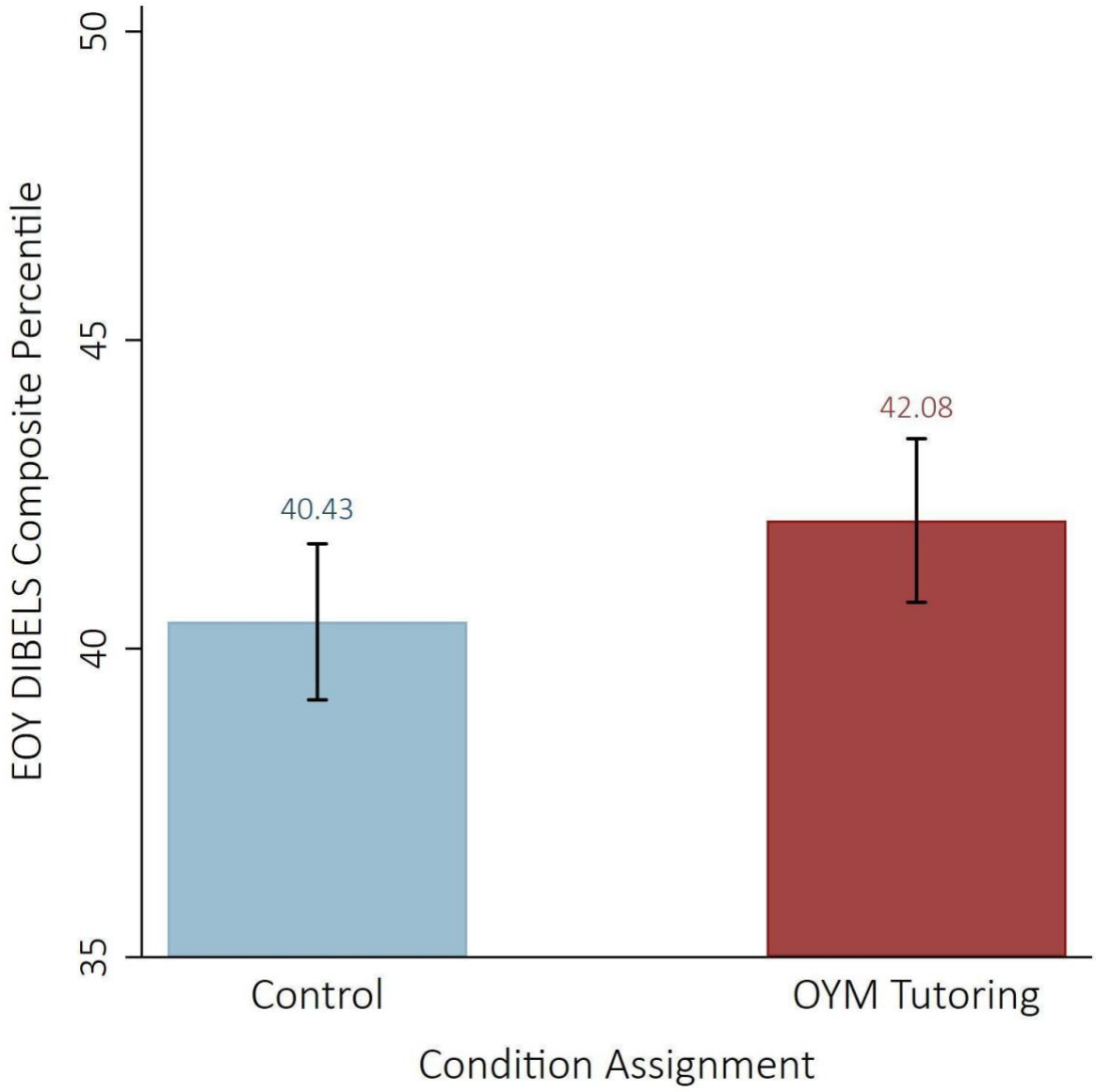
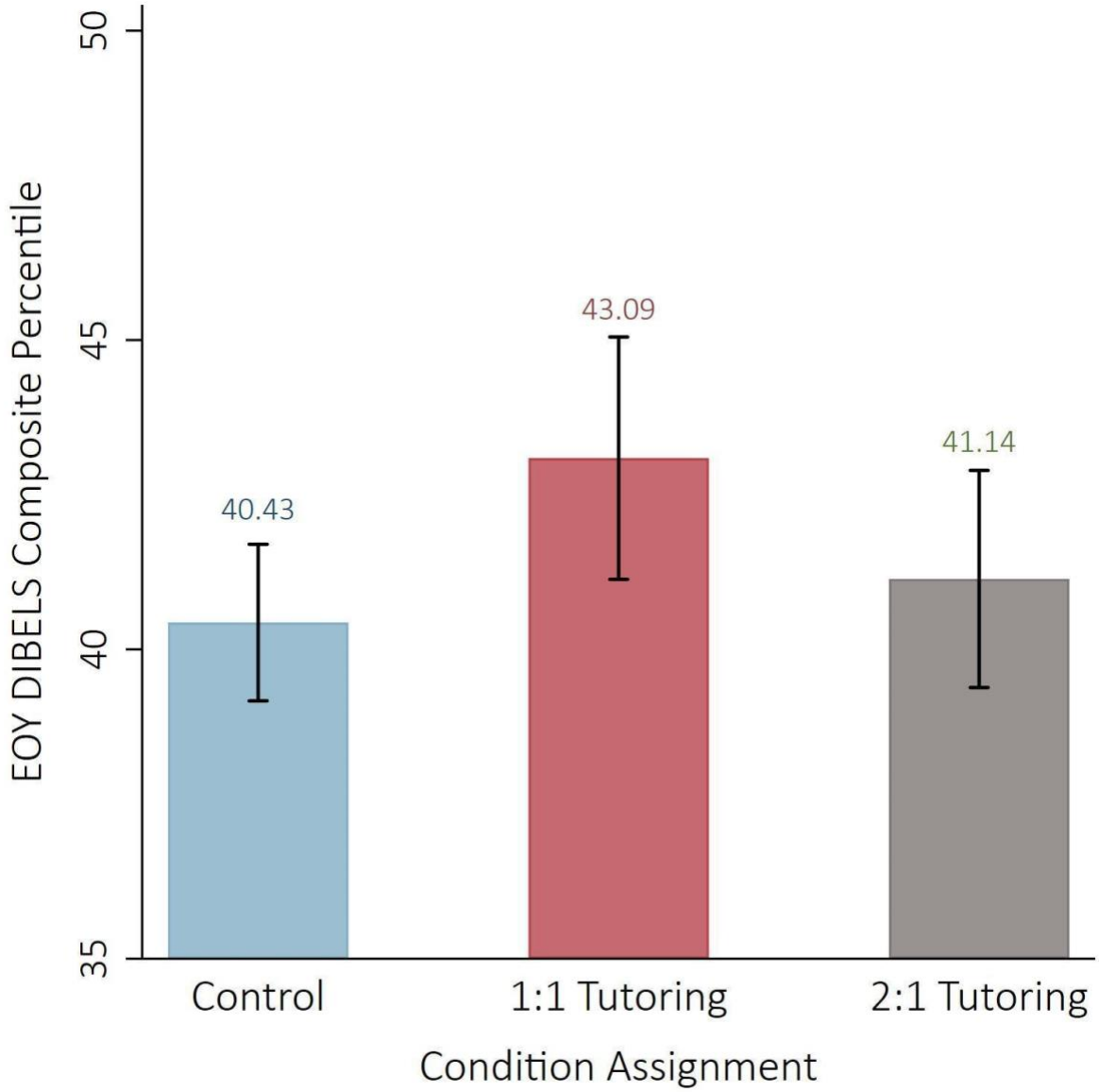


Figure 3. Average DIBELS Composite Percentile by Condition Assignment



Tables

Table 1. Descriptive Statistics and Balance Randomization

	1:1 Tutoring		2:1 Tutoring		BAU Control		
	Mean	N	Mean	N	Mean	N	P-value
Panel A. Sample of Randomized Students							
Student Demographics							
Female	.49	510	.49	570	.51	1,005	.798
Asian American	<.01	510	.01	570	.01	1,005	.437
Black	.25	510	.23	570	.26	1,005	.468
Latina/o/x	.66	510	.67	570	.66	1,005	.869
Multiracial	.03	510	.03	570	.02	1,005	.075
White	.04	510	.05	570	.03	1,005	.377
EL	.34	478	.34	532	.33	929	.869
SWD	.06	478	.05	532	.06	929	.661
ED	.93	458	.92	507	.94	899	.239
Grade							
Kindergarten	.29	510	.28	570	.30	1,005	.662
First	.39	510	.41	570	.41	1,005	.751
Second	.32	510	.32	570	.29	1,005	.401
BOY DIBELS Composite	307.80	509	306.57	570	305.05	1,004	.274
	(33.70)		(31.21)		(31.62)		
BOY DIBELS Level							
Red: Well Below Benchmark	.58	509	.57	570	.60	1,004	.442
Yellow: Below Benchmark	.18	509	.22	570	.21	1,004	.264
Green: At Benchmark	.18	509	.16	570	.14	1,004	.078
Blue: Above Benchmark	.06	509	.05	570	.05	1,004	.895
Panel B. Sample of Randomized Students Excluding English Learners and Students with Disabilities							
Student Demographics							
Female	.55	331	.50	367	.53	656	0.475
Asian American	.01	331	.01	367	.01	656	0.877
Black	.35	331	.32	367	.36	656	0.484
Latina/o/x	.51	331	.53	367	.52	656	0.918
Multiracial	.05	331	.04	367	.03	656	0.183
White	.05	331	.07	367	.05	656	0.197
ED	.89	284	.91	311	.94	560	0.101
Grade							
Kindergarten	.29	331	.27	367	.32	656	0.29
First	.40	331	.43	367	.41	656	0.72
Second	.31	331	.30	367	.27	656	0.47
BOY DIBELS Composite	310.67	330	308.99	367	305.94	656	0.08
	(35.54)		(30.62)		(32.42)		
BOY DIBELS Level							
Red: Well Below Benchmark	.53	330	.54	367	.58	656	.196
Yellow: Below Benchmark	.20	330	.24	367	.20	656	.327
Green: At Benchmark	.20	330	.17	367	.16	656	.220
Blue: Above Benchmark	.07	330	.05	367	.06	656	.707

Notes. EL= English learners; SWD= Students with disabilities. ED=economically disadvantaged. P-value is from F statistics from two-way ANOVAs testing for differences in means across the three study conditions. Numbers may not add to 100 because of rounding.

Table 2. Withdrawal from tutoring program eligibility and swapping conditions by student characteristic

	(1) Withdrawn	(2) Withdrawn	(3) Withdrawn	(4) Withdrawn	(5) Swapped In
OYM Tutoring	0.104***	0.106***	0.067***		
	(0.011)	(0.011)	(0.008)		
BOY DIBELS		-0.017**	-0.009+		
		(0.006)	(0.005)		
Female			-0.017*	-0.030+	0.044**
			(0.008)	(0.015)	(0.015)
Black			-0.002	0.001	0.020
			(0.021)	(0.039)	(0.043)
Latina/o/x			-0.005	0.006	0.020
			(0.020)	(0.037)	(0.042)
Asian American			0.046	0.101	0.169
			(0.061)	(0.162)	(0.122)
Multiracial			0.005	0.013	-0.051
			(0.036)	(0.058)	(0.054)
EL			0.033**	0.063**	0.012
			(0.011)	(0.020)	(0.018)
SWD			0.124***	0.263***	0.006
			(0.033)	(0.060)	(0.033)
School-Grade FE	Yes	Yes	Yes	Yes	Yes
R^2	0.065	0.070	0.100	0.134	0.045
Control Mean	0.016	0.015	0.004	–	–
Observations	2085	2083	1937	1010	1939

Notes. Standard errors in parentheses
+ p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Table 3. Missing end-of-year achievement outcomes

	(1) Missing EOY DIBELS	(2) Missing EOY DIBELS	(3) Missing EOY MAP	(4) Missing EOY MAP
OYM Tutoring	-0.011	-0.002	-0.012	-0.000
	(0.013)	(0.008)	(0.012)	(0.004)
Female		0.003		-0.003
		(0.008)		(0.004)
Black		-0.025		0.013*
		(0.022)		(0.007)
Latina/o/x		-0.022		0.007+
		(0.021)		(0.004)
Asian American		-0.034+		0.004
		(0.019)		(0.003)
Multiracial		-0.027		0.001
		(0.025)		(0.003)
EL		-0.010		0.000
		(0.009)		(0.004)
SWD		0.031		0.012
		(0.025)		(0.012)
School-Grade FE	Yes	Yes	Yes	Yes
R^2	0.094	0.210	0.031	0.033
Control Mean	0.109	0.037	0.081	0.006
Observations	2085	1939	2085	1939

Notes. Standard errors in parentheses
+ p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Table 4. Effect of Tutoring on DIBELS Achievement

	(1)	(2)	(3)	(4)	(5)	(6)
OYM Tutoring	0.100*	0.054+	0.060+	0.071+	0.151*	0.077*
	(0.045)	(0.030)	(0.031)	(0.037)	(0.063)	(0.038)
School-Grade FE	Yes	Yes	Yes	Yes	Yes	Yes
DIBELS Control	No	Yes	Yes	Yes	No	Yes
Student Demos	No	Yes	Yes	Yes	No	Yes
Sample						
Excludes EL	No	No	No	Yes	Yes	Yes
Excludes SWD	No	No	Yes	No	Yes	Yes
R^2	0.104	0.589	0.586	0.592	0.117	0.591
Control Mean	0.003	0.028	0.045	0.071	0.048	0.087
Observations	1869	1867	1765	1238	1164	1163

Notes. EL= English learners, SWD= Students with disabilities. Student demographic controls include variables for female, Black, Latina/o/x, EL, SWD, and economically disadvantaged. Prior DIBELS controls include students' BOY DIBELS composite score, the squared term, and an indicator for whether a student received the minimum score allowed by their grade level. Model 1 uses robust standard errors. Models 2-6 show standard errors clustered at the student pair level in parentheses.

+ $p<0.10$, * $p<0.05$, ** $p<0.01$, *** $p<0.001$

Table 5. Effect of Tutoring on MAP Reading Fluency

	(1)	(2)	(3)	(4)	(5)	(6)
OYM Tutoring	0.031	-0.009	0.008	0.016	0.109+	0.045
	(0.047)	(0.038)	(0.038)	(0.047)	(0.064)	(0.048)
School-Grade FE	Yes	Yes	Yes	Yes	Yes	Yes
DIBELS Control	No	Yes	Yes	Yes	No	Yes
Student Demos	No	Yes	Yes	Yes	No	Yes
Sample						
Excludes EL	No	No	No	Yes	Yes	Yes
Excludes SWD	No	No	Yes	No	Yes	Yes
R^2	0.062	0.380	0.382	0.373	0.077	0.385
Control Mean	-0.004	0.018	0.038	0.106	0.080	0.113
Observations	1928	1926	1818	1278	1201	1200

Notes. EL= English learners, SWD= Students with disabilities. Student demographic controls include variables for female, Black, Latina/o/x, EL, SWD, and economically disadvantaged. Prior DIBELS controls include students' BOY DIBELS composite score, the squared term, and an indicator for whether a student received the minimum score allowed by their grade level. Model 1 uses robust standard errors. Models 2-6 show standard errors clustered at the student pair level in parentheses.

+ p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Table 6. Effect of 1:1 and 2:1 Tutoring on Achievement

Panel A.	DIBELS Composite Score					
	-1	-2	-3	-4	-5	-6
1:1 Tutoring	0.127*	0.067+	0.075+	0.105*	0.208*	0.117*
	(0.057)	(0.039)	(0.040)	(0.048)	(0.083)	(0.050)
2:1 Tutoring	0.076	0.042	0.046	0.040	0.099	0.040
	(0.052)	(0.035)	(0.036)	(0.044)	(0.072)	(0.045)
School-Grade FE	Yes	Yes	Yes	Yes	Yes	Yes
DIBELS Control	No	Yes	Yes	Yes	No	Yes
Student Demos	No	Yes	Yes	Yes	No	Yes
Sample						
Excludes EL	No	No	No	Yes	Yes	Yes
Excludes SWD	No	No	Yes	No	Yes	Yes
R^2	0.105	0.589	0.587	0.593	0.119	0.592
Control Mean	0.003	0.028	0.045	0.071	0.048	0.087
Observations	1869	1867	1765	1238	1164	1163

Panel B.	MAP Reading Fluency Score					
	-1	-2	-3	-4	-5	-6
1:1 Tutoring	0.035	-0.017	0.002	0.037	0.143+	0.061
	(0.060)	(0.048)	(0.049)	(0.060)	(0.083)	(0.061)
2:1 Tutoring	0.028	-0.002	0.014	-0.004	0.077	0.030
	(0.055)	(0.046)	(0.046)	(0.057)	(0.074)	(0.057)
School-Grade FE	Yes	Yes	Yes	Yes	Yes	Yes
DIBELS Control	No	Yes	Yes	Yes	No	Yes
Student Demos	No	Yes	Yes	Yes	No	Yes
Sample						
Excludes EL	No	No	No	Yes	Yes	Yes
Excludes SWD	No	No	Yes	No	Yes	Yes
R^2	0.062	0.380	0.382	0.373	0.077	0.385
Control Mean	-0.004	0.018	0.038	0.106	0.080	0.113
Observations	1928	1926	1818	1278	1201	1200

Notes. Estimates in both panels are from models comparing each tutoring model (1:1, 2:1) to the BAU control. EL= English learners, SWD= Students with disabilities. Student demographic controls include variables for female, Black, Latina/o/x, EL, SWD, and economically disadvantaged. Prior DIBELS controls include students' BOY DIBELS composite score, the squared term, and an indicator for whether a student received the minimum score allowed by their grade level. Model 1 uses robust standard errors. Models 2-6 show standard errors clustered at the student pair level in parentheses.
+ p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Table 7. Examination of the Effects of Treatment on the Treated

Panel A. 2SLS	DIBELS Composite Score (Standardized)					
	(1) Overall Tutoring	(2) 1:1 Tutoring	(3) 2:1 Tutoring	(4) Overall Tutoring	(5) 1:1 Tutoring	(6) 2:1 Tutoring
Ever Tutored	0.072+	0.090+	0.057	0.104*	0.159*	0.064
	(0.041)	(0.051)	(0.049)	(0.053)	(0.066)	(0.062)
Sample						
Excludes EL	No	No	No	Yes	Yes	Yes
Excludes SWD	No	No	No	Yes	Yes	Yes
R^2	0.541	0.546	0.538	0.537	0.540	0.534
Panel B. First Stage	Ever Tutored					
OYM Tutoring	0.740***	0.740***	0.735***	0.742***	0.736***	0.741***
	(0.014)	(0.020)	(0.019)	(0.018)	(0.025)	(0.024)
F-Statistic	2824.28	1371.95	1475.37	1770.60	838.29	914.10
Observations	1867	1356	1407	1163	850	873

Notes. Each model includes student demographic controls (female, Black, Latina/o/x, EL, SWD, and economically disadvantaged), prior DIBELS controls (students' BOY DIBELS composite score, the squared term, and an indicator for whether a student received the minimum score allowed by their grade level). Models 4-6 exclude ELs and SWDs, however all other specifications are the same. Model 1 uses robust standard errors. Models 2-6 show standard errors clustered at the student pair level in parentheses.
+ p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Table 8. Effect of Tutoring on Reading Achievement by BOY DIBELS Level, All Grades - Excluding EL + SWD

	(1) Red (Intensive) Well Below Benchmark	(2) Yellow (Strategic) Below Benchmark	(3) Green (Core) At Benchmark	(4) Blue (Core) Above Benchmark	(5) Red+Yellow WB + B Benchmark	(6) Green+Blue At + Above Benchmark
Panel A. Effect of Tutoring Overall vs. BAU Control						
OYM Tutoring Overall	0.111*	0.050	-0.001	0.261	0.068	0.060
	(0.053)	(0.068)	(0.097)	(0.174)	(0.042)	(0.082)
Panel B. 1:1 and 2:1 Tutoring vs. BAU Control						
1:1 Tutoring	0.148*	0.080	0.072	0.135	0.097+	0.101
	(0.070)	(0.089)	(0.120)	(0.204)	(0.058)	(0.099)
2:1 Tutoring	0.076	0.025	-0.083	0.367+	0.043	0.015
	(0.060)	(0.081)	(0.124)	(0.208)	(0.046)	(0.107)
Control Mean	-0.449	0.241	0.919	1.570	-0.240	1.081
Observations	623	247	220	73	870	293

Notes. All models specified with our preferred sample excluding ELs and SWDs. Estimates from models with full sample are presented in Appendix Table S3. All models include a fixed effect for strata, student demographic controls (female, Black, Latina/o/x, and economically disadvantaged), prior DIBELS controls (students BOY DIBELS composite score, the squared term, and an indicator for whether a student received the minimum score allowed by their grade level). Standard errors clustered at the student pair level in parentheses.
+ p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Table 9. Effect of Tutoring on Reading Achievement by Group Size + Grade Level - Excluding EL + SWD

	Kindergarten		1st Grade		2nd Grade	
	(1)	(2)	(3)	(4)	(5)	(6)
	DIBELS	MAP	DIBELS	MAP	DIBELS	MAP
Panel A. Effect of Tutoring Overall vs. BAU Control						
OYM Tutoring Overall	0.077	0.006	0.090	0.132+	0.031	-0.060
	(0.073)	(0.097)	(0.058)	(0.068)	(0.065)	(0.089)
Panel B. 1:1 and 2:1 Tutoring vs. BAU Control						
1:1	0.097	0.033	0.198*	0.159+	0.009	-0.050
	(0.087)	(0.121)	(0.082)	(0.088)	(0.075)	(0.118)
2:1	0.057	-0.023	0.005	0.111	0.053	-0.070
	(0.093)	(0.117)	(0.065)	(0.082)	(0.081)	(0.105)
Control Mean	0.062	0.082	0.112	0.145	0.089	0.108
Observations	347	350	504	506	312	344

Notes. All models specified with our preferred sample excluding ELs and SWDs. Estimates from models with full sample are presented in Appendix Table S4. All models include a fixed effect for strata, student demographic controls (female, Black, Latina/o/x, and economically disadvantaged), prior DIBELS controls (students' BOY DIBELS composite score, the squared term, and an indicator for whether a student received the minimum score allowed by their grade level). Standard errors clustered at the student pair level in parentheses. The control mean is almost identical across Panel A and B, but we present the value from Panel A.
+ p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Table 10. Effect of OnYourMark on Kindergartners' DIBELS Subtest Scores - Excluding EL + SWD

	(1) Letter Naming	(2) Phoneme Segmentation	(3) Letter Sound Identification	(4) Decoding	(5) Word Reading
Panel A. Effect of Tutoring Overall vs. BAU Control					
OYM Tutoring Overall	0.038	2.563+	2.777	0.843	-0.703
	(1.624)	(1.531)	(1.690)	(0.759)	(1.085)
Panel B. 1:1 and 2:1 Tutoring vs. BAU Control					
1:1	1.706	1.536	2.964	1.198	-1.090
	(1.993)	(1.999)	(1.993)	(0.977)	(1.312)
2:1	-1.751	3.663*	2.576	0.463	-0.289
	(1.951)	(1.846)	(2.229)	(0.979)	(1.396)
Control Mean	45.555	26.371	32.109	6.442	11.709
Observations	347	347	347	347	347

Notes. All models specified with our preferred sample excluding ELs and SWDs. Estimates from models with full sample are presented in Appendix Table S5. All models include a fixed effect for strata, student demographic controls (female, Black, Latina/o/x, and economically disadvantaged), prior DIBELS controls (students' BOY DIBELS composite score, the squared term, and an indicator for whether a student received the minimum score allowed by their grade level). Standard errors clustered at the student pair level in parentheses. The control mean is almost identical across Panel A and B, but we present the value from Panel A.
+ p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Table 11. Effect of OnYourMark on 1st Graders' DIBELS Subtest Scores - Excluding EL + SWD

	(1) Letter Sound Identification	(2) Decoding	(3) Word Reading	(4) Passage Reading	(5) Reading Accuracy
Panel A. Effect of Tutoring Overall vs. BAU Control					
OYM Tutoring Overall	3.983	1.738	1.877	1.791	2.305+
	(2.774)	(1.118)	(1.319)	(2.170)	(1.362)
Panel B. 1:1 and 2:1 Tutoring vs. BAU Control					
1:1 Tutoring	9.254*	4.412**	3.725*	6.063*	2.483
	(3.915)	(1.499)	(1.832)	(2.996)	(1.677)
2:1 Tutoring	-0.203	-0.372	0.396	-1.566	2.166
	(3.106)	(1.249)	(1.429)	(2.438)	(1.568)
Control Mean	66.620	17.633	32.810	52.994	84.682
Observations	504	504	504	504	504

Notes. All models specified with our preferred sample excluding ELs and SWDs. Estimates from models with full sample are presented in Appendix Table S6. All models include a fixed effect for strata, student demographic controls (female, Black, Latina/o/x, and economically disadvantaged), prior DIBELS controls (students' BOY DIBELS composite score, the squared term, and an indicator for whether a student received the minimum score allowed by their grade level). Standard errors clustered at the student pair level in parentheses. The control mean is almost identical across Panel A and B, but we present the value from Panel A.
+ p<0.10, * p<0.05, ** p<0.01, *** p<0.001

Table 12. Effect of OnYourMark on 2nd Graders' DIBELS Subtest Scores - Excluding EL + SWD

	(1) Passage Reading	(2) Reading Accuracy	(3) Comprehension
Panel A. Effect of Tutoring Overall vs. BAU Control			
OYM Tutoring Overall	1.086	0.891	-0.357
	(2.348)	(1.106)	(0.509)
Panel B. 1:1 and 2:1 Tutoring vs. BAU Control			
1:1	-0.216	2.132	-0.686
	(2.723)	(1.355)	(0.601)
2:1	2.432	-0.415	-0.012
	(2.925)	(1.204)	(0.619)
Control Mean	69.331	90.146	5.550
Observations	312	312	312

Notes. All models specified with our preferred sample excluding ELs and SWDs. Estimates from models with full sample are presented in Appendix Table S7. All models include a fixed effect for strata, student demographic controls (female, Black, Latina/o/x, and economically disadvantaged), prior DIBELS controls (students' BOY DIBELS composite score, the squared term, and an indicator for whether a student received the minimum score allowed by their grade level). Standard errors clustered at the student pair level in parentheses. The control mean is almost identical across Panel A and B, but we present the value from Panel A.
+ $p<0.10$, * $p<0.05$, ** $p<0.01$, *** $p<0.001$