When interventions don’t move the needle: Insights from null results in education research

Elizabeth Huffaker
University of Florida

Carly D. Robinson
Stanford University

Emanuele Bardelli
Santa Rosa City Schools

Sara White
Vanderbilt University

Susanna Loeb
Stanford University

As school districts focus on improving learning, they can learn not only from when and where interventions work—but also from why they sometimes do not. Policymakers widely embraced high-impact tutoring as an evidence-supported strategy to address learning delays from the COVID-19 pandemic. However, scaling these promising practices can be difficult, and not all implementations will be effective. Many districts have turned to third-party virtual tutoring providers to deliver student supports during the school day. Using random assignment, we evaluate the impacts of one such program for 3rd through 8th grade students in a suburban Texas school district. Compared with students assigned to the comparison interventions, we find no effect of assignment to virtual tutoring on math achievement, and, for reading, we find a moderate negative effect on the state end-of-year assessment (i.e., -0.09 SD) and no effect on a low-stakes exam. Drawing from frameworks for interpreting null or unexpected results in education experiments, we find further evidence of subject-specific heterogeneity in the implementation and efficacy and identify coverage of standards-aligned material as a moderator of estimated effectiveness relative to “business-as-usual” interventions. This paper offers strategies to identify factors contributing to null or unexpected results and highlights implications for designing policy-relevant studies to assess educational interventions.

VERSION: September 2025

Suggested citation: Huffaker, Elizabeth, Carly D. Robinson, Emanuele Bardelli, Sara White, and Susanna Loeb. (2025). When interventions don’t move the needle: Insights from null results in education research. (EdWorkingPaper: 25-1295). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/58dd-6d02

When interventions don’t move the needle: Insights from null results in education research

By: Elizabeth Huffaker, Carly Robinson, Emanuele Bardelli, Sara White, and Susanna Loeb

Abstract: As school districts focus on improving learning, they can learn not only from when and where interventions work—but also from why they sometimes do not. Policymakers widely embraced high-impact tutoring as an evidence-supported strategy to address learning delays from the COVID-19 pandemic. However, scaling these promising practices can be difficult, and not all implementations will be effective. Many districts have turned to third-party virtual tutoring providers to deliver student supports during the school day. Using random assignment, we evaluate the impacts of one such program for 3rd through 8th grade students in a suburban Texas school district. Compared with students assigned to the comparison interventions, we find no effect of assignment to virtual tutoring on math achievement, and, for reading, we find a moderate negative effect on the state end-of-year assessment (i.e., -0.09 SD) and no effect on a low-stakes exam. Drawing from frameworks for interpreting null or unexpected results in education experiments, we find further evidence of subject-specific heterogeneity in the implementation and efficacy and identify coverage of standards-aligned material as a moderator of estimated effectiveness relative to “business-as-usual” interventions. This paper offers strategies to identify factors contributing to null or unexpected results and highlights implications for designing policy-relevant studies to assess educational interventions.

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.

Tutoring, defined as one-on-one or small group instruction, has become a leading strategy to support student learning. Most states used dollars dedicated to pandemic relief from the CARES Act and ESSER to fund tutoring programs (LePage & Jordan, 2021; U. S. Government Accountability Office, 2024); and a large body of causal research finds in-person tutoring has consistently large and positive effects on student test scores (Dietrichson et al., 2017; Slavin et al., 2011; Nickow et al., 2024). However, 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 (National Student Support Accelerator, 2023; White et al., 2022). A new meta-analysis finds that the benefits of tutoring tend to decline as districts expand their programs and attempt to reduce costs and streamline implementation by relaxing some program features (Kraft et al., 2024).

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 for districts of tutor recruitment

(White et al., 2021). 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. Only recently are we learning about the efficacy of virtual tutoring programs (Carlana & Ferrara, 2024; Kraft et al., 2024; Neitzel & Storey, 2024; Ready et al., 2024; Robinson et al., 2024). Carlana & Ferrara (2024) found that virtual tutoring by university students increased middle school students' scores on a researcher-administered test of math, Italian, and English by a magnitude comparable to in-person tutoring impacts (ES=0.26 SD). Other studies have identified effects that are positive but smaller than those consistently found in in-person tutoring programs (Ready et al., 2024; Robinson et al., 2024), with some indications that low dosage and technological challenges may drive the attenuated effects (Kraft et al., 2024). Robinson et al. (2024) 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. (2024) demonstrated that assignment to early literacy tutoring in 1st-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.

We contribute novel evidence to the scarce literature for this increasingly prevalent intervention by conducting 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 other recent studies, 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).

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 the treatment-control contrast, or the intervention mechanisms), and (4) implementation challenges or limitations. The rich data afforded by our research partnerships with both the tutoring provider and focal school district allow us to investigate these explanations. We address the empirical concerns statistical precision and research design in greater detail in our methods and results sections and confirm that our experiment featured successful randomization and was 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 considerations – our understanding of the counterfactual and the quality of program implementation – we draw from theory and empirical research on tutoring, which highlight four common reasons why randomized controlled trials may fail to detect effects

[1] Registry ID: 13040.1v1 at Registry of Efficacy and Effectiveness Studies.

from ostensibly well-designed tutoring programs: (a) control group supports (i.e., the counterfactual), (b) spillovers to the control group, (c) insufficient treatment group dosage, and (d) misalignment of program content with assessed material. While the effectiveness of *specific* tutoring interventions is moderated by myriad features (e.g., modality, tutor-student ratio, quality of curricular materials, tutor training etc.) these four issues surface across a wide range of tutoring evaluations (Nickow et al., 2024).

First, for a study to provide clear information on program effectiveness, researchers need to understand the experiences of the control group, as well as that of the treatment group. The growing practice of providing tutoring during the conventional school day has expanded participation (e.g., Bhatt et al., 2024) 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 sample is restricted to students below this threshold, control students also received supplementary instruction.

Second, researchers and practitioners frequently hypothesize 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 case load for the classroom teacher. Berlinski et al., (2022) find 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).

Third, a broad range of interventions have struggled to deliver tutoring in sufficient quantities (i.e., dosage). Even a well-designed program will fail to noticeably impact learning if students receive little exposure to it. Effective programs tend to provide tutoring for at least 30-minutes, three times a week, but many programs do not consistently reach those benchmarks (Huffaker et al., 2025; Robinson & Loeb, 2021; Bhatt et al., 2025). Limitations in provider capacity and high-absenteeism among tutoring-eligible students significantly drive this pattern (Nickow et al., 2024).

Fourth and finally, even appropriately dosed tutoring may have negligible measured impacts if covered content is either poorly targeted to student need or unrelated to assessed skills (TNTP, 2025). For instance, a randomized evaluation of a high-frequency, in-person, early-grade literacy program found anomalously null results even at higher dosage levels. Deeper inspection of the program revealed that the curriculum was so rigid tutors had no opportunity to adapt or differentiate the curriculum across students based on need, was misaligned with the assessment, and emphasized repetition of lower-level material over advancement to higher-level skills (Huffaker et al, 2025).

We formulate and address a secondary research question (RQ2) in order to explore these dynamics. To what extent do components of the null results mediator framework explain our topline findings? Table 1 summarizes the scope of this analysis. To interrogate the empirical relevance of control group supports and spillovers to the control group, we explore whether students in the control group either benefit directly from BaU interventions mandated by HB 4545 or benefit indirectly from spillovers generated by the virtual tutoring program. Next, to understand the role of dosage and content alignment in moderating tutoring efficacy among the treatment group, we examine the associations between session quantity, content coverage, 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. We find some evidence that spillovers and quantity of tutoring sessions may contribute to these results but are not primary drivers. Instead, our analysis suggests the primary shortcoming of this program was that the provider failed to align tutored content with the breadth and emphasis of state standards. This finding implies that districts may benefit from placing greater emphasis on the alignment of grade-level material when vetting the curricula of third-party intervention programs. 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.

Finally, this study 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 HB4545 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 (Texas Education Agency, 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 grades three through eight across 28 schools. The district identified 5,349 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).

However, tutoring was not implemented as intended. First, 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 fall to spring semester. Second, while selected students were originally scheduled to receive tutoring five days a week, attendance logs indicate that 81% of students assigned to tutoring attended three or fewer days per week. Conversations with teachers and principals suggest 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—substantially less than the 30 hours mandated by HB 4545. Third, while virtual tutoring was conducted 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. Finally, in terms of content delivery strategy, students were to receive tutoring on a subset of grade-level standards that they had not

[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 for assignment.

yet mastered. Their progress was intended to be assessed by the provider to personalize their learning.

Estimates of the effectiveness of the tutoring will also depend on the experiences of the control group. 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 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.

Data and Sample

Data and measures

Data for this study come from the district's student information system and the tutoring provider's session logs. 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 (ELL) status, 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 2 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. 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. The district administers these tests to all students grades 3 through 8 and are meant 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. For instance, STAAR scores may be used to identify students for retention in elementary school (O'Hara et al., 2023). 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. In addition, we use beginning-of-year (BOY) MAP achievement in both math and reading in exploratory quasi-experimental analysis of the effects of eligibility for small group intervention using a regression discontinuity (RD) design, and we use STAAR data to characterize the frequency with which specific standards (i.e., TEKs) are tested.

[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.

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 spring 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 analytical sample to 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.

Table 2 reports summary statistics for the pooled sample of students (N=3,398) identified as eligible for math (N=2,843) and/or reading (N=2,237) tutoring under District implementation of HB 4545 who also attended schools-grade blocks with virtual tutoring. Panel A presents baseline student traits and prior achievement, panel B presents average exposure to virtual tutoring, and panels C and D aggregates 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 ELLs. 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 (ITT) 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 be delivered to eight students in each subject in each elementary grade 3-5 and sixteen students in each 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, average expected outcomes in the absence of treatment for virtual tutoring and BaU interventions should be equal. While we cannot test this assumption directly, we can compare conditional mean differences in baseline observable traits across treatment conditions. Table 3 summarizes this assessment of whether randomization was successful. Columns 1 and 5 report the mean and standard deviation for students assigned to BaU interventions during the intervention period (Columns 2 and 6), and for those assigned to the virtual tutoring program (Columns 4 and 7). After controlling for randomization strata, students randomly assigned to participate in the virtual tutoring program are statistically indistinguishable to students assigned to the BaU condition on baseline demographic characteristics and test scores. The results are consistent with successful stratified randomization and the absence of systematic differences between treatment and control students.

Empirical Strategy

Our main, preregistered analysis estimates the effects of being assigned to the 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.

Estimation of Intent-to-Treat Effects

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

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

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 "spillover" ITT effect of being eligible for math tutoring but being assigned to reading tutoring instead. 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 strategies, and γ_s are school-grade-eligibility fixed effects that capture our randomization structure.

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.

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, language learner, and special education category. 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.

Mediator 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 (a) control group supports and (b) spillovers, or by program characteristics, including (c) quantity of tutoring received and (d) the alignment of tutored content with the assessments. 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.

[4] Scale scores are equivalent across administration languages according to the testing providers (NWEA).

Counterfactual Supports & Spillovers

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 spillovers benefits from treatment. For this study, the control condition is students who receive other interventions under HB4545. These interventions may be benefiting students. To test this possibility – and answer RQ2 with respect to parts A and B (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 to utilize a regression discontinuity (RD) design. This strategy allows us to capture the effect of *just* qualifying for math or reading support. We apply this approach across key sub-samples to draw comparative insights. Figure 1 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. First, we estimate *aggregate* "HB 4545 effects" using all these observations. This analysis indicates whether the district's implementation of the state law impacted average math and/or reading achievement among students near these thresholds.

Next, we use the RD strategy to explore the counterfactual of our virtual tutoring RCT. efficacy of BaU interventions among our control group. To do so, we restrict our sample to the schools that participated in the virtual tutoring experiment. Drawing from these observations, we separately estimate effect of marginal eligibility for 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, this analysis yields positive ITT estimates about the district's HB 4545 threshold among for both groups, then our findings would indicate that interventions received by students in both the treatment and the control group were comparably effective.

Finally, we use observations from schools that did *not* participate in the virtual tutoring experiment to draw inferences about spillovers to the control group. Specifically, we consider students near the intervention eligibility threshold from 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 on the control-group students – for instance, by freeing up teachers to work with smaller intervention groups – then we would expect HB 4545 eligibility to drive stronger learning gains in schools with virtual tutoring even among students not receiving virtual tutoring. In the absence of spillovers, we would expect similar RD estimates for students eligible to receive BaU interventions in schools with and without virtual tutoring. Because schools did select into virtual tutoring, we understand such comparisons could merely indicate differences in schools' capacity to effectively implement small-group interventions – especially with little lead time and to many students – may vary. So, we repeated our RD estimation of ITT effects about the threshold on a final subset of schools, those which never intended to implement virtual tutoring. This analysis will provide a point of comparison for the efficacy of HB-4545 associated interventions at both virtual tutoring and non-virtual-tutoring schools.

We estimate these various ITT effects (i.e., $\delta_1,$below) of HB 4545 eligibility for students near the threshold using the following equation:

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

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 $Below_{is} \times Running_i$ allows the slope we estimate for the linear function fit to the running variable to vary on either side of the cuts core. α_g controls for grade-level membership and X_i is a vector of baseline student demographic characteristics. ϵ_{ig} is a heteroskedasticity robust standard error. Because tutoring take-up closely – although not perfectly – tracks tutoring eligibility (Figure 2) we center the reduced form (i.e., ITT) RD estimates rather than performing a two-stage least squares regression to report the local average treatment effect at the threshold. For our primary results, we use bandwidths selected to minimize mean-squared error using the Calonico et al. (2020) procedure.

Following (Lee & Lemieux, 2010) we also present ITT RD estimates using the main sample across a variety of alternative bandwidths and with unrestricted data in the appendix. We also follow best practices in attending to core assumptions and research design elements, which includes establishing the integrity of the forcing variable (Figure A1) and examining covariate balance across the threshold (Table A1) (What Works Clearinghouse, 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 threshold (see Figure A1 and Table A1).

Dosage and Alignment of Tutored Content

We use a variety of approaches to explore the remainer of RQ2 – could shortcomings of program implementation in terms of dosage (i.e., because students did not receive tutoring until the second semester) and alignment (i.e., because the provider expressed that correspondence

[5] We use a structural break method (Baum, 2004) to verify these cut points in the data.

between tutored and assessed material was still developing) explain our main results? To examine aggregate dosage, we use session-log data to measure the effects of virtual tutoring scaled by hour. We then use random assignment to treatment as an instrument for quantity of virtual tutoring received. A two-stage least squares estimation strategy scales the ITT estimates from RQ1 by the difference in average minutes of virtual tutoring received by treatment assignment. If noncompliance with virtual tutoring assignment was high, this approach could reveal virtual tutoring effects not precisely captured by the ITT effects.

Next**,** to begin the role of alignment in moderating virtual tutoring effects, we estimate the relationship between virtual tutoring and achievement at the TEK (i.e., a Texas learning standard) level 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 if it addressed a tutored versus untutored standard.

Formally, we estimate the following within-student, within-TEK fixed effects regression:

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

where $Correct_{qti}$ is a binary variable that is 1 when student *i* answers question q correctly and 0 otherwise,$Tutored_{ts}$is an indicator variable that is 1 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 likelihood that a student will answer a question correctly if they have received virtual tutoring on it versus if they have 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 cover more content do 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 idiosyncratically 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 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 assessing proficiency in 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 a TEK as a measure of TEK difficulty. We then disaggregate TEK coverage by subject, grade-level, an out-of-sample measure of difficulty, frequency of assessment, and state-indicated priority level.

Results

Effects of a Virtual Tutoring Experiment

Table 4 reports the effects of being selected for the virtual tutoring program when compared to students receiving the in-person, BaU supports. 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 ($p < 0.05$). 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 A2 and A3 in the 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 cross-subject spillover or crowding out. The impact of assignment to reading tutoring on math achievement (and vice versa) are null in all Table A4 specifications except for a marginally significant negative math STAAR effect detected when baseline score controls are used.

We next consider other sources of effect heterogeneity (see Tables A5 through A7). 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 A7, Column 4, Panels B and C for more details).

Mediator Analyses (RQ2)

We next report the results of our supplementary quasi-experimental and descriptive analysis to better understand the null and negative estimated effects of this virtual tutoring experiment.

Counterfactual Supports & Spillovers: The Effects of HB 4545 Intervention Eligibility

Table 5 presents the results of our regression discontinuity estimates of the effect of eligibility for any intervention (i.e., BaU interventions *or* virtual tutoring) across the district and the subsets previously described. Broadly, these results suggest that the districts' 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) coefficient of eligibility on STAAR scores (Column 6).

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. The relevant estimated coefficient magnitudes among control group students *are* positive (0.54-0.90) but imprecisely estimated. And among schools in the experimental sample (Columns 6 and 7) the gap between RD ITT effects for treatment versus control students is only -0.046 SD, half of experimental ITT effect. This could indicate that the *relative* efficacy of BaU versus virtual tutoring on STAAR reading scores may have been driven by students farther from the eligibility threshold.

In general, however, while we cannot rule out that *very low* achieving students who are not capture in our RD design benefited from the interventions, our results provide evidence that nearly on-level students would have been better off in math, and equally well off in reading, without the prescribed HB 4545 interventions whether they were BaU or virtual tutoring. Moreover, these results – and those limited to schools in the experimental sample (Columns 2, 3, 6 & 7) strongly suggest our null and negative experimental results are not explained by control and treatment group students both receiving effective interventions.

Our results are weakly consistent with the positive spillover hypothesis, 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. The relevant comparisons, across conceptually similar schools (i.e., among those that intended to take-up tutoring, those that eventually did versus those that did not), are therefore Columns 3 versus 4 and 7 versus 8. We do find that the effect of intervention eligibility on math scores *is* approximately 0.07 to 0.08 SD *more positive* in schools with virtual tutoring[6] – though the estimated magnitudes are not statistically different from each other. For reading scores, evidence of spillovers is even more slight. We observe a nonsignificant positive difference (i.e., 0.063) between control group student achievement in schools with versus without virtual tutoring in average MAP scores but only a 0.36 difference on the STAAR assessment.

Tutoring Dosage and Alignment

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.

We directly estimate an average hourly treatment on the treated effect (ATE) in the regressions summarized by Table 6. 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

[6] Using MAP scores, the effect of marginal eligibility for HB 4545 among control group students was -0.055 in schools that offered virtual tutoring (Table 5 Panel A Column 3) and -0.139 in schools that did not offer virtual tutoring (Table 5 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 in schools that offered virtual tutoring (Table 5 Panel A Column 3) and -0.091 in schools that did not offer virtual tutoring Table 5 Panel A Column 4), a difference of 0.067.

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 6, Panel A). The effect of receiving any tutoring and of receiving an additional hour of tutoring are null for both math assessments and on the reading MAP test.

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 learning at some level of sufficient dosage, this threshold was not met. Separately, we estimated null results when we compared within-student growth from the winter to spring MAP administration by treatment assignment. That is, isolating score changes during the period in which the treatment was implemented also indicated null program impacts for both math and reading on the MAP tests.

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 6 Column 4). From these analyses we see that merely receiving some amount of virtual tutoring does not confer academic benefits. This could be because a certain threshold of sufficient dosage must be reached for gains to be observed (Bhatt et al., 2025; Huffaker et al., 2025) but also could indicate that time in virtual tutoring is simply not promoting mastery of assessed topics.

We therefore turn to a final vein of exploratory analysis to better understand *why* additional instructional time in virtual tutoring does not necessarily increase academic performance and examine the academic implications of instructional alignment by virtual tutors. We explore whether patterns of coverage for specific, assessment-aligned topics during virtual tutoring plausibly explain observed outcomes.

Table 7 reports the results from our item-level analysis. Columns 1 and 3 present ITT estimates on the probability of answering a given 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 asked about a tutored TEK. For math, we estimate that coverage of a TEK is associated with a 3-percentage point increasing in the likelihood of answering the corresponding questions correctly. In reading, the analogous coefficient estimate is statistically insignificant, though positive (1.7-percentage points). For reading, however, covering TEKs in virtual tutoring does detectably improve student performance on aligned questions. While these are not strictly causal associations – the coverage of standards was not randomly assigned within students – the inclusion of TEK fixed effects controls for systematic variation in difficulty or complexity across standards. 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. 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 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. Taken together, these within-student-and-standard estimates suggest that the null effects in math and negative effects in reading on the STAAR may be driven by non-covered TEKs. Both the coverage of standards *and* the subject area of virtual tutoring appear to be relevant in explaining our experimental results.

To further explore the relationship between TEK coverage and student performance, we disaggregate each by grade-level. Figure 4 Panel A presents the average share of total TEKs coverage 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, summarized in 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 A7). 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 in promoting state test achievement, even on tutored material.

[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.

Conclusion

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 equivalent, and in some cases 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 effects 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.

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 8 summarizes these results of the mediator 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. 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.

In general, the districts’ implementation of small group interventions – encompassing the randomized virtual tutoring program but also a variety of other practices – did not promote detectable student gains to achievement on either a low or high-stakes exam. Student eligibility for supplemental math intervention, in particular, appears to have reduced student math learning on the low-stakes exam. This initially appears in tension with our constellation of more promising observations about math tutoring - that is, math virtual tutoring was reasonably well aligned with standards emphasized by the state, and students did better on questions covering TEKs they received tutoring on. It raises the possibility that despite a 0.8 overall score correlation, MAP and STAAR assessments are misaligned on some topics emphasized by the state and in tutoring, or that students near the grade-level threshold are uniquely harmed by placement in intervention classes (e.g., through negative peer effects or missing out on other grade-level instructional supports).

Taken together, these findings contribute to both the research base and practical understanding of targeted educational interventions, such as in-school virtual tutoring. 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. As educational research moves towards a deeper understanding of not just average program effects, but of features of successful intervention *implementation*, unpacking the “black box” of null results from studies like this can still add significant value to the knowledge base – provided sufficiently detailed data and knowledge of relevant coincident policies are available.

Second, these observations carry weighty implications for policymaking 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 a 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 on a diagnostic test, virtual tutoring may generate equivalent academic benefits to BaU interventions for students. As such, virtual tutoring could be an appropriate addition to a district’s suite of interventional strategies.

More fundamentally, the failure of the statewide mandate to provide small group interventions to all below grade-level students to boost achievement carries implications for tutoring policymaking. Implementing personalized, small-group interventions on the timeline and scale required by HB 4545 would strain feasibility for most districts. In our partner district, for example, approximately 43% of all third through eighth graders in our partner district were eligible for tutoring and the bill was passed and enacted a mere two to 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.

References

Abadie, A. (2020). Statistical nonsignificance in empirical economics. *AER: Insights, 2(2)*, 193-208.

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

Baum, Christopher F. 2015. "ZANDREWS: Stata Module to Calculate Zivot-Andrews Unit Root Test in Presence of Structural Break." *Statistical Software Components*.

Berlinski, Samuel, Matias Busso, and Michele Giannola. 2022. "Helping Struggling Students and Benefiting All: Peer Effects in Primary Education."

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*. https://educationlab.uchicago.edu/resources/personalized-learning-initiative-interim-report-findings-from-2023-24/

Bloom, Howard S., Carolyn J. Hill, Alison Rebeck Black, and Mark W. Lipsey. 2008. "Performance Trajectories and Performance Gaps as Achievement Effect-Size Benchmarks for Educational Interventions." *Journal of Research on Educational Effectiveness* 1(4):289–328. doi:10.1080/19345740802400072.

Burch, P., Good, A., & Heinrich, C. (2016). Improving access to, quality, and the effectiveness of digital tutoring in k–12 education. *Educational Evaluation and Policy Analysis*, *38*(1), 65–87. https://doi.org/10.3102/0162373715592706

Burgess, S. M., Rawal, S., & Taylor, E. S. (2022). *Teachers' use of class time and student achievement* (NBER Working Paper No. 30686). National Bureau of Economic Research. https://doi.org/10.3386/w30686

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

Carlana, M., & La Ferrara, E. (2024). *Apart But Connected: Online Tutoring, Cognitive Outcomes, and Soft Skills* (SSRN Scholarly Paper 4771248). Social Science Research Network. https://papers.ssrn.com/abstract=4771248

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, Esther, Pascaline Dupas, and Michael Kremer. 2011. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." *American Economic Review* 101(5):1739–74. doi:10.1257/aer.101.5.1739.

von Hippel, Paul 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* 01623737241239677. doi:10.3102/01623737241239677.

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.*

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*(2), 615–635. https://doi.org/10.1016/j.jeconom.2007.05.001

Institute of Education Sciences. 2022. "What Works Clearinghouse Procedures and Standards Handbook, Version 5.0."

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, Matthew A., Schueler, Beth E., and Falken, Grace. 2024. "What Impacts Should We Expect from Tutoring at Scale? Exploring Meta-Analytic Generalizability." doi: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?* https://www.the74million.org/article/how-are-states-spending-their-covid-education-relief-funds/

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

Neitzel, Amanda J., and Nathan Storey. 2024. "Air Reading: A Randomized Evaluation of a Virtual Tutoring Model."

Ready, Douglas D., McCormick, Sierra G., & Shmoys, Rebecca J. (2024). *The Effects of In-School Virtual Tutoring on Student Reading Development: Evidence from a Short-Cycle Randomized Controlled Trial*. https://doi.org/10.26300/569P-WZ78

Riis-Vestergaard, M. (2023, October 4). *So, you got a null result. Now what?* The Abdul Latif Jameel Poverty Action Lab (J-PAL). https://www.povertyactionlab.org/blog/10-4-23/so-you-got-null-result-now-what

Robinson, C. D., & Loeb, S. (2021). *High-impact tutoring: State of the research and priorities for future learning* (EdWorkingPapers No. ai21-384). https://www.edworkingpapers.com/ai21-384

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

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

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

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

Table 1. Summary of Mediator Analysis (RQ2)

RQ2 Component	Mediator Category	Mediator Description
A	Features of the counterfactual	Control group supports
B		Spillover to control group
C	Features of the treatment	Quantity of tutoring (dosage)
D		Tutoring coverage (alignment)

Table 2. 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 Americai	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 Receivec	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, *** <0.010.

Table 3: 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	0.513	0.556	-0.013	0.417	0.420	0.407	-0.004
	(0.009)	(0.010)	(0.021)		(0.010)	(0.012)	(0.021)	
FRL	0.897	0.899	0.886	-0.000	0.905	0.903	0.909	-0.017
	(0.006)	(0.006)	(0.013)		(0.006)	(0.007)	(0.012)	
American Indian/Native American	0.014	0.012	0.021	-0.007	0.013	0.013	0.014	-0.002
	(0.002)	(0.002)	(0.006)		(0.002)	(0.003)	(0.005)	
Asian	0.007	0.008	0.003	0.005	0.011	0.011	0.010	-0.000
	(0.002)	(0.002)	(0.002)		(0.002)	(0.003)	(0.004)	
Black	0.410	0.412	0.402	-0.005	0.380	0.399	0.323	0.051*
	(0.009)	(0.010)	(0.021)		(0.010)	(0.012)	(0.020)	
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)		(0.001)	(0.001)	(0.002)	
Multiracial	0.020	0.019	0.024	-0.008	0.017	0.017	0.014	0.004
	(0.003)	(0.003)	(0.006)		(0.003)	(0.003)	(0.005)	
White	0.042	0.041	0.045	-0.003	0.046	0.041	0.058	-0.003
	(0.004)	(0.004)	(0.009)		(0.004)	(0.005)	(0.010)	
English Learner	0.329	0.326	0.343	-0.014	0.385	0.377	0.411	-0.019
	(0.009)	(0.010)	(0.020)		(0.010)	(0.012)	(0.021)	
Special Education	0.196	0.208	0.150	0.029	0.232	0.241	0.205	0.011
	(0.007)	(0.009)	(0.015)		(0.009)	(0.010)	(0.017)	
N Students	2843	2271	572	2843	2237	1665	572	2237
(B) Pre-Intervention Test Scores								
MAP Fall 21 Read (sd)	-0.522	-0.566	-0.353	-0.018	-0.963	-0.981	-0.910	-0.009
	(0.017)	(0.019)	(0.037)		(0.014)	(0.016)	(0.026)	
N Students	2809	2240	569		2235	1665	570	
MAP Fall 21 Math (sd)	-0.789	-0.808	-0.712	-0.016	-0.701	-0.754	-0.548	-0.033
	(0.012)	(0.014)	(0.027)		(0.018)	(0.021)	(0.036)	
N Students	2838	2267	571		2214	1651	563	
F-stat, joint-significance				0.768				0.774
Number of school-grade blocks	57	57	57	57	58	58	58	58

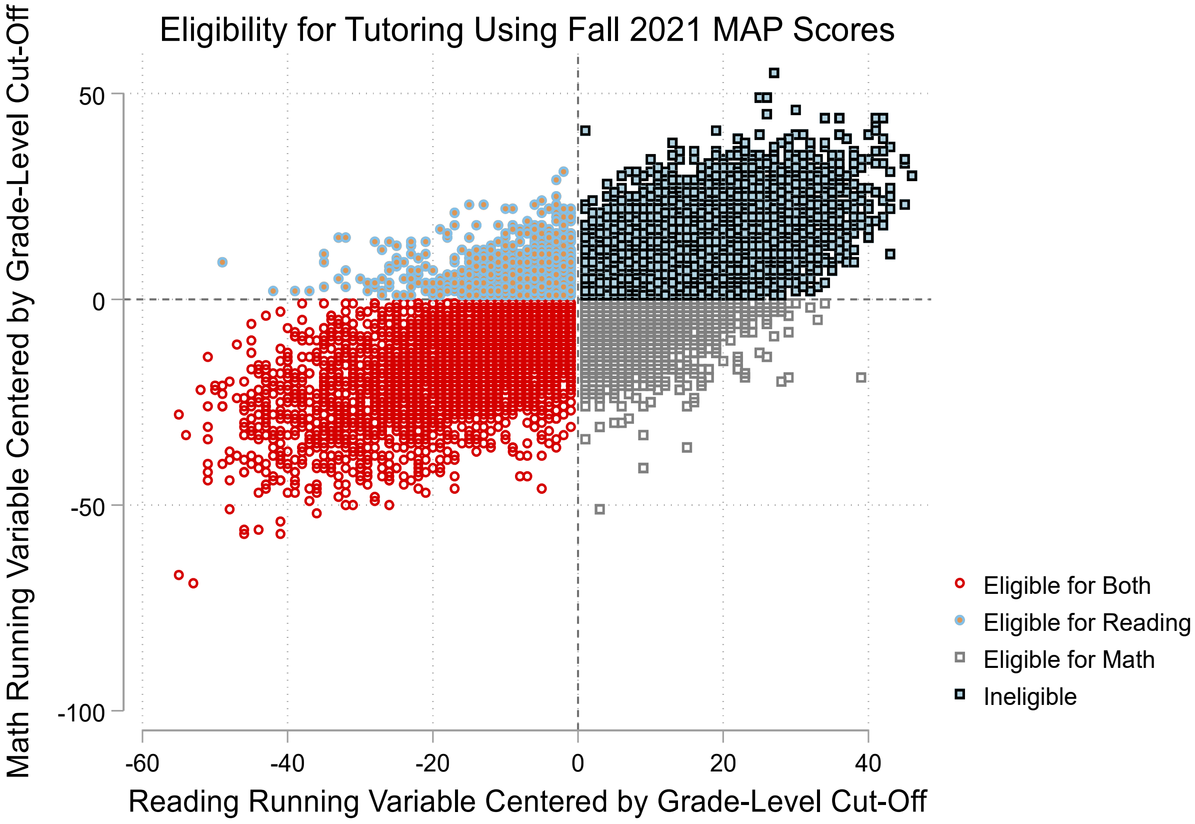
This table reports descriptive statistics for the sample of students identified for tutoring. Columns 1 through 3 report the variable means and standard deviations for each group with respect to math tutoring; Column 4 reports the adjusted difference between the group of students participating in the business-as-usual tutoring program and students participating in the virtual tutoring program after controlling for randomization strata. Columns 5-8 report corresponding information with respect to reading tutoring. The demographic categories used are what is reported in the district's administrative data set and we recognize are not representative of the full range of student identities and experiences. We do not know how students are placed into these categories. All students are marked as either Male or Female. All students are marked as one of the racial/ethnic categories listed in the table or Multiracial. No students are listed in more than one racial/ethnic category. * p<0.10, ** p<0.05, *** p<0.010.

Table 4: 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 Efffct	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 contrc	No	Yes	Yes	No	Yes	Yes
Includes student baseline controls	No	No	Yes	No	No	Yes
Observations	1981	1965	1965	2012	1995	1995

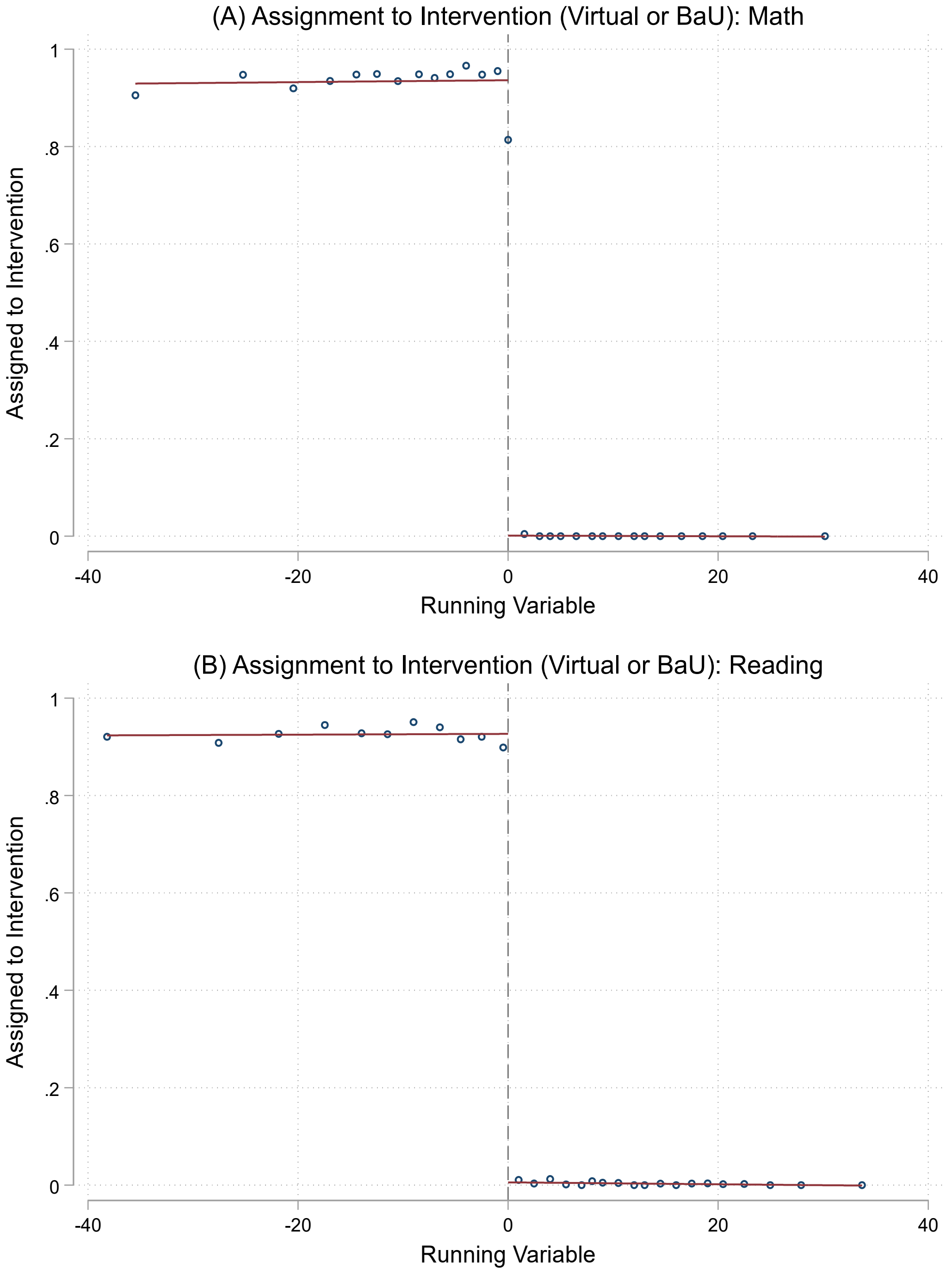
This table presents the intent-to-treat effects of being randomized into either math or reading tutoring on achievement in that subject. Columns 1 through 3 summarize standardized MAP scores while columns 4 through 6 do so for the STAAR assessment. All specifications include fixed effects to control 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. * p<0.10, ** p<0.05, *** p<0.010.

Figure 1. Distribution of 2021 Assignment Scores for HB4545 Intervention Elgibility



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.

Figure 2. Treatment Status for HB 4545 Intervention by 2021 Assignment Variable


(A) Assignment to Intervention (Virtual or BaU): Math

(B) Assignment to Intervention (Virtual or BaU): Reading

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.

Table 5: Regression Discontinuity Estimates on the Effect of Eligibility for Tutoring

	MAP				STAAR			
		Schools Participating in Virtual Tutoring Experiment		Schools Not in Virtual Tutoring Experiment		Schools Participating in Virtual Tutoring Experiment		Schools Not in Virtual Tutoring Experiment
	All HB 4545 Eligible vs. Ineligible	Virtual vs. Ineligible	Control (BaU) vs. Ineligible	HB 4545 Eligible vs. Ineligible at Schools that Intended to Offer Virtual Tutoring	All HB 4545 Eligible vs. Ineligible	Virtual vs. Ineligible	Control (BaU) vs. Ineligible	HB 4545 Eligible vs. Ineligible at Schools that Intended to Offer Virtual Tutoring
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(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 Observatio	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 Observatio	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. * p<0.10, ** p<0.05, *** p<0.010.

Table 6. 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 Tutori	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 Tutorinք	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

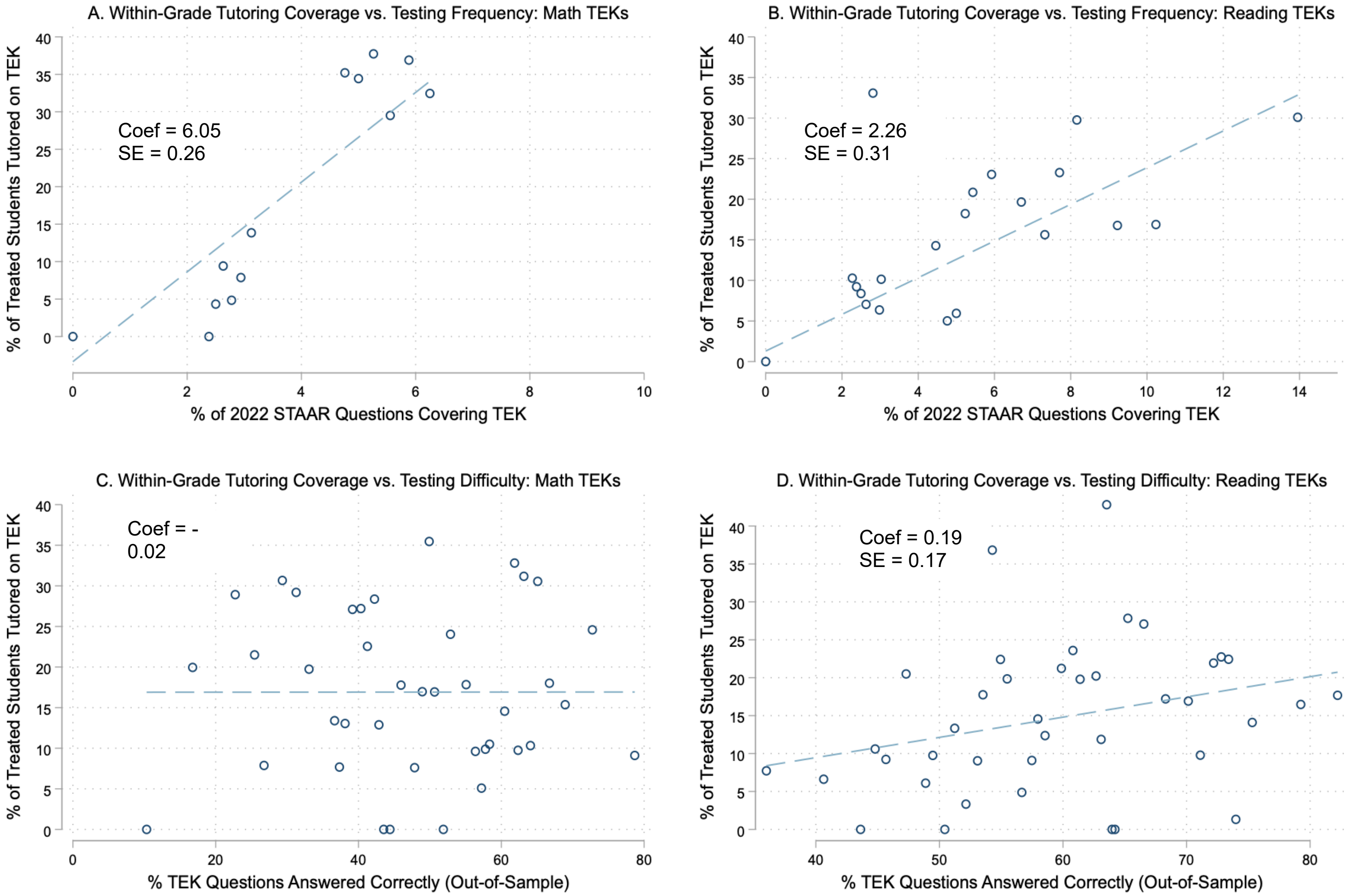
Note: Panels (A) and (B) respectively report first stage and 2SLS results on the intensive margin for each math and reading assesment among the relevant sub-sample. Panels (C) and (D) report the same estimates on the extensive margin of tutoring receipt. Each regression includes controls for baseline student covariates and pre-intervention MAP test scores. Heteroskedasticity robust errors are in parentheses. * p<0.10, ** p<0.05, *** p<0.010

Table 7: 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	16,774	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. * p<0.10, ** p<0.05, *** p<0.010.

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=1680 (i.e., 840 observations per panel). Panels C and D plot the same tutoring coverage data 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.

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

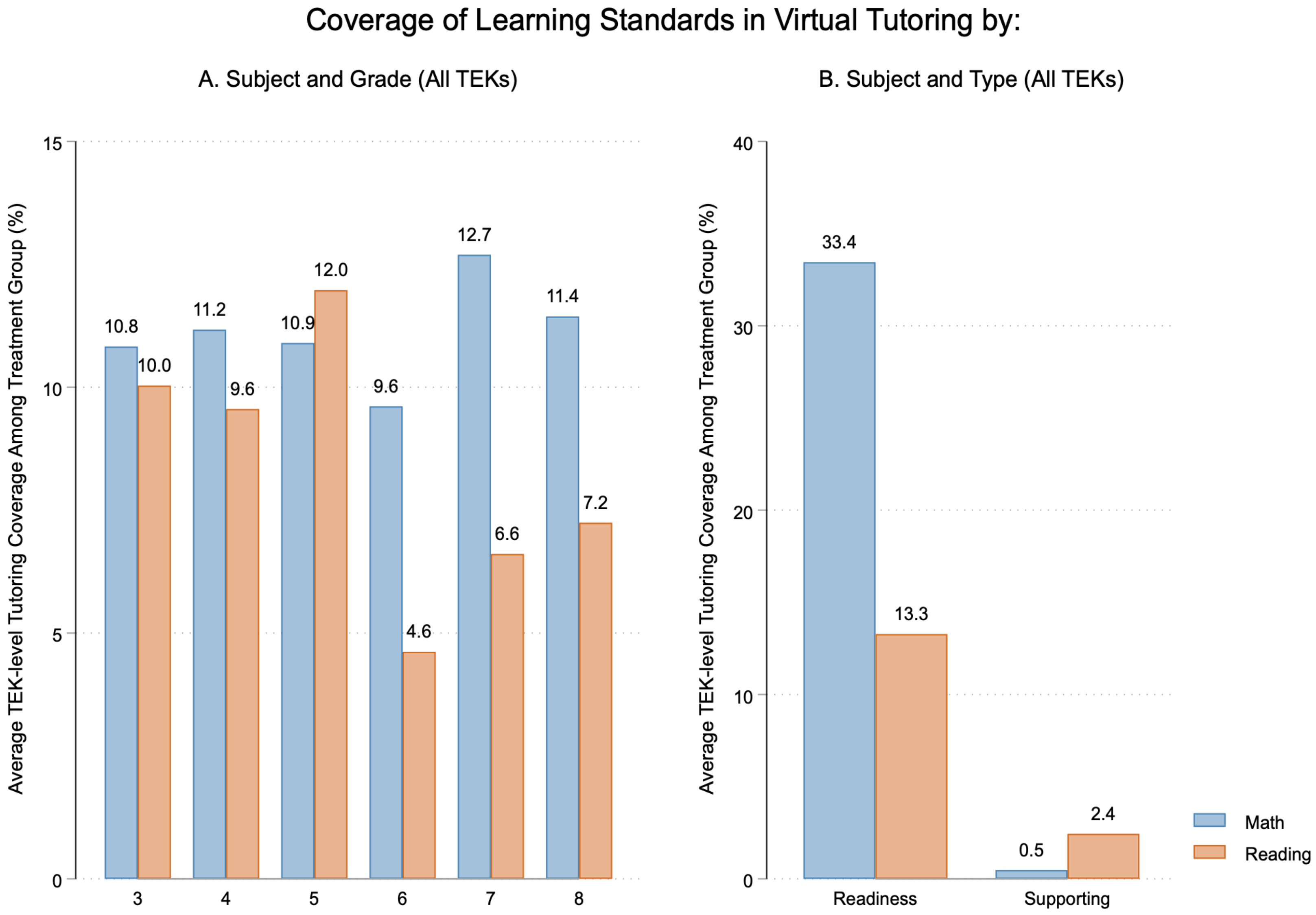


Table 8. Summary of Mediator Analysis (RQ2) & Results

RQ2 Component	Mediator Category	Mediator Description	Analytic Strategy	Key Findings: Math	Key Findings: Reading
A	Features of the counterfactual	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 the treatment	Quantity of tutoring (dosage)	Two-stage least squares	Average dosage low: ~450 minutes, null ATE estimates	Average dosage low: ~500 minutes, negative ATE estimate on (STAAR)
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

DATA APPENDIX

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 A2 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 A3. Null results are replicated and the negative STAAR reading effect is very similar (i.e., ES=-0.081 SD).

Turning to Table A5 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 A6, 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: Auxiliary Regression Discontinuity Estimimates 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
	(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 A2: Intent-to-Treat (ITT) Effects of Virtual Tutoring on Student Outcomes, Alternative Standard Errors

	MAP		STAAR	
	Clustered by Strata	Randomizatio n 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 A3: 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 A4: Intent-to-Treat (ITT) Spillover Effects of Virtual Tutoring on 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 Tutoring						
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
(A) Impacts of Assignment to Math Tutoring on Reading Achievement, Among Students Eligible for Reading Tutoring in Blocks that Received Reading Tutoring						
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 spillover 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. 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 A5. 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 A6. Estimates of the Intent-to-Treat (ITT) Effect of Virtual Tutoring by Assesment Language

	Math		Reading	
	MAP	STAAR	MAP	STAAR
	(1)	(2)	(3)	(4)
(A) English Language Assesments				
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

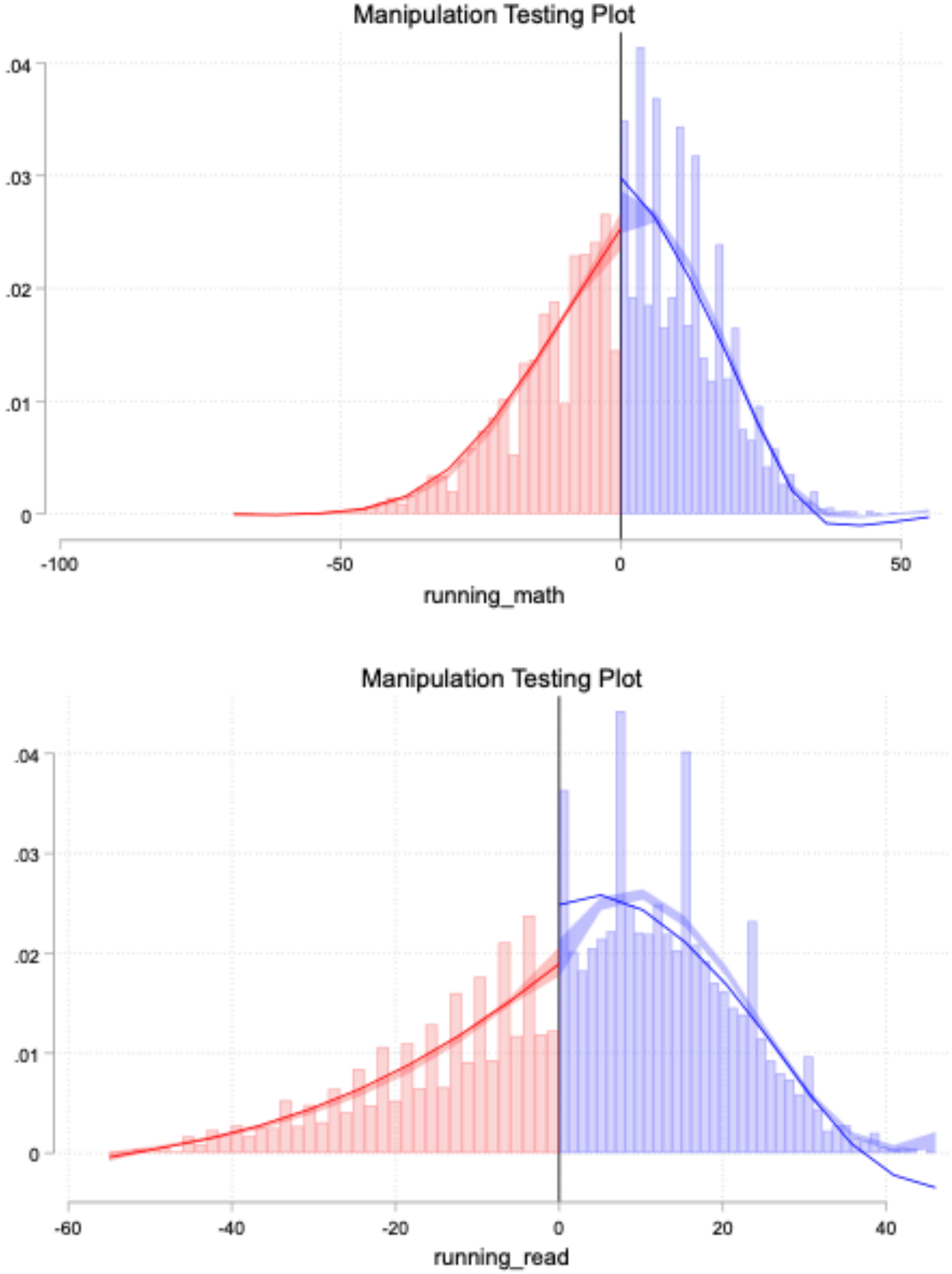
Note: This table presents the intent-to-treat effects of being randomized into either math or reading tutoring on achievement in that subject, by baseline demographic traits. All models include controls for baseline characteristics, pre-intervention test scores, an indicator for assignment to virtual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

Table A7. 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

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


Manipulation Testing Plot
running_math
Manipulation Testing Plot
running_read