Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 25-1295

Interpreting Null Effects in Education: Insights from a Tutoring Experiment

Elizabeth Huffaker
University of Florida

Carly D. Robinson
Stanford University

Emanuele Bardelli
Santa Rosa City Schools

Sara White
Vanderbilt University

Susanna Loeb
Stanford University

We learn not only from when interventions work—but also from why they do not. Policymakers embraced high impact tutoring to address pandemic learning loss, but scaling promising practices is difficult, and not all implementations prove effective. Using random assignment, we find null (math and a low-stakes reading test) and negative (-0.09 SD on state reading scores) impacts from assignment to a virtual tutoring program. Using frameworks for interpreting null results in education experiments, we identify subject-specific heterogeneity and contribute to the literature on dosage challenges for in-school tutoring. We also identify alignment of tutoring content with classroom instruction as a moderator warranting further study. This paper offers strategies for uncovering mechanisms behind unexpected results and designing theory-driven, policy-relevant evaluations.

VERSION: November 2026

Suggested citation: Huffaker, Elizabeth, Carly D. Robinson, Emanuele Bardelli, Sara White, and Susanna Loeb. (2026). Interpreting Null Effects in Education: Insights from a Tutoring Experiment. (EdWorkingPaper: 25-1295). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/58dd-6d02

Interpreting Null Effects in Education: Insights from a Tutoring Experiment

Elizabeth Huffaker[1*]

Carly D. Robinson[2]

Emanuele Bardelli[3]

Sara May White[4]

Susanna Loeb[2]

[1] College of Education, University of Florida, 1221 SW 5th Avenue, Gainesville, FL 32601
[2] Graduate School of Education, Stanford University, 520 Galvez Mall, Stanford, CA 94305
[3] Santa Rosa City Schools, 110 Stony Point Road, Santa Rosa, CA 95401
[4] Peabody College, Vanderbilt University, 230 Appleton Place, Nashville, TN 37203
[*] Corresponding author. Email: ehuffaker@ufl.edu

November 2026

Abstract – We learn not only from when interventions work—but also from why they do not. Policymakers embraced high impact tutoring to address pandemic learning loss, but scaling promising practices is difficult, and not all implementations prove effective. Using random assignment, we find null (math and a low-stakes reading test) and negative (-0.09 SD on state reading scores) impacts from assignment to a virtual tutoring program. Using frameworks for interpreting null results in education experiments, we identify subject-specific heterogeneity and contribute to the literature on dosage challenges for in-school tutoring. We also identify alignment of tutoring content with classroom instruction as a moderator warranting further study. This paper offers strategies for uncovering mechanisms behind unexpected results and designing theory-driven, policy-relevant evaluations.

Acknowledgments – We are grateful to our partners in this research. We received insightful feedback and support from the National Student Support Accelerator team, particularly Nancy Waymack. We thank the Smith Richardson Foundation and Arnold Ventures/Accelerate for their support of our full research program.

AI Disclosure – The authors used Claude (Anthropic) as a coding and editing assistant for figure generation, reference formatting, and manuscript preparation.

Note – This paper has been accepted for publication in the American Educational Research Journal. The version of record will be available at https://doi.org/10.3102/00028312261492656.

Understanding not only when and where interventions are effective—but also why they sometimes fall short—can offer guidance for future implementation. We present findings from a district-wide randomized controlled trial (RCT) of a virtual math and reading tutoring program, highlighting not only the average null effects but also the critical implementation dynamics and contextual factors that likely explain them. The paper contributes to a growing literature that investigates how interventions are implemented, how they interact with existing support systems, and how well their content aligns with local standards and student needs. Examining the mechanisms behind both successful and unsuccessful interventions can generate insights to guide policy and strengthen the design of future educational interventions.

However, the literature on understanding null and unexpected results (e.g., Jacob et al., 2019) is often underutilized in applied practice. Null results papers often languish in the file drawer (Franco et al., 2014), but a blanket dismissal of such findings as uninformative conceals considerable and revealing heterogeneity. While well identified treatment effects share a common underpinning (i.e., a statistically significant contrast in average outcomes is observed across a treatment and control group), null results papers are unhappy in their own way – variously distinguished by insufficient statistical power, improper study design, flawed theory, weak implementation, or all the above. To demonstrate the value of using a null-results framework to guide exploratory analysis of educational experiments, we provide a practical illustration for how these concepts can be rigorously implemented in an applied context.

The focal evaluation of an online tutoring program is well-situated to serve as an exemplar case thanks to the considerable base of empirical research on tutoring programs. Despite education being a field in which few interventions consistently “work”, the core theory of action for tutoring is uniquely well supported (e.g., Nickow et al., 2024). At the same time, a

growing body of literature has documented implementation challenges that arise when programs attempt to scale, including insufficient take-up and dosage (Bhatt et al., 2025; Heinrich et al., 2014; Huffaker et al., 2025). The large tutoring research base, therefore, provides both an established theory of action as well as information about common implementation failures for our focal intervention. These insights from the literature facilitate our mapping of the null results framework onto features of our study. To implement our aligned diagnostic analyses, we then leverage rich, granular data available through research partnerships with both the online provider and school district. As a result of this process, we both contribute to evidence on drivers of effective, and ineffective, tutoring interventions and provide an accessible model for theory-driven analysis of unexpected findings in experimental settings.

Tutoring in Theory and Practice

Tutoring, defined as one-on-one or small group instruction, has become a leading strategy to support student learning. During the pandemic, most states directed federal relief dollars from the CARES Act and ESSER toward tutoring programs (LePage & Jordan, 2021; U. S. Government Accountability Office, 2024). This proliferation reflects a robust body of causal research demonstrating consistently large and positive effects on students' achievement, with effect sizes typically ranging from 0.20 to 0.40 SD and exceeding those associated with other common in-school supports (Baye, 2019; Dietrichson et al., 2017; Nickow et al., 2024; Slavin et al., 2011).

Researchers attribute tutoring's outsized effects to several core features (see Nickow et al., 2024; Robinson & Loeb, 2021). The small instructor-to-student ratio enables a degree of instructional personalization that whole-class instruction cannot replicate. Tutors can identify and address individual misconceptions in real time, adapt pacing to student need, and provide

immediate corrective feedback. Tutoring is also embedded in a sustained, caring relationship between tutor and student, which research suggests is itself a meaningful driver of motivation, engagement, and learning. Additionally, tutoring provides students with additional targeted instruction and practice time, allowing for the kind of repeated, scaffolded exposure to content that struggling learners in particular benefit from. Programs that attend deliberately to these features are referred to as "high-impact tutoring" (Robinson & Loeb, 2021).

The high-impact tutoring model validated by these studies (i.e., delivered by a trained and caring adult during the school day at least three-times a week) requires intensive inputs that may be difficult to provide at scale (White et al., 2023). Moreover, a developing literature has begun to examine additional dimensions along which tutoring efficacy is moderated, such as dosage and alignment with classroom instruction (Huffaker et al., 2025; Jackson & Shakeel, 2025; Kraft et al., 2024). Supporting these concerns, a recent meta-analysis finds that the benefits of tutoring tend to decline as districts expand their programs. In attempt to reduce costs and streamline implementation, districts often relax some program features (Kraft et al., 2024). The rapid expansion of virtual tutoring represents one attempt to implement high-impact tutoring at scale, but the effects of changing the delivery modality are only beginning to be understood.

Virtual Tutoring

Virtual tutoring – in which students are in school but they meet with their tutors through video conference – has become a popular approach for providing tutoring during the school day, at least in part because it helps to address a persistent challenge of tutor recruitment (White et al., 2023). Virtual tutoring allows for recruitment from a larger geographic area and eliminates commuting costs for tutors, which likely increases local supply as well. However, due to their relative newness, prominent meta-analyses of tutoring do not describe the effectiveness of virtual

tutoring programs. We are only recently learning about the efficacy of virtual tutoring programs (Carlana & La Ferrara, 2025; Kraft et al., 2024; Neitzel & Storey, 2024; Ready et al., 2026; Robinson et al., 2025). Carlana & La Ferrara (2025) found that virtual tutoring by university students increased middle school students' scores on a researcher-administered test of math by a magnitude comparable to in-person tutoring impacts (ES=0.20 SD).

Other studies have identified effects that are positive but smaller than those consistently found for in-person tutoring programs (Ready et al., 2026; Robinson et al., 2025), with some indications that low dosage and technological challenges may drive the attenuated effects (Kraft et al., 2024). Robinson et al. (2025) found that K-2 students assigned to receive virtual tutoring increased their early literacy skills by 0.05-0.08 SD, on average. Similarly, a study conducted by Ready et al. (2026) demonstrated that assignment to virtual early literacy tutoring in 1st through 4th grade increased student achievement on a literacy assessment by about 0.05 SD. Finally, the Personalized Learning Initiative (PLI) research team recently released interim results from the 2023-24 school year demonstrating that virtual tutoring can increase middle school student math achievement by approximately 0.10-0.12 SD (Bhatt et al., 2025). To our knowledge, no studies have rigorously examined the impact of school-based virtual tutoring on reading achievement for late elementary and middle school students and on math achievement for late elementary school students.

The Present Study

To further this emerging evidence on virtual tutoring, we conducted an RCT of a virtual math and reading tutoring program which upper elementary and middle school students attended during the school day as part of a district-wide intervention block. We address a preregistered[1] confirmatory research question (RQ1): What is the effect of being assigned to receive

synchronous, small-group virtual math and reading tutoring on the end-of-year math and reading test scores of students grades 3 through 8? In contrast to the broadly positive evaluation literature for tutoring programs (Nickow et al., 2024), we estimate null results on three of four key outcomes, and negative impacts on one outcome. Despite prevailing norms in the research community that elevate the value of statistically significant estimates, null results – especially if precise and unexpected – are often highly informative (Abadie, 2020).

Null Results in Social Science Experiments

We therefore draw on frameworks for learning from null and negative findings in social science experiments to better understand our results (Alrababa'h et al., 2023; Jacob et al., 2019; Riis-Vestergaard, 2023). Prior research highlights four categories of explanation for null results to an affirmative hypothesis: (1) a lack of statistical precision, (2) flawed or inappropriate research design, (3) an insufficient theory of action (e.g., a misunderstanding of features of the counterfactual, intervention mechanisms, or other determinants of the treatment-control contrast), and (4) implementation challenges or limitations. However, how these categories look in practice, particularly (2)-(4), will vary depending on the field of study and program of interest. The rich data afforded by our research partnerships with both the provider and focal school district allow us to investigate these explanations in the context of a virtual tutoring field experiment.

Applying the Framework

The two primarily empirical categories – (1) statistical precision and (2) research design – are addressed in our methods and results sections, and do not explain our topline findings. Rather, we confirm that our experiment featured successful randomization and was sufficiently

powered to detect effect sizes approximately one quarter of those identified, on average, by a high-impact tutoring meta-analysis (Nickow et al., 2024).

To unpack the latter pair of potential explanations for null results – (3) the theory of action and (4) the quality of program implementation – we draw on insights from tutoring research. First, with respect to the theory of action (category 3), the experiences of control-group students may lead to null results, even for well-designed tutoring programs, in two ways: (a) counterfactual supports and (b) spillover. Students in the control group often also receive supplemental academic supports. The growing practice of providing tutoring during the conventional school day has expanded participation (e.g., Bhatt et al., 2025) but also made it more difficult for researchers to conceptualize an appropriate counterfactual. Students *not* assigned to in-school tutoring are likely supported by "business-as-usual" (BaU) intervention practices during an "intervention block" or "math/reading support" section. BaU practices could include receiving informal small-group instruction from a teacher or using adaptive online educational software. Thus, many studies of in-school tutoring programs do not measure tutoring versus *no supportive instruction*, but tutoring versus a different form of supportive instruction, which will tend to attenuate estimated tutoring effect. This study typifies this dynamic: tutoring was adopted in large part due to a Texas state mandate, House Bill 4545 (HB 4545), to provide small group interventions during the school day to all students who failed to meet math and/or reading benchmarks. As our intent-to-treat (ITT) sample is restricted to students below this threshold, control students were also supposed to receive supplementary instruction.

Second, researchers and practitioners have hypothesized that positive spillovers to non-tutored students may occur when a large proportion of their classmates receive tutoring, which might, again, bias effect estimates towards zero (Nickow et al., 2024). A direct pathway for this

effect would be if in-class tutoring effectively reduces the size and heterogeneity in the instructional caseload for the classroom teacher. For example, if half the class receiving online tutoring frees the classroom teacher to provide more individualized attention to the control students, they may see gains in achievement that shrink the gap in treatment and control group performance, even if the tutoring program is effective. Berlinski et al., (2023) also finds evidence that a highly effective reading tutoring program *indirectly* induced positive externalities via peer effects for non-tutored students by improving the average achievement of their classmates and possibly reducing peer misbehavior. To the extent that targeted tutoring interventions reduce variation in classroom skill-level by raising the skills of the lowest achieving students, all students may benefit from more efficient instructional targeting from the classroom teacher (Duflo et al., 2011). Thus, a pair of compelling mechanisms – counterfactual supports and spillovers to the control group – are the basis for our exploratory analysis of category 3.

The quality of program implementation for the treatment group (category 4) can also be explored along multiple dimensions. Indeed, there are so many potential ways that tutoring programs can struggle to deliver their programming as intended (e.g., insufficient tutor training, technological challenges, not tailoring instruction to student need) that it is neither feasible nor wise to pursue every channel. As such, we focus on two dimensions that are both measurable using our study data and have surfaced as moderators for the efficacy of a wide range of tutoring interventions in the extant literature: (c) insufficient treatment group dosage, and (d) misalignment between intervention content and classroom instruction (Huffaker et al., 2025; Kraft et al., 2024; Shmoys et al., 2026).

Despite the overall strong research base for tutoring (e.g., Nickow et al., 2024) dosage (i.e., the delivery of tutoring in sufficient quantity) has long posed challenges for otherwise well-

designed programs. Effective programs tend to provide tutoring for at least 30 minutes, three times a week, but few consistently reach those benchmarks (Bhatt et al., 2025; Huffaker et al., 2025; White et al., 2021), and students cannot benefit from tutoring if they receive little of it. In earlier tutoring studies, dosage tended to be constrained because tutoring was offered outside of the school day and the students who needed it most did not attend consistently (Heinrich et al., 2014; Huffaker et al., 2025; White et al., 2021). In recent years, schools and providers have moved tutoring into the school day to address this problem. However, this has introduced new challenges that perpetuate the dosage problem (e.g., Kraft et al., 2024). School-day tutoring requires the re-allocation of scarce instructional time, as well as complex student scheduling and staffing choices. For example, tutoring-eligible students often also receive supportive services aligned with IEP or English language acquisition goals, and these supports can compete and crowd each other out (see Huffaker et al., 2025). Planning and coordination across providers and schools becomes essential but is often overlooked during the roll out of new programs (Huffaker et al., 2025; Ready et al., 2026). Limitations in provider capacity, delayed implementation timelines, and high absenteeism among tutoring-eligible students also drive low tutoring dosage (Nickow et al., 2024).

The final challenge to tutoring implementation that we explore is the alignment of tutored content with core classroom instruction. A small but growing body of research (e.g., Huffaker et al., 2025; Jackson & Shakeel, 2025; TNTP, 2025) recognizes this pathway as related to, but distinct from, mere dosage. That is, students who receive a lot of tutoring that is not aligned with the material covered in their core, classroom instruction still may not show growth on grade-level standards-aligned assessments. And students who receive little tutoring but on classroom-aligned instruction may improve on narrow subtopic scores, if not summatively. Jackson and

Shakeel (2025) provide direct, experimental evidence on the importance of core instruction-aligned tutoring. Students who were randomly assigned to receive literacy tutoring using the same instructional materials as whole-group instruction outperformed those tutored with other high-quality intervention materials by 0.12 SD. Additionally, a simulation study by TNTP (2025) found that even with equalized dosage, a strategy which aligns instruction to student need by remediating key below-level prerequisites while also addressing grade-level material outperforms a pure remediation strategy. While students with substantial unfinished learning may still grow from tutoring programs that do not reach grade-level material, this growth is unlikely to be detected on their grade-level standards-aligned assessments.

Exploratory Research Questions

We explore each tutoring-specific explanation for null effects in turn. Broadly, we ask: (RQ2) To what extent do components of our applied null results framework explain our topline findings? Table 1 summarizes the scope of this analysis. To interrogate the empirical relevance of features of the counterfactual, we ask: Did students in the control group either benefit directly from BaU interventions mandated by HB 4545 (RQ 2a) or indirectly from spillovers generated by the virtual tutoring program (RQ 2b)? Next, to understand the role of implementation in moderating tutoring efficacy among the treatment group, we ask: What are the associations between tutoring quantity (RQ 2c), content covered in tutoring (RQ 2d), and student achievement? Insights from these analyses supplement our topline experimental findings to sharpen a theory of action for virtual tutoring, inform future research, and improve implementation for similar interventions. We rule out control group supports as explaining our topline findings and we find limited evidence of spillovers. Instead, our analysis suggests the primary shortcomings of this program were insufficient dosage in literacy and a failure to align

tutored content with the breadth and emphasis of state standards. While misalignment is also a problem for in-person programs, it could be exacerbated when districts contract with virtual tutoring providers as neither their developers nor tutors are often embedded in the local educational context. This study also provides evidence that top-down statewide tutoring mandates may fail to replicate the large effects identified by the tutoring literature if the scaling efforts that ensue dilute implementation.

More broadly, in seeking to understand why a particular intervention yielded null results, this paper highlights the importance of interrogating not just when and where interventions succeed, but also why they sometimes fail to produce expected effects—offering valuable insights for theory, policy, and future research.

The Virtual Tutoring Intervention

We study the implementation of a virtual tutoring program embedded during the school day in a suburban public school district in Texas during the 2021-22 school year. Previously, this district held tutoring initiatives after school, on Saturdays, or during the summer. These tutoring programs focused primarily on students who did not meet grade level standards and were funded from a variety of sources, including federal and state funding for economically disadvantaged students (i.e., Title I) and compensatory education. Following the onset of the COVID-19 pandemic, in June 2021, the Texas legislature approved HB 4545 which mandated – but did not fund – a minimum of 30 hours of small group or individual instruction *during the school day* in each subject a student did not pass in the previous year's State of Texas Assessments of Academic Readiness (STAAR) end-of-course (EOC) tests (Tex. H.B. 4545, 2021). District leaders instructed all schools to include an intervention block for all students during the 2021-2022 school year, and these intervention blocks, which ran five days a week for at least 45

minutes at the elementary level and for 60 minutes at the secondary level, became the primary way in which the district sought to provide the state-mandated in-school tutoring hours under HB 4545.

Our study tests the effectiveness of one of the programs used during these intervention blocks. During the spring of 2021, the district partnered with a virtual tutoring provider to pilot small-group tutoring to students in five elementary and middle schools, as a supplement to school-based interventions. For the 2021-2022 school year, the district expanded access to virtual tutoring with the goal of supporting 1,680 students in grades three through eight across 28 schools.

The research practice partnership (RPP) that facilitated study of this scale up began in early 2021, prior to the expanded implementation. In this partnership, the research team worked closely with both the district and the tutoring provider throughout the study, collaborating on research design, data collection protocols, and key programmatic decisions. This close collaboration afforded the research team ongoing visibility into implementation details as they unfolded—including delays in program delivery and low student participation—and facilitated access to detailed administrative and program data that would not have been otherwise available. Critically, the partnership was established before the scale-up began, allowing the research team to prospectively design data collection systems and ensure that the measures needed to answer the study's core questions were in place from the outset. At the same time, the depth of the ongoing partnership proved equally valuable once results were known. Discussions with district and provider partners after analyses were complete surfaced additional data sources and contextual details that helped the research team interpret null findings and situate them within the program's implementation history.

The district identified 7,096 students as eligible for math tutoring, reading tutoring, or both based on whether they scored below grade level on the fall 2021 Measures of Academic Progress (MAP) tests.[2] As designed, the virtual tutoring program mirrored many of the common characteristics of effective tutoring. Impactful programs tend to have three or more tutoring sessions per week, focus on building student-tutor relationships, employ formative assessments to personalize instruction, align with school curriculum, and provide tutors training and support (Robinson & Loeb, 2021).

We were also interested in the experiences of students in the control group (RQ 2a and 2b). District administrators reported that students who were eligible for small group intervention, but not assigned to virtual tutoring, received a range of instruction from teachers, trained interventionists, and computer-based instructional software. Instructional group sizes varied and were reportedly larger than virtual tutoring groups, although district leaders indicated that most were no larger than 1:6. We refer to the range of intervention services provided to students in the control group as BaU supports.

The virtual tutoring, specifically, was not implemented as intended, providing additional motivation for our exploration dosage (RQ 2c) and alignment (RQ 2d). First, with respect to dosage, the program faced unanticipated staffing delays as tutors could not begin work before completing background checks, despite being employed remotely. This delay pushed the scaled-up implementation from the fall to spring semester. Additionally, while selected students were originally scheduled to receive tutoring five days a week, attendance logs indicate that about 80% of students assigned to tutoring attended three or fewer days per week. Conversations with teachers and principals suggested that some schools used their intervention periods for reading and math instruction two days a week each. This approach conflicted with the district's plan for

students assigned to virtual tutoring to receive daily support in just one subject area. On average, students who attended at least one tutoring session received 10 hours of tutoring total—substantially less than the 30 hours mandated by HB 4545. Second, in terms of content alignment, students were supposed to receive tutoring on a subset of grade-level standards that they had not yet mastered. Their progress was intended to be assessed by the provider to personalize their learning. However, the provider expressed that aligning student tutoring with standards emphasized in classroom instruction and assessed on the end-of-year state test was a work in progress. Finally, while virtual tutoring was conducted as planned in small groups (i.e., an average tutor-student ratio of 1:3) students were not assigned to a consistent group or tutor. Most students received a new tutor for each session they attended, reducing opportunities for relationship building.

Data and Sample

Data for this study come from the district's student information system and the tutoring provider's session logs. Thanks to the joint RPP with both the district and the provider, unusually rich and granular data on both student's experience in tutoring and their academic performance could be combined into our analytic dataset.

Data and Measures

We use enrollment data for the 2021-2022 school year for students in grades 3 through 8, which describe school and grade membership as well as characteristics including gender, free and reduced lunch status (FRL), race and ethnicity, English Language Learner (EL) status, and Special Education status. The demographic categories used are what is reported in the district's administrative data set and may not represent the full range of student identities and experiences. All students are marked as either Male or Female. All students are marked as one of the

racial/ethnic categories listed in the Table A1 or as Multiracial. No students are listed in more than one racial/ethnic category.

The district files also include fields on students' eligibility and treatment status. They indicate whether a student was eligible for a math and/or reading intervention under their interpretation of HB 4545 and whether they were required to receive additional supports, encompassing virtual tutoring and BaU interventions. Dummy variables for assignment to virtual tutoring (i.e., the experimental treatment) reflect the random assignment we conducted among eligible students.

We then merge in data on students' achievement. We observe students' achievement on the NWEA Measures of Academic Progress (MAP) and State of Texas Assessments of Academic Readiness (STAAR) standardized test for the 2020-2021 and 2021-2022 school years. The MAP assessments for math and reading measure student achievement and growth at multiple time points during the school year (see online appendix for additional detail). The district administers STAAR tests to all students grades 3 through 8 to measure students' mastery of the Texas Essential Knowledge and Skills (TEKS) curriculum standards. Unlike the MAP assessments, STAAR exams can be consequential for students and schools. We standardize both MAP and STAAR test scores by grade-level and subject using districtwide assessment data.[3] In addition to using test score data as dependent variables, we use pre-intervention MAP (fall 2021) and STAAR (spring 2021) scores as control variables in our primary models and in exploratory quasi-experimental analyses. STAAR data are also used to characterize the frequency with which specific standards (i.e., TEKs) are tested.

Finally, we leverage session-log metrics from the tutoring provider to better understand implementation. We use indicators for session attendance, session duration, and coverage of TEKs to capture content coverage and dosage in the intervention.

Analytic Sample

Our analytic sample is the pool of all 3rd through 8th grade students who were eligible for tutoring (i.e., they scored below grade-level on fall 2021 MAP reading and/or math) and enrolled in a school and grade-level that randomized a subset of students into virtual tutoring. Because some schools declined to implement virtual tutoring, we restrict our analytic sample to school-grade strata where at least one student participated in at least one virtual tutoring session. Additionally, because eligibility for tutoring was subject dependent, we estimate the effects of assignment to tutoring across subject specific sub-samples. A CONSORT diagram (Figure 1) summarizes the flow of students in each subject from eligibility through randomization and block-level implementation to the analytic and estimation samples.

Table A1 reports summary statistics for the pooled sample of students (N=3,398) identified as eligible for virtual tutoring in math (N=2,843) and/or reading (N=2,237). Panel A presents baseline student traits, panel B presents average exposure to virtual tutoring, and panels C and D aggregate pre- and post-intervention test-scores, respectively. The experimental sample is almost 90% FRL eligible, 18% classified for special education services, and 35% classified as ELs. Half of tutoring-eligible students are categorized as Hispanic, and 38% as Black. These shares align with the overall ethnoracial composition of the district but reflect a greater concentration of poor students (i.e., 90% versus approximately 60%) than the districtwide population.

Research Design

This section details the stratified randomization process and how we account for this assignment structure to estimate the intent-to-treat effects of the virtual tutoring program. The share of students randomly assigned to the treatment group was constrained by the district's contract with the tutoring provider. Within each participating school, virtual tutoring was supposed to be delivered to eight students per subject in each elementary grade 3-5 and sixteen students per subject in each middle grade 6-8. However, because of large variation in the number of eligible students between schools, before randomization occurred, we reassigned some tutoring seats from school-grade-subject blocks with few eligible students to blocks with many eligible students. Students could only be assigned to virtual tutoring in one subject and therefore students eligible for tutoring in both math and reading were less likely to be assigned to treatment for a particular subject compared with a student only eligible for tutoring that subject. Therefore, randomization effectively occurred not just within school-grade-subject strata but within school-grade-subject *eligibility* strata.

Conditional on strata membership (same school, grade, and subject-eligibility), average expected outcomes under no treatment should be equal across the virtual tutoring and BaU conditions. We cannot test this directly for unobservable traits, but we assess balance on observable baseline covariates using both statistical tests and standardized effect sizes, following What Works Clearinghouse guidance (Institute of Education Sciences, 2022). Table 2 reports means and standard errors for all eligible students (Columns 1 and 5), the BaU group (Columns 2 and 6), and the virtual tutoring group (Columns 3 and 7). Columns 4 and 8 present the strata-adjusted treatment–BaU difference with Hedges' g beneath. No individual difference is statistically significant, and we fail to reject joint. In terms of effect sizes, no characteristic

exceeds the WWC threshold of 0.25 for g. The characteristics in each sample that fall in the 0.05–0.25 SD adjustment band are included as covariates in our impact models, satisfying WWC's adjustment requirement. These results establish average baseline equivalence on both significance and effect-size criteria.

Empirical Strategy

Our main, preregistered analysis estimates the effects of being assigned to the virtual tutoring program (i.e., the ITT effect) instead of the BaU interventions on student MAP and STAAR test scores (RQ 1). We then deploy quasi-experimental and descriptive strategies to unpack our core results (RQ 2).

Estimation of Intent-to-Treat Effects

To answer our confirmatory research question, we use the following estimating equation:

$$Y_{is} = \beta_0 + \beta_1 MathTutor_{is} + \beta_2 ReadTutor_{is} + \alpha X_i + \gamma_s + \mu_{is} \quad (1)$$

Where, Y_{is} is our outcome of interest, either spring 2022 MAP or STAAR test scores in either math or reading. $MathTutor_{is}$ and $ReadTutor_{is}$ are indicator variables taking the value of 1 when student *i* in randomization strata *s* was selected for tutoring in either subject. β_1 and β_2 are the ITT coefficients of interest. They capture the estimated effects of being selected to math or reading tutoring respectively on outcome Y_{is}. When Y_{is} is a math test score, β_1 estimates the direct effect of assignment to virtual math tutoring on math achievement. β_2 captures any ITT effect of being eligible for math tutoring but being assigned to reading tutoring instead. Such an effect might appear if, for instance, tutoring in one subject crowds out intervention time in another. Interpretations are reversed when Y_{is} is a reading outcome. X_i is a student-level vector of demographic characteristics including administrative ethnoracial and gender categories, special education, and EL identification, and γ_s are school-grade-eligibility fixed effects that

capture our randomization structure. μ_{is} is the residual term. Standard errors are heteroskedasticity-robust (Abadie et al., 2023; see online appendix for additional details).

We explore potential heterogeneity in virtual tutoring effects with respect to RQ 1 across a few dimensions. Because prior research has found that the efficacy of educational interventions varies by subject and age (Bloom et al., 2008), we subset our sample to estimate grade-specific ITT effects. Additionally, we present ITT estimates by ethnoracial category, EL, and special education identification status. By examining heterogeneity across baseline traits, we can identify for whom virtual tutoring may have differential effects versus a BaU intervention. Finally, we estimate ITT effects by tested-language subsample. We do so because test scores are standardized across both languages within each grade-level[4] and disaggregation allows us to observe whether estimates are similar for English and Spanish-language administrations.

Null Results Analysis

We conduct a series of follow-up analyses to answer our exploratory research question RQ2 and better understand our main results. Specifically, we investigate whether our experimental results are plausibly explained by features of the control condition, including (2a) control group supports and (2b) spillovers, or by treatment characteristics, including (2c) quantity of tutoring received and (2d) the alignment of tutored content with grade-level state standards. These analyses are exploratory and descriptive in nature, and we caution against drawing causal conclusions from them. Instead, we use them to further develop hypotheses to guide the implementation and study of future virtual tutoring interventions. Additionally, we note that while student-level dosage and coverage research questions (i.e., associated with components 2c and 2d) were *included* in the exploratory section of the analysis plan, we did not pre-register the item-level analysis for 2c or the regression discontinuity analyses used in 2a and

2b. We pre-registered our analysis before receiving actual data on the implementation context that made such analyses feasible. In the Discussion, we draw from this experience to propose how a null results framework can help researchers develop structured and theory-grounded pre-analysis plan in the face of data and design uncertainties.

Counterfactual Supports & Spillovers (RQ 2a and 2b): One reason a program may not look effective is if it is compared to a particularly effective control condition, or if students in the control condition receive spillover benefits from treatment. For this study, the control condition is students who receive other interventions under HB 4545. These interventions may be benefiting students. To test this possibility – and answer RQ2 with respect to components 2a and 2b (Did the control group benefit either directly from non-tutoring HB-4545 eligibility or indirectly from the virtual tutoring program?) – we leverage the arbitrary threshold for small group intervention eligibility introduced by HB 4545 within a regression discontinuity (RD) design. This strategy allows us to capture the effect of *just* qualifying for math or reading support under HB 4545. We apply this approach across key sub-samples to draw comparative insights. Figure A1 presents the eligibility surface for both interventions using the district-wide sample of students with non-missing fall scores (N=13,187). Students with scores just below the x-axis are eligible for math tutoring, while those with scores just left of the y-axis are eligible for reading supports.

To assess RQ 2a, we use the RD strategy to estimate the impact of BaU interventions among only students not assigned to virtual tutoring. This comparison addresses the question: did control group students benefit from receiving the non-tutoring (i.e., BaU) small group interventions? We construct this comparison by restricting our sample to the schools that participated in the virtual tutoring experiment. We then separately estimate the effect of marginal

eligibility for HB 4545 small-group intervention among students assigned to receive virtual tutoring (i.e., our treatment group) and students assigned to receive BaU interventions (i.e., our control group). If, for example, these analyses yield positive estimates for students just under the district's HB 4545 in both groups, then our null RCT results would be explained by students in the virtual tutoring treatment and BaU control groups both receiving effective interventions.

Next, to test RQ 2b, we use observations from schools that did *not* participate in the virtual tutoring experiment in order to infer whether virtual tutoring did in fact create positive spillovers to control group students. To do this, we estimate the effect of a student being just barely eligible for HB 4545 interventions in schools that had intended to implement virtual tutoring but never did so. These schools are plausibly similar to those in our main sample, and they allow us to observe whether BaU interventions were similarly effective in schools that did and did not *also* provide virtual tutoring. If the provision of virtual tutoring to *some* students in an intervention classroom has positive spillovers – for instance, by freeing up teachers to work with smaller intervention groups – then the effect of HB 4545 eligibility should be greater in schools that took part in the virtual tutoring experiment versus in schools that did not. Meanwhile, in the absence of spillovers, we would expect similar RD estimates for students eligible for HB 4545 interventions in schools without virtual tutoring and for similarly HB 4545-eligible, control group students in schools with virtual tutoring.

We estimate these various effects (i.e., δ_1) of HB 4545 eligibility for students near the threshold using the following equation:

$$Y_{ig} = \delta_0 + \delta_1 Below_{ig} + \delta_2 Running_i + \delta_3 Below_{ig} \times Running_i + \alpha X_i + \alpha_g + \xi_{ig} \quad (2)$$

$Where\ Running_i$ is the continuous forcing variable centered at 0 within each grade level (i.e., a transformed MAP fall 2021 score). $Below_i$ is an indicator variable taking the value of 1 when

the fall 21 MAP test score for student *i* in grade-level *g* was below the eligibility cut-off for intervention.[5] The interaction term $Below_{ig} \times Running_i$ allows the relationship between the running variable and the outcome to vary on either side of the cut score. α_g controls for grade-level membership and X_i is a vector of baseline student demographic characteristics. ξ_{ig} is the residual term. Standard errors are heteroskedasticity-robust. For our primary results, we use bandwidths selected to minimize mean-squared error using the Calonico et al. (2014) procedure. Following Lee & Lemieux (2010), we also present ITT RD across a variety of alternative bandwidths and with unrestricted data in the online appendix. We follow best practices in attending to core assumptions, including establishing the integrity of the forcing variable (Figure A2), the strength of the first stage (Figure A3), and examining covariate balance across the threshold (Table A3) (Institute of Education Sciences, 2022). We find that our data are consistent with the maintained assumption of an RD design – treatment assignment is as-good-as-randomly distributed near the cut-score.

Dosage and Alignment of Tutored Content (RQ 2c and 2d): We use a variety of approaches to explore the remaining components of RQ2: could implementation shortcomings in terms of dosage (component 2c) or alignment (component 2d) explain our main results? We first use random assignment to treatment as an instrument for the quantity of virtual tutoring received, aggregated at the student-level from session-log data. On the intensive margin, the two-stage least squares (2SLS) estimation strategy scales the ITT estimates from RQ1 by the difference in average hours of virtual tutoring received by students in the treatment versus control condition. On the extensive margin, the estimates are scaled by the difference in the share of students that received any tutoring in the treatment versus the control condition. If noncompliance with virtual

tutoring assignment was high, this approach could reveal virtual tutoring effects not precisely captured by the ITT effects.

Because compliance (i.e., the average treatment-control contrast in quantity of tutoring received) was known to be low, the 2SLS strategy still might not recover the true effect of the program when delivered at the recommended dosage (i.e., 30 hours). Bhatt et al. (2025) and Huffaker et al. (2025), for instance, begin consistently identifying tutoring effects when dosage is 800-1000 minutes (i.e., 13-16 hours). Therefore, we also leverage school-grade block-level variation in virtual tutoring dosage among students assigned to treatment to explore effect heterogeneity by tutoring quantity. We do this by introducing an interaction between a continuous block-level dosage measure and treatment assignment indicator to our experimental (RQ1) model and by repeatedly restricting the sample of blocks to those meeting increasingly higher dosage thresholds (e.g., 50% of the recommended dose, 60%, etc.). Thus, we estimate the experimental ITT effect among increasingly high-fidelity blocks. While these estimates may not be generalizable to low fidelity school-grade blocks, they do capture internally valid causal estimates. This is because randomization occurred at the student-level within blocks and thus block-level dosage is exogenous to individual student characteristics (i.e., due to motivation, attendance, etc.).

Next**,** to explore the role of curricular alignment in moderating virtual tutoring effects (2d), we estimate the relationship between receiving virtual tutoring on a particular learning standard (i.e., TEK) and achievement on test items related to that same standard among treatment group students. Leveraging both item-level state-test data and the fact that the tutoring logs indicated which TEKs were covered in a particular session, we compare the probability of

correctly answering a question on the spring STAAR test for TEKs covered in a students' virtual tutoring sessions versus TEKs not covered in tutoring.

Formally, we estimate the following linear probability model with within-student and within-TEK fixed effects:

$$Correct_{qti} = \beta_0 + \beta_1 Tutored_{qti} + \delta_i + \tau_t + \sigma_{qti} \quad (3)$$

where $Correct_{qti}$ is a binary variable that is 1 when student *i* answers question *q* correctly, and 0 otherwise. $Tutored_{qti}$ is an indicator variable that is 1 when student *i* received tutoring on the TEK *t,* tested by question *q.* β_1 is the coefficient of interest and can be interpreted as the difference in the probability of answering an item correctly if a student received virtual tutoring on that TEK, versus if they did not. σ_{qti} is the error term clustered at the student level. δ_i and τ_t represent, respectively, the student and TEK fixed effects that undergird the quasi-experimental logic of this estimation strategy. Crucially, these fixed effects address two central sources of potential selection bias. First, because this is a within-student estimator, we are not merely affirming that students who attend more tutoring, and thus cover more content, did better on the test. Second, because this is a within-TEK estimator, we are implicitly controlling for TEK-level characteristics, such as difficulty, using observations from students that both did and did not receive virtual tutoring in that standard. While idiosyncratic, non-random pairing between virtually tutored TEKs and students is still a concern, rendering these estimates not strictly causal, this approach attends to the most concerning sources of systematic bias.

Finally, we explore the role of curricular coverage strategy in moderating the detected program effects by characterizing the emphasis and breadth of learning standards in tutoring versus on the state assessment. To do so, we compare TEK tutoring coverage with three measures of TEK emphasis by the state. First, we use Texas's classification of TEKs as either

"readiness" (i.e., tested every year) versus "supporting" (i.e., only tested some years). Second, we calculate the share of the 2022 STAAR items that assess a particular TEK. Third, using item-level test data from students *not* included in the experimental sample, we calculate the average rate of correctly answering 2022 STAAR questions on each TEK as a measure of TEK difficulty. Using these measures, we describe the TEKs covered in virtual tutoring sessions by subject, grade-level, our out-of-sample measure of difficulty, frequency of assessment, and state-indicated priority level.

Results

Table 3 reports the effects of assignment to the virtual tutoring program compared to students assigned to the in-person, BaU supports.

Effects of a Virtual Tutoring Experiment (RQ1)

We do not find evidence that being assigned to receive virtual tutoring has a meaningful effect on students' end-of-year (EOY) MAP and STAAR test scores, except for their reading STAAR scores. In our preferred regression model with baseline controls, null results from assignment to treatment are precisely estimated for both math tests (Columns 3 and 6, Panel A). However, we find that assignment to virtual tutoring reduced student reading STAAR scores by 0.091 standard deviations ($p < 0.01$). Conversions equate an elementary reading score decline of this magnitude to approximately 3.5 fewer months of learning (Bloom et al., 2008) or a 3.4 percentile point drop (von Hippel, 2024). This pattern of null to negative results is stable across alternative specifications (see Tables A4 and A5 in the online appendix). A plausible hypothesis for the divergence in MAP and STAAR reading effects is that BaU interventions were likely to be systematically aligned with the high-stakes STAAR, rather than the low-stakes MAP test. We interrogate the relative efficacy of virtual versus BaU tutoring with respect to standards-aligned

proficiency in more detail below. Finally, we do not find evidence of meaningful between-subject crowding out. The impact of assignment to reading tutoring on math achievement (and vice versa) are null in all Table A6 specifications except for a marginally significant negative effect on math STAAR when baseline controls are used.

We also consider other sources of effect heterogeneity based on students' academic and demographic characteristics (see Tables A7 through A9). Among our most notable observations from these results is that the STAAR reading effect is predominantly driven by substantial negative impacts in 4th (ES=-0.19 SD) and 5th (ES=-0.16 SD) grade (see Table A9, Column 4, Panels B and C for more details).

Null Results Analyses (RQ2)

Next, we report the results of our supplementary quasi-experimental and descriptive analysis to understand the null and negative estimated effects of this virtual tutoring experiment.

Counterfactual Supports (2a) & Spillovers (2b)

Table 4 presents the results of our regression discontinuity estimates of the effects of eligibility for HB 4545 small-group intervention (i.e., BaU interventions *or* virtual tutoring) across multiple tests and sub-samples. Before proceeding to the comparisons constructed to reveal any counterfactual and spillover effects, we present, for context, the aggregate ITT estimates for HB 4545 supports across the entire district in Columns 1 and 5. These results indicate that the district's 2021-22 implementation of state-mandated, small-group interventions was ineffective for marginally eligible students. In math (Panel A), eligibility caused a statistically significant 0.068 SD (Column 1) decline in MAP scores. We also estimate a negative but imprecise (-0.051 SD) effect on STAAR scores (Column 5). Marginal eligibility for reading tutoring (Panel B) had no statistically significant impact on either STAAR or MAP scores. This

null result holds even for STAAR reading scores, where we had identified a negative relative effect of assignment to virtual tutoring versus BaU in our experimental analysis.

Next, we present estimates of the effect of marginal eligibility for HB 4545 small group intervention among the experimental sample, separately for students assigned to the virtual tutoring treatment (Columns 2 and 6) and for those assigned to the BaU control (Columns 3 and 7). The estimated ITT coefficients among the control group students are null on all outcomes, with small positive magnitudes for reading and small negative magnitudes for math. Among treated students (Columns 2 and 6), the HB 4545 eligibility effect was negative on the math MAP (ES = -0.105 SD), but the negative STAAR reading effect from the experimental results does not appear. This indicates that the *relative* inefficacy of virtual vs BaU tutoring on STAAR reading may have been driven by low proficiency students farther from the eligibility threshold.

Taken together, while we cannot rule out that very low-achieving students who are not captured in our RD design benefited from the interventions, our results strongly suggest the null and negative experimental results are not explained by control and treatment group students both receiving effective interventions (RQ2a). Instead, it appears that nearly on-level students would have been better off in math, and equally well off in reading, without being eligible to receive the prescribed HB 4545 supports – BaU or virtual tutoring – during the intervention period.

Our results are also only weakly consistent with the positive spillover hypothesis (RQ2b), especially for math tutoring. In the presence of positive spillovers, we would expect students receiving BaU interventions to benefit more from those supports in schools that also provided virtual tutoring. However, when we compare HB 4545 effects across otherwise plausibly similar schools with and without virtual tutoring (i.e., Columns 3 vs. 4 and 7 vs. 8) we find only small differences. The effect of intervention eligibility on math scores is approximately 0.07 to 0.08

SD more positive in schools with virtual tutoring, which is directionally consistent with positive spillovers[6] – though the estimated magnitudes are not statistically distinct from each other. For reading scores, we observe a nonsignificant positive difference (i.e., 0.063 SD) between control group student achievement in schools with versus without virtual tutoring in average MAP scores, but a nonsignificant negative difference of 0.036 SD on the STAAR assessment.

Tutoring Dosage (2c) and Alignment (2d)

An additional hypothesis for the null and negative estimated ITT effects is that they understate the true effects of *receiving* – as opposed to just being assigned – virtual tutoring. We know that challenges in tutor clearance delayed implementation until spring semester of the implementation year. Few students were tutored for the target number of hours. Table 5 presents the results of supplemental quasi-experimental approaches to estimating the effect of treatment.

First, we estimate an average hourly treatment on the treated effect (ATE) in the regressions summarized by Table 5. Specifically, we report the results of a two-stage least squares regression that estimates the effect of being assigned to virtual tutoring on hours of virtual tutoring received. ITT effects are then scaled by this first stage, yielding results directionally consistent with our experimental effect estimates. Students assigned to virtual reading tutoring received 8.25 hours (or just under 500 minutes) of this intervention on average, compared with approximately 7.6 hours (or about 450 minutes) received by students assigned to math tutoring (Table 5, Panel A). We find that the effect of receiving any virtual tutoring and of receiving an additional hour of virtual tutoring are null for both math assessments and on the reading MAP test. On the reading state assessment, consistent with our topline findings, we estimate that every hour of virtual tutoring in reading generates a statistically significant 0.01 SD *decline* in STAAR achievement relative to a BaU reading intervention (Table 5 Column 4).

We see that merely receiving some amount of virtual tutoring does not confer academic benefits. Given that Bhatt et al. (2025) only identify statistically significant tutoring effects in 2 of 10 programs they examined where students received approximately 1000 or fewer minutes of annual tutoring, it is plausible that even if virtual tutoring could boost math or reading learning at some level of sufficient dosage, this threshold was not met. Alternatively, these results could indicate that time in virtual tutoring is simply not promoting mastery of assessed topics.

To further explore the relationship between the effectiveness of virtual tutoring and dosage, we leverage variation in average dosage across randomized blocks (Figure A4). While estimated coefficients on the interaction between mean block-level dosage and treatment are not significant for any outcome, Figure 2 (also Table A10), which summarizes disaggregated experimental effects at increasing, discrete thresholds for block-level dosage, is revealing. Math results remain null at all dosage levels, with the imprecisely estimated coefficients becoming more negative. The negative STAAR reading effect is stable in magnitude but become statistically insignificant at the 50% of recommended dosage minimum threshold. However, the efficacy of reading tutoring with respect to the reading MAP does appear to improve as block-level dosage increases. Overall, these results are suggestive that while the "dosage problem" is real for this and many other tutoring programs, it does not provide a decisive explanation for our pattern of findings. Indeed, the continued divergence in STAAR and MAP reading scores even at equivalent levels of dosage indicate that alignment in tutored content with assessed, grade-level material is a plausible moderator of tutoring effects.

We therefore turn to a final vein of exploratory analysis to better understand *why* additional instructional time in virtual tutoring may not necessarily increase academic performance and examine the academic implications of instructional alignment by virtual tutors

(RQ 2d). Table 6 reports the results from our item-level analysis of the STAAR assessments. Columns 1 and 3 present ITT estimates on the probability of answering a given item correctly, which expectedly mirror our topline experimental results. Columns 2 and 4 report the within-student difference in answering a question correctly based on whether a question is asking about a tutored versus non-tutored TEK. For math, we estimate that coverage of a TEK is associated with a 3-percentage point increase in the likelihood of answering the corresponding questions correctly ($p<0.10$). In reading, covering TEKs in virtual tutoring does not detectably improve student STAAR performance on aligned questions. The coefficient estimate is statistically insignificant, though positive (1.7-percentage points).

Although student fixed effects absorb all student-level selection (e.g., total tutoring volume, motivation, and baseline achievement) and the TEK fixed effects controls for systematic variation in difficulty or complexity across standards, these are still not strictly causal associations because the coverage of standards was not randomly assigned within students. For instance, students who progress quickly through tutoring sessions cover more standards, such that mastery may produce coverage rather than the reverse. Tutors may also appropriately target below-grade-level skills (e.g., phonics rather than grade-level reading comprehension) for students with foundational deficits, and such students may exhibit both lower measured alignment to grade-level standards and slower growth on standards-aligned assessments. We note that these threats do not operate in a single direction, as readiness-based progression through content would inflate the coverage estimates, whereas targeting of students' weakest skills would attenuate them; implementation reports also suggest fidelity to the planned diagnostic-adaptive targeting was low. Still, when we directly control for a measure of TEK difficulty, we obtain very similar estimates, and Figure 3 Panels C and D illustrate no association between assessed

difficulty and TEK coverage in virtual tutoring. Altogether, these results are suggestive that virtual tutoring from this program may be effective on covered standards in math but not in reading.

As illustrated in Figure 3 Panels A and B, we also observe a notable difference across subjects in how tutoring coverage aligns with STAAR emphasis: a one-percentage point increase in the share of STAAR questions covering a particular TEK is associated with a six-percentage point increase in math tutoring coverage, but only a two-percentage point increase in reading tutoring coverage. To further explore the relationship between TEK coverage and student performance, we disaggregate each by grade-level. Figure 4 Panel A presents the average percent of TEKs covered during virtual tutoring sessions for students in each grade-level. TEK coverage in virtual tutoring for math is largely stable across grades, ranging from 10 to 13% of standards. Nearly all math virtual tutoring sessions focused on the "readiness" TEKs prioritized by the state – the average student assigned to math tutoring received lessons covering approximately a third of these TEKs (Figure 4, Panel B). For reading tutoring, however, TEK coverage ranged from only 6% in 6th grade to 15% in 5th grade. Furthermore, only an average of 21% of "readiness" TEKs but approximately 3% of the more numerous but less frequently tested "supporting TEKs" were covered during reading tutoring sessions.[7] This is congruent with the stronger association between TEK-coverage and assessed frequency in math than reading (Figure 3, Panels A and B).

Surprisingly, however, the variation in grade-level reading TEK coverage does not appear to explain ITT effects of virtual tutoring by grade (Table A9). The negative effect of assignment to virtual tutoring on reading STAAR scores is driven by negative, statistically significant effects of -0.19 and -0.16 in 4th and 5th grade, respectively. Yet, the elementary grade levels had the largest share of the reading TEK curriculum covered in the treatment group.

Considered alongside the within-student-and-TEK estimates summarized above, this is more evidence that, while standards-aligned coverage was poor in reading compared to math, virtual reading tutoring may have simply been less effective than either math virtual tutoring or reading BaU supports at promoting state test achievement, even on tutored material.

Discussion

This study examines the effectiveness and underlying mechanisms of virtual tutoring, a potential approach to support high need students without relying on locally recruited tutors. Despite heightened post-COVID-19 demand for interventionists to facilitate high-impact tutoring, there is still limited empirical evidence on the effects of districts contracting with an online provider to deliver these services. Our evaluation shows that students randomly assigned to virtual tutoring as part of a state-mandated intervention performed on par with or worse than students who received the district's "business-as-usual" mix of support strategies.

On three of four endline assessments, we estimate precise null impacts from assignment to virtual tutoring compared with assignment to conventional small group instruction. However, on the state STAAR reading test, we estimate a statistically significant -0.091 SD intent-to-treat effect for virtual tutoring. This could imply that reading tutoring is, in general, less effectively delivered in a virtual modality than math tutoring. However, the heterogeneity in our estimates across subject, assessment, and grade-level motivated deeper exploration of the potential mechanisms underlying these topline effects.

Insights for Online Tutoring Programs

By combining data available through research-practice partnership from the partner district and tutoring provider, we weigh potential explanations for the topline null and negative results. We also leverage district-specific implementation details within the statewide policy

context of this experiment to deploy related quasi-experimental analyses. Table 7 summarizes the results of these exploratory analyses. We find compelling evidence that, relative to BaU interventions, insufficient dosage of aligned material best explains our results in math, while reading tutoring was marked by both poor alignment and inefficacy even on tutored content with respect to the state assessment. We do note these exploratory results warrant interpretive caution with respect to their causal validity. Unlike the experimental and regression discontinuity estimates, the within-student TEK estimates rest on tutored content that was not randomly assigned, reflecting both students' pace through content and tutors' appropriate targeting of below-grade-level skills. We read this uniquely granular, within-student coverage evidence as suggestive of the alignment mechanism rather than as strictly causal effects of tutored content.

However, among only the randomized blocks where over half of the recommended dosage was received by students assigned to treatment, we observe positive effects in line with those observed in effective tutoring programs (i.e., ~0.2 SD) on reading MAP scores. A possible explanation for the persistent divergence of MAP and STAAR reading effect estimates is that the tutoring program, delivered in sufficient quantities, was effective at developing foundational literacy skills but, relative to BaU supports, not the higher-order organizational and writing skills that are only assessed on the STAAR.

The null and negative regression discontinuity results also rule out beneficial counterfactual practices as a dominant explanation for our findings, at least for students near the "on-level" threshold. Our results are weakly consistent with the existence of small positive spillovers among BaU intervention students not assigned to virtual tutoring, possibly because virtual tutoring reduced caseloads of intervention teachers. However, the relevant differences are not statistically significant and should be considered speculative.

These observations carry weighty implications for policymakers seeking to support high-need learners. While the "dosage problem" is already central to discussions of educational research and practice (Bhatt et al., 2025; Huffaker et al., 2025), the importance and challenge of crafting alignment across student need, instructional materials, classroom instruction, and grade-level state standards is understudied. This study uses uniquely rich data to empirically highlight the centrality of this consideration for the effective design of virtual interventions, like tutoring. Virtual providers, especially if located out-of-state, may not deliver content aligned or responsive to state standards. In this case, our analyses indicated that BaU reading supports more effectively prepared students for the state assessment, but not a low-stakes exam.

The normative implications of the MAP vs. STAAR reading effects gap may depend on whether achievement on a high stakes standardized examination is privileged by policymakers, practitioners, and families. In the case of HB 4545, STAAR achievement was the intended eligibility measure and therefore the implicit metric for success. However, if the STAAR score gap is really a "test prep" gap that disappears and may even be reversed on a diagnostic test, virtual tutoring may generate academic benefits equivalent to BaU interventions. Virtual tutoring could be an appropriate addition to a district's suite of interventional strategies in that case.

More fundamentally, the failure of the statewide mandate to provide small group interventions to all below-grade-level students and to boost their academic achievement carries implications for tutoring policymaking. Implementing personalized, small-group interventions on the timeline and scale required by HB 4545 would strain most districts. In our partner district, for example, approximately 43% of all third through eighth graders were eligible for tutoring when the bill was passed, a mere three months before the school year began. Given such constraints, it is understandable that many districts will contract with third-party providers, as they did here, to

meet urgent demands. This approach may be especially prevalent in districts with less developed internal systems for delivering targeted student support. While external providers can play a critical role in expanding capacity, our study suggests that contracting with a tutoring provider alone does not automatically confer the academic benefits observed in prior, often more tightly controlled, tutoring RCTs. Instead, our findings offer new evidence that states considering top-down requirements for tutoring or similar interventions should invest in thoughtfully assessing and building schools' capacity to support sustainable and effective scaling of these practices.

Insights for Research

This study also provides broader insights for empirical education research, especially with respect to field experiments. First, this study highlights the value of research-practice partnerships in educational policy evaluation, particularly for interventions where empirical evidence on underlying mechanisms is limited. Such partnerships afford access to rich data, institutional context, and implementation details that enhance the insights gained from a preregistered, experimental evaluation. Second, as educational research moves towards a deeper understanding of not just average program effects, but also of features of successful intervention implementation, unpacking the "black box" of null results from studies like this can add significant value to the knowledge base — provided sufficiently detailed data and knowledge of relevant coincident policies are available. A structured null results framework that links potential sources of null effects to the treatment-specific theory of action can help researchers develop a structured plan for the exploratory analysis of specific, relevant mechanisms.

Third, we have used this study to further develop an integration of open science principles with null results analysis. Most of the exploratory analyses in this paper (i.e., including student-level dosage and session coverage questions) were preregistered. However, we believe

that future pre-analysis plans, including our own, can be substantively strengthened if researchers develop a more concrete framework for understanding the null hypothesis. Pre-analysis plans are typically structured around the theory of action: What should the treatment do, and for whom? While the theory of action and the theory of the null are in a sense inverses, reflecting explicitly on the latter may clarify what diagnostic analyses and data would be needed to distinguish among competing explanations for null effects (e.g., insufficient dosage, weak curricular alignment, implementation failures). This exercise also provides researchers with a roadmap to extract meaning in the event of null or unexpected results, which could reduce file drawer bias.

However, imposing this structure on preregistered, but exploratory, research questions presents challenges. A persistent barrier to the broader adoption of rigorous open science principles in education research is the tension between practical considerations — where design-relevant implementation details may not be understood until data are observed — and the scientific discipline that ought to be reflected in preregistered analyses (Dee, 2025; Olken, 2015; Banerjee et al., 2020). An aim of our study is to help researchers navigate this tension: learning from rich data while avoiding fishing expeditions. Understanding and responsibly negotiating this balance is likely to become even more central as advances in computational tools reduce the cost of running additional analyses (Gelman & Loken, 2014). The disciplinary role of pre-registration may become an increasingly central researcher contribution to the analytic process.

These three insights are also mutually reinforcing. The null results framework provides a structured basis for the conversations that research-practice partnerships make possible: rather than asking partners open-ended questions about available data, researchers can frontload the research process by linking “where might null results come from in this experiment?” with “how would we study such mechanisms?” And then, asking the practice partner “do such data exist,

and in what form?" This specificity, grounded in both the theory of action and open science principles, can strengthen empirical education research as an applied area of rigorous social science by continuing to shift consequential intellectual work into the design stage.

Notes

1. Registry ID: 13040.1v1 at Registry of Efficacy and Effectiveness Studies.
2. Here, the district deviated from the state stipulation that assignment to small group instruction be based on the spring 2021 STAAR scores. Given lower than typical statewide STAAR participation rates for that administration (i.e., 86% in spring 2021 vs. 98% in spring 2022), using fall 2021 MAP scores ensured more complete baseline data.
3. While national norms exist for the NWEA English language assessments and state norms exist for STAAR exams, transforming scores for both tests using the same sample holds the underlying distribution of latent student ability constant.
4. Scale scores are equivalent across administration languages according to the testing providers (NWEA, n.d.).
5. We use a structural break method (Baum, 2015) to verify these cut points in the data.
6. Using MAP scores, the effect of marginal eligibility for HB 4545 among control group students was -0.055 SD in schools that offered virtual tutoring (Table 4, Panel A, Column 3) and -0.139 SD in schools that did not offer virtual tutoring (Table 4, Panel A, Column 4), a difference of 0.084. Using STAAR scores, the effect of marginal eligibility for HB 4545 among control group students was -0.024 SD in schools that offered virtual tutoring (Table 4, Panel A, Column 7) and -0.091 SD in schools that did not offer virtual tutoring (Table 4, Panel A, Column 8), a difference of 0.067.

7. Across grades 3-8, there are 250 math TEKs, of which 80 are readiness and 170 are supporting. In reading, excluding sub-TEKs, there are 71 readiness and 88 supporting TEKs. However, because tutoring logs do not classify tutoring sub-TEKs, we categorize TEKs as "readiness" if any sub-component was "readiness". As such, we undercount the number of supporting TEKs tutored and assessed in reading.

References

Abadie, A. (2020). Statistical nonsignificance in empirical economics. *American Economic Review: Insights, 2(2)*, 193–208. https://doi.org/10.1257/aeri.20190252

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2023). When should you adjust standard errors for clustering? *The Quarterly Journal of Economics*, *138*(1), 1–35. https://doi.org/10.1093/qje/qjac038

Alrababa'h, A., Williamson, S., Dillon, A., Hainmueller, J., Hangartner, D., Hotard, M., Laitin, D. D., Lawrence, D., & Weinstein, J. (2023). Learning from null effects: A bottom-up approach. *Political Analysis*, *31*(3), 448–456. https://doi.org/10.1017/pan.2021.51

Banerjee, A., Duflo, E., Finkelstein, A., Katz, L. F., Olken, B. A., & Sautmann, A. (2020). *In praise of moderation: Suggestions for the scope and use of pre-analysis plans for RCTs in economics* (Working Paper No. 26993). National Bureau of Economic Research. https://doi.org/10.3386/w26993

Baum, C. F. (2015). ZANDREWS: Stata module to calculate Zivot-Andrews unit root test in presence of structural break *[Statistical Software Components*]. Boston College Department of Economics.

Baye, A., Inns, A., Lake, C., & Slavin, R. E. (2019). A synthesis of quantitative research on reading programs for secondary students. *Reading Research Quarterly, 54*(2), 133–166. https://doi.org/10.1002/rrq.229

Berlinski, S., Busso, M., & Giannola, M. (2023). Helping struggling students and benefiting all: Peer effects in primary education. *Journal of Public Economics, 224*, 104925. https://doi.org/10.1016/j.jpubeco.2023.104925

Bhatt, M. P., Chau, T., Condliffe, B., Davis, R., Grossman, J., Guryan, J., Ludwig, J., Magnaricotte, M., Mattera, S., Momeni, F., Oreopoulos, P., & Stoddard, G. (2025). *Personalized Learning Initiative interim report: Findings from 2023–24. University of Chicago Education Lab.*

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, 1(4), 289–328. https://doi.org/10.1080/19345740802400072

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust data-driven inference in the regression-discontinuity design. *The Stata Journal*, *14*(4), 909–946. https://doi.org/10.1177/1536867X1401400413

Carlana, M., & La Ferrara, E. (2025). Apart but connected: Online tutoring, cognitive outcomes, and soft skills. *American Economic Review, 115*(10), 3487–3513. https://doi.org/10.1257/aer.20240401

Dee, T. S. (2025). The case for preregistering quasi-experimental program and policy evaluations. *Evaluation Review*, *49*(5), 931–945. https://doi.org/10.1177/0193841x251326738

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

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American Economic Review, 101*(5), 1739–1774. https://doi.org/10.1257/aer.101.5.1739

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science,* 345(6203), 1502–1505. https://doi.org/10.1126/science.1255484

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*(6), 460–465. https://www.americanscientist.org/article/the-statistical-crisis-in-science

Heinrich, C. J., Burch, P., Good, A., Acosta, R., Cheng, H., Dillender, M., Kirshbaum, C., Nisar, H., & Stewart, M. (2014). Improving the implementation and effectiveness of out-of-school-time tutoring. *Journal of Policy Analysis and Management, 33*(2), 471–494.

Huffaker, E., Lee, M., Zhou, H., Robinson, C., & Loeb, S. (2025). *Beyond the one-teacher model: Experimental evidence on using embedded paraprofessionals as personalized instructors* (EdWorkingPaper: 25-1326). Annenberg Institute at Brown University. https://doi.org/10.26300/pzy2-wr51

Institute of Education Sciences. (2022). *What Works Clearinghouse procedures and standards handbook, Version 5.0*.

Jackson, C., & Shakeel, A. (2025). *Creating coherence: Does instructional alignment affect the impact of tutoring?* (EdWorkingPaper: 25-1132). Annenberg Institute at Brown University. https://doi.org/10.26300/en9j-xp53

Jacob, R. T., Doolittle, F., Kemple, J., & Somers, M.-A. (2019). A framework for learning from null results. *Educational Researcher*, *48*(9), 580–589. https://doi.org/10.3102/0013189x19891955

Kraft, M. A., Schueler, B. E., & Falken, G. (2024). *What impacts should we expect from tutoring at scale? Exploring meta-analytic generalizability* (EdWorkingPaper: 24-1031). Annenberg Institute at Brown University. https://doi.org/10.26300/zygj-m525

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, *48*(2), 281–355. https://doi.org/10.1257/jel.48.2.281

LePage, B., & Jordan, P. W. (2021, July 14). *How are states spending their COVID education relief funds?* The 74. https://www.the74million.org/article/how-are-states-spending-their-covid-education-relief-funds/

Neitzel, A. J., & Storey, N. (2024). *Air Reading: A randomized evaluation of a virtual tutoring model*. Center for Research and Reform in Education, Johns Hopkins University.

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*, *61*(1), 74–107. https://doi.org/10.3102/00028312231208687

NWEA. (n.d.). *MAP tests in Spanish*. Retrieved April 1, 2026, from https://teach.mapnwea.org/impl/maphelp/Content/AboutMAP/SpanishTests.htm

Olken, B. A. (2015). Promises and perils of pre-analysis plans. *Journal of Economic Perspectives*, *29*(3), 61–80. https://doi.org/10.1257/jep.29.3.61

Ready, D. D., McCormick, S. G., & Shmoys, R. J. (2026). The effects of in-school virtual tutoring on student reading development: Evidence from a short-cycle randomized controlled trial. *Journal of Education for Students Placed at Risk.* https://doi.org/10.1080/10824669.2026.2615787

Riis-Vestergaard, M. (2023, October 4). *So, you got a null result. Now what?* The Abdul Latif Jameel Poverty Action Lab (J-PAL).

Robinson, C. D., & Loeb, S. (2021). *High-impact tutoring: State of the research and priorities for future learning* (EdWorkingPaper: 21-384). Annenberg Institute at Brown University. https://doi.org/10.26300/qf76-rj21

Robinson, C. D., Pollard, C., Novicoff, S., White, S., & Loeb, S. (2025). The effects of virtual tutoring on young readers: Results from a randomized controlled trial. *Educational Evaluation and Policy Analysis*, *47*(4), 1245–1265. https://doi.org/10.3102/01623737241288845

Shmoys, R., McCormick, S., Bretas, S., Ready, D. D., McCarty, G., & Matulewicz, E. (2026). The dosage dilemma: Implementation evidence on the barriers to high-dosage tutoring. Educational Researcher, 55(7), 432–441. https://doi.org/10.3102/0013189X261437761

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

Tex. H.B. 4545, 87th Leg., R.S. (2021). https://capitol.texas.gov/tlodocs/87R/billtext/pdf/HB04545F.pdf

TNTP. (2025). *Unlocking algebra*. https://tntp.org/publication/unlocking-algebra/

U.S. Government Accountability Office. (2024). *K-12 education: School districts reported spending initial COVID relief funds on meeting students' needs and continuing school operations* (Report No. GAO-24-106913). https://www.gao.gov/products/gao-24-106913

von Hippel, P. T. (2024). Multiply by 37 (or divide by 0.027): A surprisingly accurate rule of thumb for converting effect sizes from standard deviations to percentile points.

Educational Evaluation and Policy Analysis, 47(3), 960–969. https://doi.org/10.3102/01623737241239677

White, S., Carey, M., O'Donnell, A., & Loeb, S. (2021). *Early lessons from implementing high-impact tutoring at scale*. National Student Support Accelerator. https://nssa.stanford.edu/briefs/early-lessons

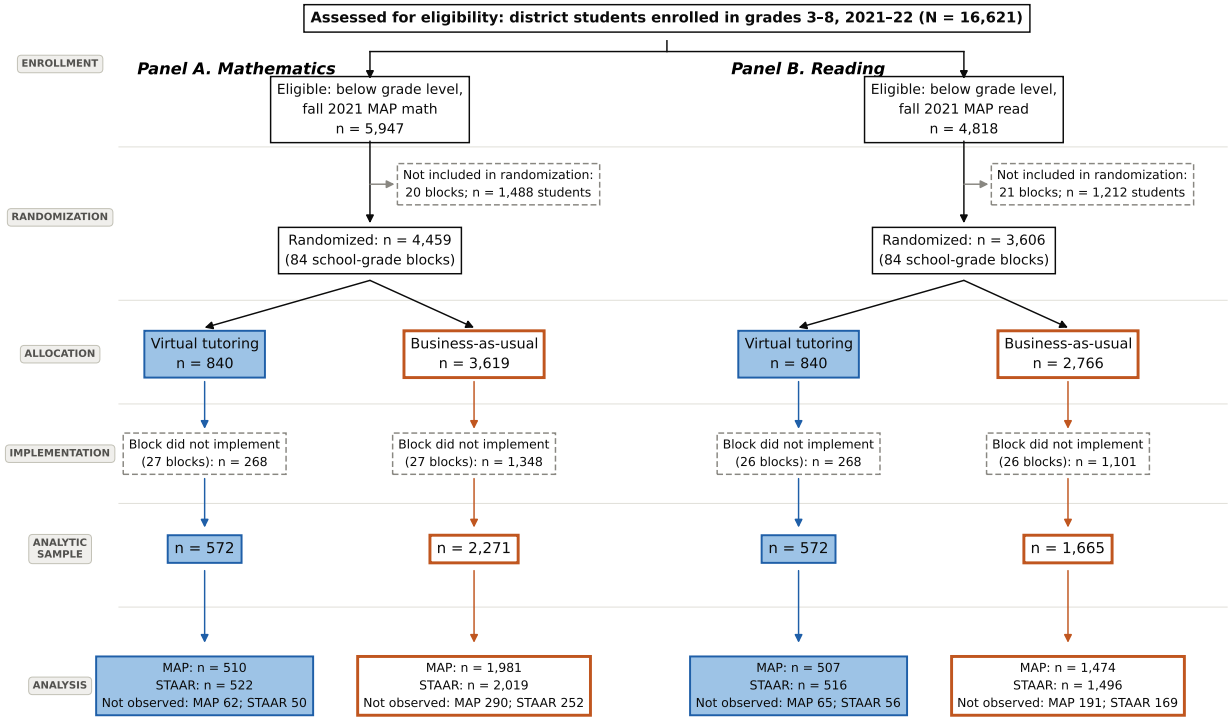
White, S., Groom-Thomas, L., & Loeb, S. (2023). *A systematic review of research on tutoring implementation: Considerations when undertaking complex instructional supports for students* (EdWorkingPaper: 22-652). Annenberg Institute at Brown University. https://doi.org/10.26300/wztf-wj14

Table 1. Summary of Null Results Analysis (RQ2)

RQ2 Component	Null Results Framework Category	Description of Application to Tutoring Context
A	Features of the counterfactual (3)	Control group supports
B		Spillover to control group
C	Features of the treatment implementation (4)	Quantity of tutoring (dosage)
D		Tutoring coverage (alignment)

Note. RQ2 = research question 2.

Figure 1. Participant Flow Through the Virtual Tutoring Experiment (CONSORT Diagram)



Note: Students eligible in both subjects appear in both panels (n = 1,682; pooled analytic N = 3,398). Randomization occurred within school × grade × eligibility strata. Inclusion in randomization and program implementation were determined at the school-grade block level. Analysis boxes count students with the indicated outcome observed (the unadjusted intent-to-treat samples, Table 3). MAP = Measures of Academic Progress; STAAR = State of Texas Assessments of Academic Readiness.

Table 2. Baseline Characteristics by Tutoring Assignment

	Eligible for Math Tutoring				Eligible for Reading Tutoring			
	All	BaU Tutoring	Virtual Tutoring	Strata Adjusted Difference	All	BaU Tutoring	Virtual Tutoring	Strata Adjusted Difference
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(A) Baseline Characteristics								
Female	0.521	0.512	0.555	0.013	0.417	0.420	0.409	0.006
	(0.009)	(0.011)	(0.021)	g=0.087 †	(0.010)	(0.012)	(0.021)	g=-0.023
FRL	0.896	0.899	0.886	-0.001	0.904	0.902	0.909	0.019
	(0.006)	(0.006)	(0.013)	g=-0.044	(0.006)	(0.007)	(0.012)	g=0.024
American Indian/Native American	0.014	0.013	0.021	0.007	0.013	0.013	0.012	0.000
	(0.002)	(0.002)	(0.006)	g=0.073 †	(0.002)	(0.003)	(0.005)	g=-0.008
Asian	0.007	0.008	0.004	-0.005	0.011	0.011	0.011	0.001
	(0.002)	(0.002)	(0.002)	g=-0.058 †	(0.002)	(0.003)	(0.004)	g=-0.002
Black	0.410	0.412	0.402	0.005	0.379	0.399	0.321	-0.054**
	(0.009)	(0.010)	(0.021)	g=-0.019	(0.010)	(0.012)	(0.020)	g=-0.160 †
Native Hawaiian/Pacific Islander	0.002	0.002	0.002	0.000	0.003	0.003	0.002	-0.000
	(0.001)	(0.001)	(0.002)	g=-0.010	(0.001)	(0.001)	(0.002)	g=-0.024
Multiracial	0.020	0.019	0.025	0.008	0.017	0.018	0.014	-0.005
	(0.003)	(0.003)	(0.007)	g=0.042	(0.003)	(0.003)	(0.005)	g=-0.026
White	0.042	0.041	0.046	0.003	0.046	0.042	0.059	0.003
	(0.004)	(0.004)	(0.009)	g=0.025	(0.004)	(0.005)	(0.010)	g=0.080 †
English Learner	0.330	0.327	0.343	0.013	0.385	0.376	0.412	0.021
	(0.009)	(0.010)	(0.020)	g=0.033	(0.010)	(0.012)	(0.021)	g=0.075 †
Special Education	0.195	0.207	0.149	-0.028	0.230	0.240	0.201	-0.015
	(0.007)	(0.009)	(0.015)	g=-0.146 †	(0.009)	(0.011)	(0.017)	g=-0.094 †
MAP Fall 21 Read (sd)	-0.522	-0.566	-0.353	0.018	-0.962	-0.979	-0.909	0.010
	(0.017)	(0.019)	(0.037)	g=0.242 †	(0.014)	(0.016)	(0.026)	g=0.110 †
MAP Fall 21 Math (sd)	-0.787	-0.806	-0.712	0.010	-0.701	-0.754	-0.548	0.033
	(0.012)	(0.014)	(0.027)	g=0.143 †	(0.018)	(0.021)	(0.036)	g=0.238 †
N Students	2809	2240	569	2809	2214	1651	563	2214
F-stat, joint-significance				0.684				0.678
Number of school-grade blocks				2809				2214

Note. Columns 1-3 (5-7) show means with (standard errors) for the math (reading) sample. Columns 4 and 8 report the strata-adjusted treatment-BaU difference; Hedges' g is shown beneath each coefficient. BaU = business-as-usual; FRL = free or reduced-price lunch; IES = Institute of Education Sciences; MAP = Measures of Academic Progress; WWC = What Works Clearinghouse. Stars: * p<0.10, ** p<0.05, *** p<0.01. Daggers follow WWC baseline equivalence thresholds (IES 2022): † for 0.05<|g|<=0.25. No variable has |g|>0.25. All † variables are included as covariates in our impact models. Demographic categories reflect district administrative data and may not capture the full range of student identities.

Table 3. Intent-to-Treat (ITT) Effects of Virtual Tutoring on Student Outcomes

	MAP			STAAR		
	(1)	(2)	(3)	(4)	(5)	(6)
(A) Impacts of Assignment to Math Tutoring on Math Achievement						
ITT Effect	0.034	0.030	0.009	0.016	0.016	0.004
	(0.037)	(0.037)	(0.028)	(0.030)	(0.030)	(0.028)
Observations	2491	2471	2471	2541	2517	2517
(B) Impacts of Assignment to Reading Tutoring on Reading Achievement						
ITT Effect	0.002	0.010	-0.023	-0.078**	-0.075**	-0.091***
	(0.044)	(0.045)	(0.037)	(0.034)	(0.035)	(0.031)
Includes design controls	Yes	Yes	Yes	Yes	Yes	Yes
Exclude observations missing baseline controls	No	Yes	Yes	No	Yes	Yes
Includes student baseline controls	No	No	Yes	No	No	Yes
Observations	1981	1965	1965	2012	1995	1995

Note. This table presents the intent-to-treat effects of being randomized into either math or reading tutoring on achievement in that subject. Columns 1 through 3 summarize standardized MAP scores while columns 4 through 6 do so for the STAAR assessment. All specifications include fixed effects to control for the design of randomization strata. Columns 3 and 6 include additional controls for student baseline characteristics. Columns 2 and 5 present regression results using the model from Columns 1 and 4 and the sub-sample used for Columns 3 and 6 (i.e., the group of students with complete data on pre-intervention traits). Heteroskedasticity robust standard errors are in parentheses. MAP = Measures of Academic Progress; STAAR = State of Texas Assessments of Academic Readiness. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table 4. Regression Discontinuity Estimates on the Effect of Eligibility for Tutoring

	MAP				STAAR			
		Schools Participating in Virtual Tutoring Experiment		Schools That Initially Intended to Use Virtual Tutoring but Did Not		Schools Participating in Virtual Tutoring Experiment		Schools That Initially Intended to Use Virtual Tutoring but Did Not
	All HB 4545 Eligible vs. Ineligible (1)	Virtual vs. Ineligible (2)	Control (BaU) vs. Ineligible (3)	HB 4545 Eligible vs. Ineligible (4)	All HB 4545 Eligible vs. Ineligible (5)	Virtual vs. Ineligible (6)	Control (BaU) vs. Ineligible (7)	HB 4545 Eligible vs. (8)
(A) Math Outcomes								
Direct ITT Effect	-0.068**	-0.105**	-0.055	-0.139**	-0.051	-0.083	-0.024	-0.091
	(0.032)	(0.053)	(0.042)	(0.063)	(0.036)	(0.067)	(0.049)	(0.072)
Bandwidth +/-	7.61	8.367	9.152	7.265	8.944	9.03	9.731	7.805
Effective Observations	4514	1800	2722	1183	5124	2035	2765	1201
(B) Reading Outcomes								
Direct ITT Effect	0.015	0.060	0.060	-0.003	0.045	0.008	0.054	0.090
	(0.034)	(0.068)	(0.045)	(0.068)	(0.039)	(0.068)	(0.056)	(0.072)
Bandwidth +/-	11.24	11.51	13.89	9.667	9.511	13.17	10.7	10.61
Effective Observations	5620	2149	3123	1215	4750	2595	2466	1384

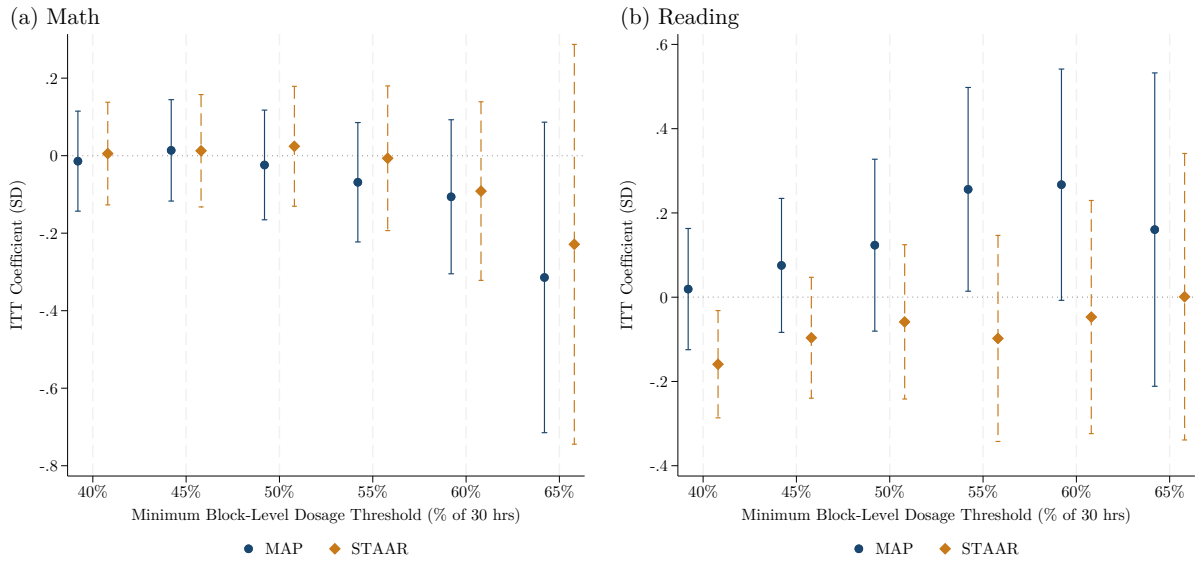
Note. Each cell presents the regression discontinuity results of tutoring eligibility and includes controls for grade level, baseline demographic and test-score covariates, and an indicator for whether students were eligible for tutoring in the other subject. Heteroskedasticity robust standard errors are in parentheses. BaU = business-as-usual; HB 4545 = Texas House Bill 4545; ITT = intent-to-treat; MAP = Measures of Academic Progress; STAAR = State of Texas Assessments of Academic Readiness. * p<0.10, ** p<0.05, *** p<0.01.

Table 5. Local Average Treatment Effects of Virtual Tutoring Among Randomized Blocks

	Math		Reading	
	MAP	STAAR	MAP	STAAR
	(1)	(2)	(3)	(4)
(A) First Stage: Hours of Virtual Tutoring	7.612***	7.571***	8.260***	8.247***
	(0.308)	(0.302)	(0.330)	(0.326)
(B) Hourly ATE	0.002	0.002	-0.003	-0.011***
	(0.003)	(0.003)	(0.004)	(0.004)
(C) First Stage: Ever Received Tutoring	0.75935***	0.76029***	0.75004***	0.74922***
	(0.018)	(0.018)	(0.019)	(0.018)
(D) LATE	0.01991	0.01774	-0.02958	-0.11812***
	(0.035)	(0.035)	(0.046)	(0.039)
Observations	2471	2517	1965	1995

Note. Panels (A) and (B) respectively report first stage and 2SLS results on the intensive margin for each math and reading assessment among the relevant sub-sample. Panels (C) and (D) report the same estimates on the extensive margin of tutoring receipt. Each regression includes controls for baseline student covariates and pre-intervention MAP test scores. Heteroskedasticity robust errors are in parentheses. 2SLS = two-stage least squares; MAP = Measures of Academic Progress; STAAR = State of Texas Assessments of Academic Readiness. * p<0.10, ** p<0.05, *** p<0.01.

Figure 2. Intent-to-Treat Effects of Virtual Tutoring by Subject and Block-Level Dosage Threshold



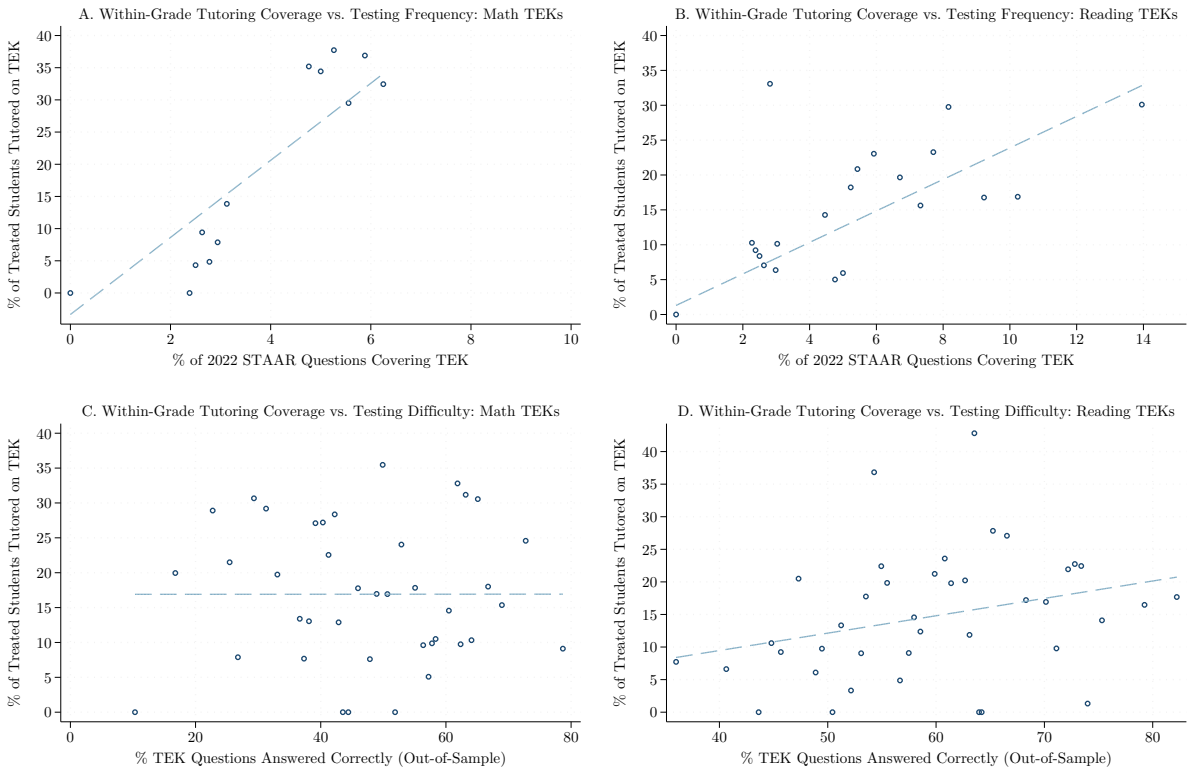
Note: Each point is the intent-to-treat (ITT) coefficient from a separate regression restricted to school-grade blocks whose treated students received at least the indicated share of the recommended 30 hours of tutoring. Vertical bars are 95% confidence intervals. Estimates control for randomization strata, baseline test scores, and student demographic characteristics, with heteroskedasticity-robust standard errors. MAP = Measures of Academic Progress; SD = standard deviation; STAAR = State of Texas Assessments of Academic Readiness.

Table 6. The Effects of Virtual Tutoring on Likelihood of Correct STAAR Responses

	STAAR Math		STAAR Reading	
	Intent-to-treat	Within-student	Intent-to-treat	Within-student
	(1)	(2)	(3)	(4)
Any item	0.005		-0.012**	
	(0.01)		(0.00)	
Item on a covered TEK		0.029*		0.017
		(0.01)		(0.01)
Strata FEs	Yes	No	Yes	No
TEK FEs	No	Yes	No	Yes
Student FEs	No	Yes	No	Yes
Number of Students	2,843	459	2,237	444
Number of Items	78094	16774	64087	16777

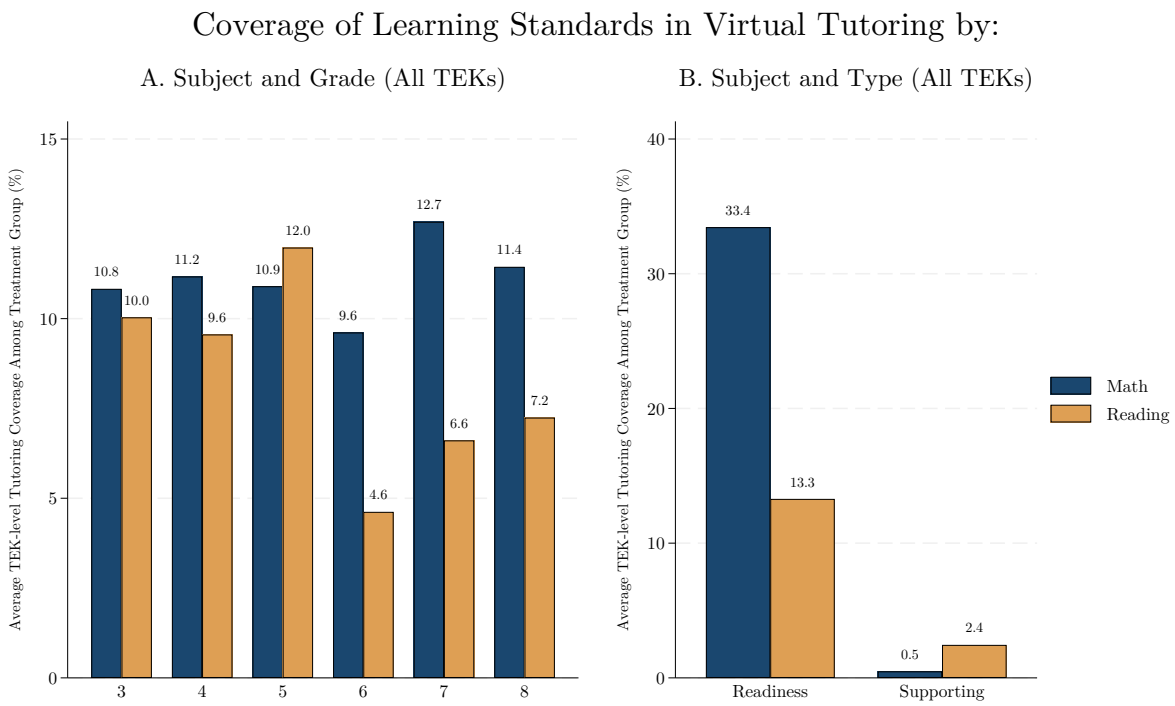
Note. Each cell presents the intent-to-treat estimates of the impact of assignment to virtual tutoring to answering any particular question on the spring 2022 STAAR assessment correctly. Columns 1 and 3 summarize the overall impact and use heteroskedasticity robust standard errors. Columns 2 and 4 estimate the within-student difference in the probability of providing a correct answer for covered relative to non-covered TEKs, with standard errors clustered by student id. Only students assigned to tutoring are used to estimate columns (2) and (4). FE = fixed effects; STAAR = State of Texas Assessments of Academic Readiness; TEK = Texas Essential Knowledge and Skills. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Figure 3. The Relationship Between Tutor Coverage of Standards (TEKs) with Tested Frequency and Difficulty



Note: Panels A and B use TEK-level data derived from students assigned to virtual tutoring, N = 1,680 (i.e., 840 observations per panel). Panels C and D plot the same tutoring coverage data as A and B against the average rate of answer correctness by TEK on the 2022 STAAR exam among district students not in the experimental sample. STAAR = State of Texas Assessments of Academic Readiness; TEK = Texas Essential Knowledge and Skills.

Figure 4. Average TEK Coverage Among Treated Students by Grade Level



Note: Bars show the mean percentage of treated students tutored on each standard, averaged across standards within the indicated group. Panel A disaggregates by subject and grade level; Panel B by subject and standard type, where the state designates each standard as either a readiness or a supporting standard. TEK = Texas Essential Knowledge and Skills.

Table 7. Summary of Null Results Analysis (RQ2) & Results

RQ2 Component	Null Results Framework Category	Description of the Application to Tutoring Context	Analytic Strategy	Key Findings: Math	Key Findings: Reading
A	Features of the counterfactual (3)	Control group supports	Regression Discontinuity (RD)	Control group supports ineffective	Control group supports null to weakly positive
B		Spillover to control group	RD sub-sample comparison	Weak evidence of positive spillovers	Very weak evidence of positive spillovers
C	Features of treatment implementation (4)	Quantity of tutoring (dosage)	Two-stage least squares; heterogeneity analysis by block-level dosage	Average dosage low: ~450 minutes, null ATE estimates; no effects at higher block-level dose	Average dosage low: ~500 minutes, negative STAAR ATE; small pos. MAP effects at higher block-level dose
D		Tutoring coverage (alignment)	Standards-level fixed effects and descriptive analyses	Coverage better aligned with testing emphasis; students perform better on tutored standards	Coverage worse aligned with testing emphasis; students perform equivalently on tutored standards

Note. ATE = average treatment effect; MAP = Measures of Academic Progress; RQ2 = research question 2; STAAR = State of Texas Assessments of Academic Readiness.

Online Appendix

MAP Administration Language

Some students completed a MAP Reading test in Spanish. NWEA indicates that scale scores are equivalent across languages. The second set of tests are the State of Texas Assessments of Academic Readiness (STAAR) exams in mathematics and reading. While both the MAP and STAAR Reading tests primarily assessed fluency and reading comprehension, the STAAR also assessed components of writing process through "revising and editing" multiple choice response items.

Standard Error Specification

Given recent guidance that clustering standard errors given individual-level randomization may generate overly conservative estimates, our main analysis uses heteroskedasticity robust standard errors (Abadie et al., 2023). However, we check the robustness of these results in response to alternative standard error specifications in the appendix. This analysis includes implementing a design-based randomization inference procedure wherein estimate uncertainty is derived from the distribution of hypothetical treatment assignment rather than sampling error (Table A4). Our results are unchanged by this approach.

Additional experimental specifications

When baseline covariates are excluded in Columns 4 and 5, the reading STAAR effect is slightly attenuated toward zero (ES=-0.075 to -0.078). Table A4 shows that alternative constructions of the standard error (i.e., allowing for strata-level clustering and using randomization inference) do not meaningfully influence the precision of our results. And, because our preferred regression described by Equation 1 differs slightly from our originally planned estimation approach, we present the analogous pre-registered effect estimates in Table

Online Appendix

A5. Null results are replicated and the negative STAAR reading effect is very similar (i.e., ES=-0.081 SD).

Turning to Table A7 replicates our null findings across key subgroups for the math and MAP reading assessments, but we do observe some sources of heterogeneity in the negative reading STAAR ITT effects. Specifically, the consequences of assignment to virtual tutoring on reading STAAR scores are most pronounced for students categorized as Hispanic, female, and not eligible for special education services.

Finally, in Table A8, while not precisely estimated for the Spanish-language assessment due to the small sample size, we identify STAAR reading effects of equal magnitude across administered languages. While math STAAR and reading MAP ITT effects for Spanish-language testers differ (i.e., are positive and larger in magnitude) from the more precise nulls estimated on the English-language assessments, they carry large standard errors.

Table A1. Descriptive Statistics

Variables	Mean	SD	Minimum	Maximum	*N*
A. Fall 2021 Student Characteristics					
Female	0.49	0.50	0	1	3398
FRL	0.89	0.31	0	1	3398
American Indian/Native America	0.01	0.12	0	1	3398
Asian	0.01	0.10	0	1	3398
Black	0.38	0.49	0	1	3398
Hispanic	0.52	0.50	0	1	3398
Native Hawaiian/Pacific Islander	0.00	0.05	0	1	3398
Multiracial	0.02	0.14	0	1	3398
White	0.05	0.21	0	1	3398
ELL	0.35	0.48	0	1	3398
Special Ed	0.18	0.39	0	1	3398
B. Virtual Tutoring					
Assigned to Reading Tutoring	0.17	0.37	0	1	3398
Assigned to Math Tutoring	0.17	0.38	0	1	3398
Hours Reading Tutoring Receive	1.29	4.37	0	44	3398
Hours Math Tutoring Received	1.23	4.19	0	40	3398
C. Pre-Intervention Math Scores					
STAAR Spring 21 Read (scale)	-0.57	0.71	-5	3	2039
STAAR Spring 21 Math (scale)	-0.54	0.62	-5	3	2042
MAP Fall 21 Read (sd)	-0.55	0.83	-3	2	3364
MAP Fall 21 Math (sd)	-0.62	0.75	-4	2	3372
D. Post-Intervention Math Scores					
MAP Spring 22 Read (sd)	-0.43	0.87	-4	2	2998
STAAR Spring 22 Read (sd)	-0.48	0.78	-3	3	3042
MAP Spring 22 Math (sd)	-0.48	0.85	-4	4	2989
STAAR Spring 22 Math (sd)	-0.47	0.69	-3	5	3042

Notes: This table reports summary statistics for the sample of students eligible to be randomized into virtual tutoring (i.e., the Intent-to-Treat sample) in math, reading, or both subjects. All data are sourced from District administative data or the tutoring providor. * p<0.10, ** p<0.05, *** p<0.010.

Table A2: Outcome Attrition by Tutoring Assignment

	All	BaU	Virtual	Strata Adj. Difference
	(1)	(2)	(3)	(4)
Panel A: Math Sample				
MAP Spring 22 Math	12.4%	12.8%	10.8%	-1.06
	(0.62)	(0.70)	(1.30)	(1.56)
STAAR 2022 Math	10.6%	11.1%	8.7%	-1.88
	(0.58)	(0.66)	(1.18)	(1.42)
N Students (Math eligible)	2843	2271	572	
Panel B: Reading Sample				
MAP Spring 22 Reading	11.4%	11.5%	11.4%	1.02
	(0.67)	(0.78)	(1.33)	(1.61)
STAAR 2022 Reading	10.1%	10.2%	9.8%	0.83
	(0.64)	(0.74)	(1.24)	(1.52)
N Students (Reading eligible)	2237	1665	572	

Notes: Each row reports the share of eligible students missing the indicated outcome. Columns 1–3 show mean attrition (%) with (standard errors). Column 4 reports the strata-adjusted Virtual-minus-BaU differential from a linear probability regression of the attrition indicator on treatment with strata fixed effects and heteroskedasticity-robust standard errors; the coefficient and (standard error) are shown in percentage points. Stars denote significance: * p<0.10, ** p<0.05, *** p<0.01. Eligibility is defined as in the baseline balance table (Table 2).

Table A3: Auxiliary Regression Discontinuity Estimates of Baseline Covariate Balance

	Math				Read			
			Bandwidth				Bandwidth	
	Full sample	N	+/- 10	N	Full sample	N	+/- 10	N
	(1)		(2)		(3)		(4)	
Female	0.021	13499	0.030	6280	-0.016	13452	-0.009	5219
	(0.015)		(0.024)		(0.013)		(0.025)	
FRL	0.011	13499	-0.007	6280	-0.012	13452	-0.026	5219
	(0.010)		(0.016)		(0.009)		(0.017)	
American Indian/Native	-0.001	13499	-0.001	6280	-0.001	13452	0.006	5219
	(0.004)		(0.005)		(0.004)		(0.006)	
Asian	0.020***	13499	-0.003	6280	0.018***	13452	0.001	5219
	(0.005)		(0.005)		(0.004)		(0.005)	
Black/African American	0.002	13499	0.001	6280	0.021	13452	0.009	5219
	(0.016)		(0.027)		(0.014)		(0.025)	
Native Hawaiian/Pacific Islander	-0.002	13499	0.000	6280	-0.002	13452	-0.005	5219
	(0.001)		(0.002)		(0.002)		(0.003)	
Multiracial	-0.002	13499	-0.005	6280	0.003	13452	-0.010	5219
	(0.002)		(0.003)		(0.005)		(0.007)	
White	-0.012*	13499	-0.010	6280	0.003	13452	-0.009	5219
	(0.007)		(0.010)		(0.006)		(0.012)	
English Language Learner	-0.018	13499	-0.022	6280	-0.023	13452	0.002	5219
	(0.015)		(0.022)		(0.017)		(0.028)	
Special Education	-0.020***	13499	-0.013	6280	0.020**	13452	-0.000	5219
	(0.007)		(0.012)		(0.009)		(0.015)	
STAAR 21 Math (Standardized)	0.045	8625	-0.052	3865	0.072***	8605	-0.009	3326
	(0.027)		(0.038)		(0.026)		(0.048)	
STAAR 21 Reading (Standardized)	-0.070**	8540	0.039	3831	-0.104***	8519	0.029	3308
	(0.030)		(0.044)		(0.022)		(0.035)	
Observations	2491		2294		1833		1792	

Note: Cells contain ITT estimates from individual regressions where the dependent variable is a baseline student covariate. The running variable is fitted with flexible linear splines. Heteroskedasticity robust standard errors are in parentheses. * p<0.10, ** p<0.05, *** p<0.010.

Table A4: Intent-to-Treat (ITT) Effects of Virtual Tutoring on Student Outcomes, Alternative Standard Errors

	MAP		STAAR	
	Clustered by Strata	Randomization Inference	Clustered by Strata	Randomization Inference
	(1)	(2)	(3)	(4)
(A) Impacts of Math Tutoring on Math Achievement				
ITT	0.009	0.009	0.004	0.004
	(0.031)	(0.029)	(0.027)	(0.027)
Observations	2471	2471	2517	2517
(B) Impacts of Reading Tutoring on Reading Achievement				
ITT	-0.023	-0.023	-0.091***	-0.091***
	(0.034)	(0.036)	(0.033)	(0.032)
Observations	1965	1965	1995	1995

This table presents the intent-to-treat effects of being randomized into either math or reading tutoring on achievement in that subject, with alternatively constructed standard errors. All models include controls for baseline characteristics, pre-intervention test scores, an indicator for assignment to virtual tutoring in the other subject, and randomization strata * $p<0.10$, ** $p<0.05$, *** $p<0.010$.

Table A5: Intent-to-Treat (ITT) Effects of Virtual Tutoring on Student Outcomes, Pre-Registered Specification

	MAP		STAAR	
	Preferred	Pre-registered	Preferred	Pre-registered
	(1)	(2)	(3)	(4)
(A) Impacts of Math Tutoring on Math Achievement				
ITT	0.009	0.006	0.004	-0.001
	(0.028)	(0.026)	(0.028)	(0.026)
Observations	2471	2471	2517	2517
(B) Impacts of Reading Tutoring on Reading Achievement				
ITT	-0.023	-0.019	-0.091***	-0.081**
	(0.037)	(0.031)	(0.031)	(0.031)
Controls for grade-school block	No	Yes	No	Yes
Includes "eligible for both" indicator control	No	Yes	No	Yes
Controls for grade-school-subject-eligibility bl	Yes	No	Yes	No
Observations	1965	1965	1995	1995

This table compares intent-to-treat effects of being randomized into either math or reading tutoring on achievement in that subject across the preferred (e.g., Table 3, Column 3) versus pre-registered specifications. All models include controls for baseline characteristics, pre-intervention test scores, and an indicator for assignment to virtual tutoring in the other subject,. In Columns 1 and 3, heteroskedasticity robust standard errors are in parentheses. In Columns 2 and 4, standard errors are clustered at the grade-school block level. * p<0.10, ** p<0.05, *** p<0.010

Table A6: Intent-to-Treat (ITT) Effects of Virtual Tutoring on Cross-Subject Student Outcomes

	MAP			STAAR		
	(1)	(2)	(3)	(4)	(5)	(6)
(A) Impacts of Assignment to Reading Tutoring on Math Achievement, Among Students Eligible for Math Tutoring in Blocks that Received Math Tutorir						
Spillover ITT Efffct	0.027	0.026	-0.035	-0.018	-0.018	-0.055*
	(0.048)	(0.048)	(0.033)	(0.036)	(0.036)	(0.032)
Observations	2491	2471	2471	2541	2517	2517
(B) Impacts of Assignment to Math Tutoring on Reading Achievement, Among Students Eligible for Reading Tutoring in Blocks that Received Reading T						
Spillover ITT Effect	0.040	0.043	-0.003	0.009	0.010	-0.012
	(0.058)	(0.058)	(0.047)	(0.043)	(0.043)	(0.039)
Includes design controls	Yes	Yes	Yes	Yes	Yes	Yes
Exclude observations missing baseline contrc	No	Yes	Yes	No	Yes	Yes
Includes student baseline controls	No	No	Yes	No	No	Yes
Observations	1981	1965	1965	2012	1995	1995

This table presents the intent-to-treat effects summarized of being randomized into either math or reading tutoring on achievement in the other subject area if the student was also eligible for tutoring in that subject (e.g., for evidence of crowding out supports). Columns 1 through 3 summarize standardized MAP scores while columns 4 through 6 do so for the STAAR assessment. All specifications include fixed effects to control for the design of randomization strata. Columns 3 and 6 include additional controls for student baseline characteristics. Columns 2 and 5 present regression results using the model from Columns 1 and 4 and the sub-sample used for Columns 3 and 6 (i.e., the group of students with complete data on pre-intervention traits). Heteroskedasticity robust standard errors are in parentheses.
* p<0.10, ** p<0.05, *** p<0.010

Table A7. Estimates of the Intent-to-Treat (ITT) Effect of Virtual Tutoring by Student Traits

	Math		Reading	
	MAP	STAAR	MAP	STAAR
	(1)	(2)	(3)	(4)
Male	-0.017	-0.025	0.039	-0.022
	(0.050)	(0.045)	(0.052)	(0.041)
Observations	1165	1193	1132	1151
Female	0.028	-0.002	-0.076	-0.191***
	(0.034)	(0.037)	(0.059)	(0.052)
Observations	1306	1324	1323	1338
Hispanic	0.009	-0.007	0.016	-0.101**
	(0.039)	(0.039)	(0.059)	(0.048)
Observations	1291	1303	1313	1316
White	0.216	-0.376	-0.205	-0.070
	(0.261)	(0.239)	(0.449)	(0.329)
Observations	94	100	97	100
Black	-0.012	0.034	0.026	-0.074
	(0.047)	(0.043)	(0.079)	(0.057)
Observations	978	1005	987	1011
Other	0.203	-0.142	0.033	0.076
	(0.122)	(0.142)	(0.240)	(0.175)
Observations	202	209	206	210
English Learner	-0.020	-0.019	-0.003	-0.106*
	(0.049)	(0.050)	(0.076)	(0.058)
Observations	870	879	884	884
Not Classified English I	0.025	0.009	0.013	-0.096**
	(0.036)	(0.035)	(0.057)	(0.044)
Observations	1601	1638	1622	1653
Special Education	0.002	-0.064	0.055	-0.076
	(0.091)	(0.072)	(0.102)	(0.072)
Observations	476	491	487	500
Not Classified Special E	0.022	-0.001	-0.003	-0.102**
	(0.030)	(0.031)	(0.048)	(0.040)
Observations	1995	2026	2019	2037

Note: This table presents the intent-to-treat effects of being randomized into either math or reading tutoring on achievement in that subject, by baseline demographic traits. All models include controls for baseline characteristics, pre-intervention test scores, and indicator for assignment to virtual tutoring in the other subject, and randomization strata. * p<0.10, **

Table A8. Estimates of the Intent-to-Treat (ITT) Effect of Virtual Tutoring by Assessment Language

	Math		Reading	
	MAP	STAAR	MAP	STAAR
	(1)	(2)	(3)	(4)
(A) English Language Assessments				
ITT Effect	0.009	-0.001	-0.014	-0.095***
	(0.028)	(0.026)	(0.034)	(0.034)
Observations	2471	2371	1817	1830
(B) Spanish Language Assessments				
ITT Effect	-	0.109	0.142	-0.092
	-	(0.168)	(0.098)	(0.171)
Observations	-	146	148	165

Note: This table presents the intent-to-treat effects of being randomized into either math or reading tutoring on achievement in that subject, by baseline demographic traits. All models include controls for baseline characteristics, pre-intervention test scores, an indicator for assignment to virtual tutoring in the other subject, and randomization strata. * p<0.10, ** p<0.05, *** p<0.010.

Table A9. Estimates of the Intent-to-Treat (ITT) Effect of Virtual Tutoring by Grade

	Math		Reading	
	MAP	STAAR	MAP	STAAR
	(1)	(2)	(3)	(4)
(A) Grade 3 Only				
ITT Effect	0.071	0.012	-0.114	-0.021
	(0.075)	(0.073)	(0.081)	(0.068)
Observations	560	564	464	465
(B) Grade 4 Only				
ITT Effect	0.007	-0.060	-0.100	-0.188**
	(0.061)	(0.059)	(0.091)	(0.073)
Observations	420	422	337	342
(C) Grade 5 Only				
ITT Effect	0.024	0.009	0.072	-0.159***
	(0.053)	(0.059)	(0.075)	(0.058)
Observations	520	533	443	451
(D) Grade 6 Only				
ITT Effect	-0.080	0.051	0.022	0.064
	(0.078)	(0.081)	(0.116)	(0.112)
Observations	287	296	220	222
(E) Grade 7 Only				
ITT Effect	-0.008	-0.022	-0.088	-0.004
	(0.072)	(0.063)	(0.086)	(0.086)
Observations	334	342	235	241
(F) Grade 8 Only				
ITT Effect	-0.088	-0.004	0.113	-0.131
	(0.086)	(0.086)	(0.108)	(0.095)
Observations	350	360	266	274

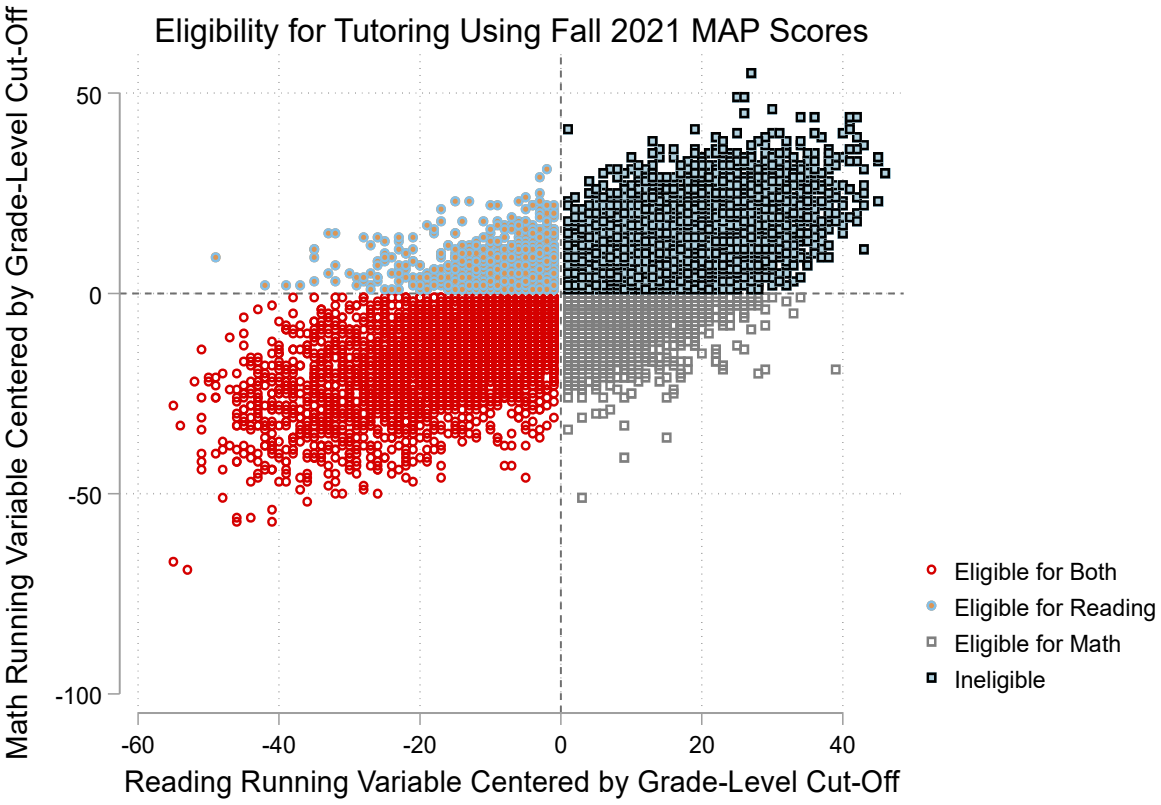
Note: This table presents the intent-to-treat effects of being randomized into either math or reading tutoring on achievement in that subject, by student grade. All models include controls for baseline characteristics, pre-intervention test scores, an indicator for assignment to virtual tutoring in the other subject, and randomization strata. * p<0.10, ** p<0.05, *** p<0.010

Table A10. Intent-to-Treat (ITT) Effects of Virtual Tutoring by Block-Level Dosage Threshold

	40% (12 hrs) (1)	45% (13.5 hrs) (2)	50% (15 hrs) (3)	55% (16.5 hrs) (4)	60% (18 hrs) (5)	65% (19.5 hrs) (6)
(A) Impacts of Math Tutoring on Math MAP Achievement						
ITT Effect	−0.014	0.014	−0.024	−0.069	−0.106	−0.314
	(0.066)	(0.067)	(0.072)	(0.079)	(0.101)	(0.204)
Observations	395	339	307	203	144	54
Blocks	12	11	10	7	5	2
(B) Impacts of Math Tutoring on Math STAAR Achievement						
ITT Effect	0.005	0.013	0.024	−0.007	−0.091	−0.229
	(0.068)	(0.074)	(0.079)	(0.095)	(0.118)	(0.263)
Observations	399	343	311	206	145	55
Blocks	12	11	10	7	5	2
(C) Impacts of Reading Tutoring on Reading MAP Achievement						
ITT Effect	0.019	0.076	0.123	0.256**	0.267*	0.160
	(0.073)	(0.081)	(0.104)	(0.123)	(0.140)	(0.190)
Observations	484	400	255	170	133	67
Blocks	17	14	10	7	6	4
(D) Impacts of Reading Tutoring on Reading STAAR Achievement						
ITT Effect	−0.159**	−0.096	−0.059	−0.098	−0.047	0.001
	(0.065)	(0.073)	(0.093)	(0.125)	(0.141)	(0.173)
Observations	488	400	256	170	132	67
Blocks	17	14	10	7	6	4

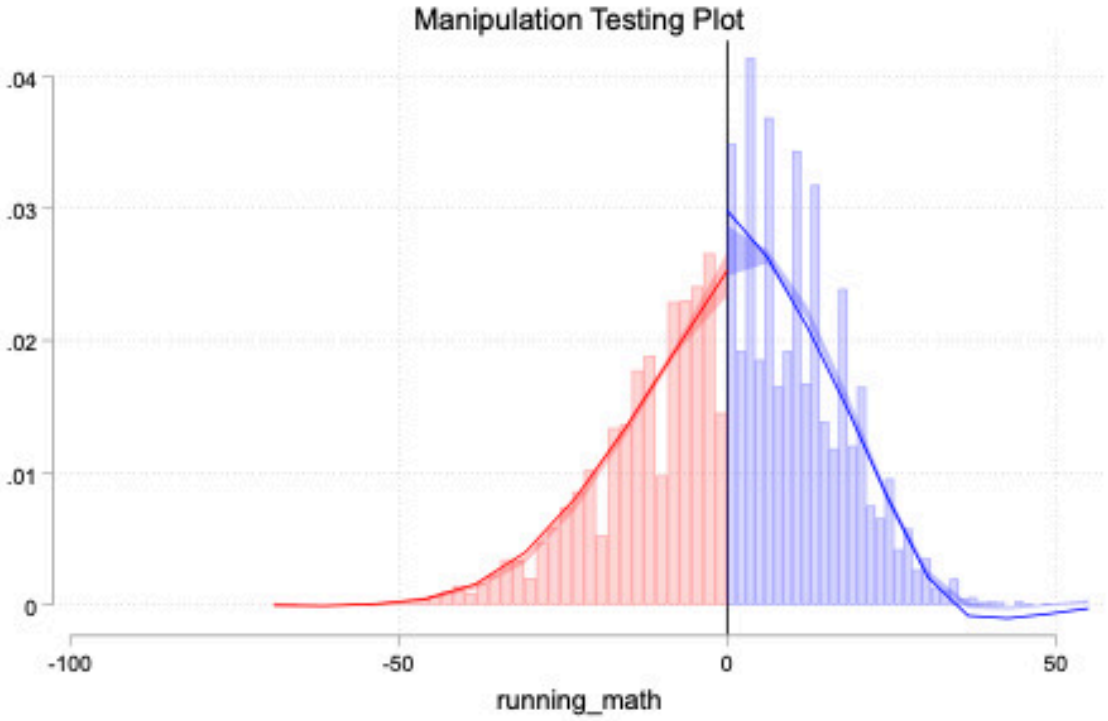
Note: This table presents the intent-to-treat effects of being randomized into either math or reading tutoring on achievement in that subject, restricting to randomization blocks where treated students averaged at least the indicated percentage of the recommended 30 hours of tutoring. All models include controls for baseline characteristics, pre-intervention test scores, an indicator for assignment to virtual tutoring in the other subject, and randomization strata fixed effects. Heteroskedasticity robust standard errors are in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Figure A1. Distribution of 2021 Assignment Scores for HB4545 Intervention Eligibilit



Note: This figure plots district-wide (N=13,187) fall 2021 MAP scores for math and reading (including Spanish language reading assesments) centered about the grade-specific intervention-eligibility thresholds.

Appendix Figure A1. Density Tests for 2021 Assignment Variable to HB 4545 (BaU or Virtual)



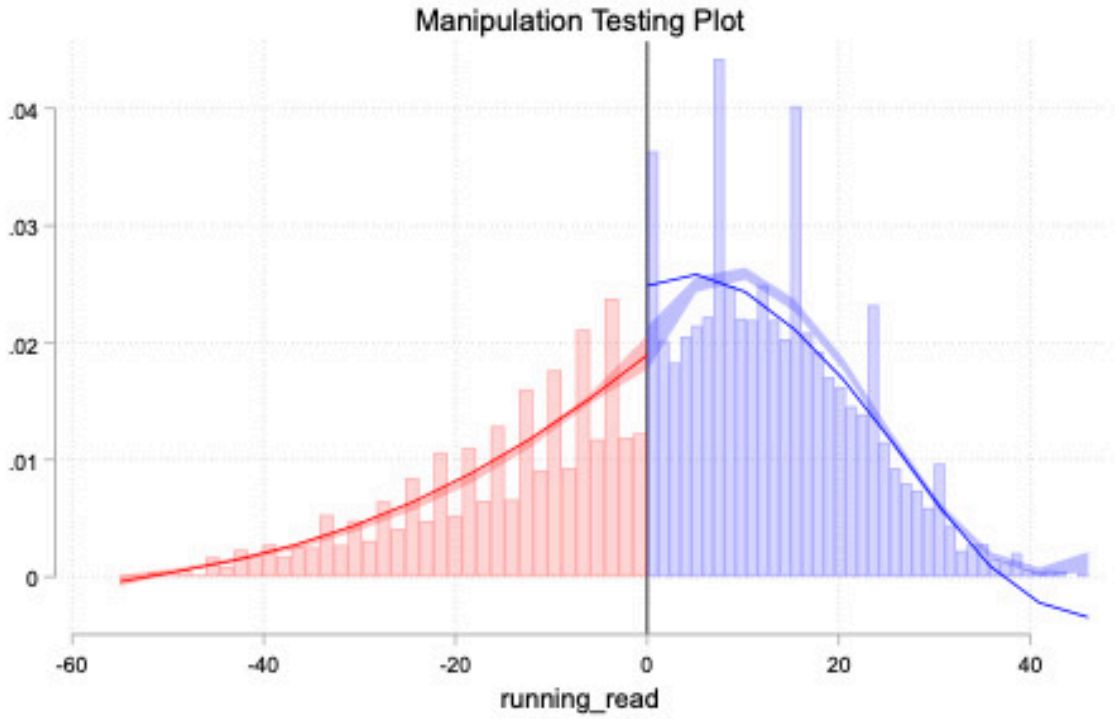
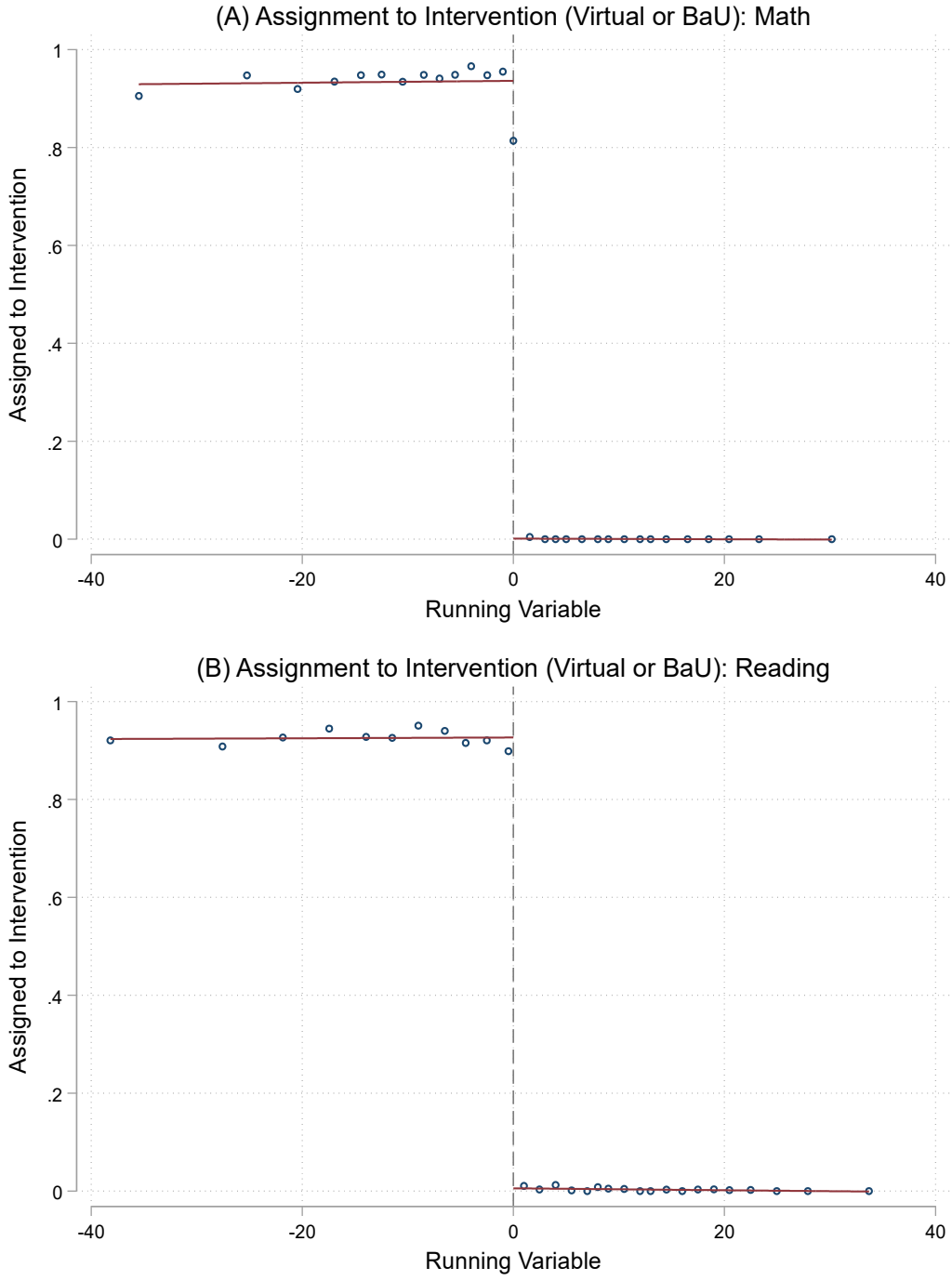


Figure A3. Treatment Status for HB 4545 Intervention by 2021 Assignment Variable



Note: These show the relationship betwee district-indicated eligibility for academic intervention (i.e., aligned with HB 4545) and fall 2021 math (N=13,500) and reading (N=13,453) map scores, centered about grade-specific cut-scores. Bin number = 30.

Figure A4. Distribution of Average Dosage by Randomization Block

