WestEd

OBC | The Center for Outcomes Based Contracting

EVALUATION OF OUTCOMES BASED CONTRACTING (OBC) FINAL REPORT

Kelsey Krausen, Sean Tanner, Ruthie Caparas,
Betsy Garcia, and Hope Harrington

July 2026

© 2026 WestEd. All rights reserved.

Suggested citation: Krausen, K., Tanner, S., Caparas, R., Garcia, B., & Harrington, H. (2026). *Evaluation of outcomes based contracting: Final report*. WestEd.

WestEd is a nonpartisan, nonprofit organization that aims to improve the lives of children and adults at all ages of learning and development. We do this by addressing challenges in education and human development, reducing opportunity gaps, and helping build communities where all can thrive. WestEd staff conduct and apply research, provide technical assistance, and support professional learning. We work with early learning educators, classroom teachers, local and state leaders, and policymakers at all levels of government.

For more information, visit WestEd.org. For regular updates on research, free resources, solutions, and job postings from WestEd, subscribe to the E-Bulletin, our semimonthly e-newsletter, at WestEd.org/subscribe.

Table of Contents

Executive Summary

This report presents findings from an independent mixed-methods evaluation of outcomes based contracting (OBC), conducted by WestEd between August 2024 and March 2026 in partnership with the Center for Outcomes Based Contracting (COBC). The evaluation examined how school districts implemented tutoring and educational technology (EdTech) interventions under OBC agreements and explored the effects of OBC on district–provider relationships, implementation quality, student outcomes, and district spending.

The evaluation primarily focused on eight early-adopter districts in California, Florida, Mississippi, and Texas that implemented tutoring or EdTech interventions under outcomes based contracts. Using interviews, observations, document reviews, and quantitative analyses of student achievement and implementation data, the study assessed the conditions under which OBC appears most effective and the organizational changes associated with the model.

Key Findings

Overall, the findings suggest that OBC has strong potential to improve implementation integrity,[1] strengthen district–provider partnerships, and contribute to positive student outcomes when implemented with clear target populations, sufficient dosage and duration, and strong organizational capacity. Specific findings by research question are as follows.

[1] Implementation integrity is similar to implementation fidelity except that implementation fidelity upholds a stricter adherence to the intended implementation model, whereas implementation integrity is more "responsive to varied conditions and contexts" (LeMahieu, 2011). To emphasize that OBC's continuous improvement process includes the flexibility to make adjustments that improve the intervention's implementation, the COBC prefers the term "implementation integrity."

- **Impact on Students' Academic Performance:** OBC interventions targeted students with the greatest need, and in the districts where causal impact analyses were possible and the intervention was delivered for the intended duration, OBC generated positive causal effects on outcomes. OBC also built organizational capacity that may extend benefits beyond the intervention itself.
- **Bidding Process:** The contract development stage provided a key opportunity for building the district–provider partnership's foundation under OBC. OBC required internal district collaboration and external guidance for those new to the process, and some providers appeared better positioned to participate in OBC than others, with particular challenges for nonprofit providers and smaller providers.
- **Relationships Between Districts and Providers:** OBC created stronger accountability for both parties, though it explicitly placed more implementation responsibilities on districts than on providers.
- **District Spending on Interventions:** In the small sample of districts where such an analysis was possible, district spending on interventions under an outcomes based contract was similar to spending on comparable interventions under traditional contracts, suggesting that districts may receive greater intensity of services for a comparable cost. Across all three interventions in the sample, a substantial share of spending was directly tied to students meeting outcomes.
- **Components of OBC That Contribute to Program Effectiveness:** Clear and detailed contracts, enhanced communication, and stronger data use between districts and providers were important for enabling shared problem-solving and improving classroom implementation.
- **Conditions That Support Success:** Successful OBC requires baseline district capacity, adherence to implementation integrity, school-level buy-in, and external support. In addition, engaging in communication with school leaders that empowers them to make necessary site-level implementation adjustments is important for continuous improvement.

Limitations

Several limitations should be considered when interpreting the findings. Districts and school sites often exercised substantial discretion in assigning students to interventions, which made it impossible to conduct causal impact analyses. Specifically, only 4 out of 10 district–provider pairs across the 6 districts assigned students to the intervention in a way that allowed for a rigorous causal analysis. In addition, because the COBC provides intensive coaching and implementation support alongside districts' use of the OBC model, it is difficult to fully isolate the effects of OBC from the effects of the accompanying capacity-building support.

Conclusion

Taken together, the findings indicate that OBC represents a promising approach for strengthening school districts' procurement, implementation integrity, and accountability for student outcomes. The model appears especially promising when paired with strong organizational capacity, implementation support, and sustained attention to continuous improvement. At the same time, the study underscores the importance of continued research, careful implementation planning, and sustained investment in district and school capacity as OBC expands to additional districts, intervention types, and state-level initiatives.

Introduction

This report presents findings from an independent evaluation of outcomes based contracting (OBC) as developed and supported by the Center for Outcomes Based Contracting (COBC). The evaluation was conducted by WestEd from August 2024 through March 2026 and examined the implementation of tutoring and educational technology (EdTech) interventions in eight school districts that used the OBC model to contract with external providers. The goal of this evaluation is to provide rigorous, evidence-based insights that can inform the ongoing work of the COBC and the decisions of education leaders, policymakers, providers, and researchers considering or currently engaged in OBC implementation.

The evaluation was designed to assess the potential impact and conditions for success of the COBC's OBC model. To do so, it draws on both quantitative analyses of student achievement and implementation data and qualitative analyses of data collected through interviews, observations, and document review. The sections that follow provide background on OBC and the COBC, describe the evaluation's research design and methods, present findings organized around six research questions, and suggest implications for policy and practice.

Background on OBC

School districts across the United States routinely contract with external vendors to deliver a wide array of services. Contracted services range from physical operations like transportation, food services, and facility maintenance to instruction-focused services such as EdTech, tutoring, and professional development. After staff compensation, contracted, or purchased, services constitute the next largest proportion of schools' current expenditures—11% of current expenditures, or roughly $100 billion per year (National Center for Education Statistics, 2024). Using outside contractors enables school districts and charter schools to meet student needs that require dedicated resources or specialized knowledge that may not be available in-house. Since schools are funded by taxpayer dollars and play an essential public role—educating the next generation—they are expected to use funds effectively regardless of whether services are provided by district employees or by external vendors. Accordingly, having clear data on vendors' performance and their services' impact on student outcomes serves a dual purpose—it provides public accountability and equips district leaders with the information needed to make informed decisions about which contracted services to sustain, scale, or discontinue.

Traditional procurement approaches in education typically emphasize inputs—which services will be provided, how many staff will be deployed, and/or which materials will be supplied—with payments made regardless of whether desired outcomes are achieved. In contrast, OBC represents a fundamental shift in how districts and vendors (referred to as providers) structure their agreements. Rather than simply purchasing services, districts implementing an intervention under an outcomes based contract tie a portion of provider compensation directly to the achievement of specific, measurable results for a targeted group of students.

Under an outcomes based contract, providers receive a baseline payment for providing services, with additional compensation—representing at least 40% of the overall contract value—contingent upon meeting predetermined success metrics that the district and provider agree upon collaboratively (Center for Outcomes Based Contracting, 2025a). This structure creates mutual accountability through aligned incentives: Districts gain assurance that payments align with results, and providers gain opportunities to demonstrate their effectiveness and earn additional revenue by achieving strong outcomes. OBC is based on the "Pay for Success" model that has been used in numerous projects within public health, social services, and early childhood education (Iovan et al., 2018; Rangan & Chase, 2015). However, its use in K–12 education remains relatively new (Waymack et al., 2024).

Role of the COBC

In 2021, to advance the use of OBC within education, the Center for Education Policy Research at Harvard University launched a multiyear program to support school districts in piloting OBC partnerships with tutoring providers (Barshay, 2024). Shortly thereafter, the

Southern Education Foundation took over leadership of this initiative. In 2024, the Southern Education Foundation established the COBC, which continues to grow in size and in the scope of its support for the implementation of OBC in education.

The COBC supports school districts in adopting, implementing, refining, and scaling an OBC approach when contracting with external partners. The COBC's support includes technical assistance, coaching, research insights, toolkits and guidance materials, and opportunities to learn from peers that are implementing OBC.

Beginning in 2025, the COBC also began implementing a state cohort model, which brings together multiple districts within a state with support from the state education agency and a regional intermediary organization.

Standards of Excellence for OBC

To promote consistency, ensure quality, and foster efficiency in OBC implementation, the COBC developed the Standards of Excellence for OBC™: a framework describing the five key domains of a well-designed OBC (Center for Outcomes Based Contracting, 2025a). These domains are as follows:

1. **Clearly defined population:** identification of a specific group of students who will receive the intervention
2. **Clearly defined outcomes and metrics:** explicit expectations for what will be achieved and how it will be measured
3. **Contingent outcomes payments:** payment structure tied to achievement of outcomes
4. **Mutual accountability:** shared responsibility between districts and providers for success
5. **Continuous improvement:** ongoing monitoring, reflection, and adjustment throughout implementation

Fuller descriptions of each Standard of Excellence are included in Appendix A. Although each of these standards is addressed and documented within the OBC contract, the standards play an essential role throughout implementation, well beyond the procurement stage.

Purpose and Scope of This Evaluation

WestEd's evaluation was conducted in partnership with the COBC and in coordination with two other research organizations: Digital Promise and the National Student Support Accelerator.

The evaluation employed a mixed-methods design combining

- qualitative data collection through interviews, observations, and document review;
- quantitative analysis of student achievement data, implementation metrics, and fiscal information; and
- coordination across research partners to maximize learning and minimize the burden on participating districts.

In addition to this final evaluation report, WestEd provided quarterly update reports to the COBC. These quarterly update reports were intended to document WestEd's research activities and key findings on an ongoing basis and, in combination with WestEd's biweekly meetings with COBC staff, inform the continuous improvement of OBC implementation and the COBC's support.

Study Context

This evaluation focused primarily on districts participating in yearlong COBC cohorts that began between the 2022/23 and 2024/25 school years. One district included in the study did not participate in a COBC cohort but received intermittent technical assistance from COBC staff. Quantitative and/or qualitative data were used from eight[2] early-adopter districts, spanning four states: California, Florida, Mississippi, and Texas. Participating districts ranged in size from approximately 10,000 students to as high as approximately 100,000 students. In all of the districts, more than 70% of the student body consisted of students of color. The population of English Learners ranged from 3% in one district to over 38% in another. In each district, between approximately 700 and 2,500 students received an intervention under an outcomes based contract.

Participating districts implemented two types of interventions under outcomes based contracts:

- **High-Dosage Tutoring:** one-on-one or small-group tutoring focused on literacy or mathematics, delivered virtually or in-person by a live tutor
- **Educational Technology:** technology-based interventions for reading or math, typically implemented during the school day

Most districts piloted one type of intervention, but two of the districts began with one intervention in 2023 and then expanded the use of OBC to another type of intervention in 2024.

[2] For two of these eight districts, the evaluation team was only able to collect qualitative data.

Research Design

Research Questions

WestEd's evaluation of OBC focused on the following research questions:

1. How does OBC impact the bidding process?
2. How does OBC impact the relationship between districts and providers?
3. What impact has OBC participation had on students' academic performance, including spillover effects beyond the students targeted for intervention?
4. How does OBC impact district spending on interventions?
5. Which components of OBC contribute to program effectiveness? (For example, program components include the procurement process, collaboration between the district and contractors, accountability for achieving specific outcomes, data collection and use, and/or implementation of the intervention itself.)
6. Under what conditions is OBC successful? (For example, conditions might include district/school characteristics, leadership, relationships with families, contractor characteristics, and support from the COBC team)

Research Methods

To address the research questions above, WestEd's evaluation centered on three key research activities:

- **descriptive analysis of quantitative data** on implementation, student outcomes, and spending
- **causal impact analysis**, which used a quasi-experimental design to assess OBC's impact on student achievement
- **collection and analysis of qualitative data** on the OBC procurement process, implementation, and other outcomes (including district capacity-building outcomes)

The evaluation employed a mixed-methods approach that integrated qualitative and quantitative data to provide a comprehensive understanding of OBC implementation and impacts. The quantitative analysis aimed to measure the magnitude of OBC's impact on student outcomes and to understand patterns in implementation and spending. The

qualitative work provided rich contextual information about how OBC functions in practice, the experiences of participants, the conditions that support or hinder success, and other factors that may have affected the observed quantitative outcomes.

Quantitative Data Sources

Student Achievement Data: WestEd submitted data requests to participating districts to obtain student-level achievement data, including the following:

- state assessment results for math and reading (including scale score, conditional growth scores, and performance levels)
- state assessment results for English Learners
- interim diagnostic results

Student-Level Demographic and Implementation Data: Districts provided local demographic and implementation data, such as the following:

- grade
- school
- race and ethnicity
- English Learner status
- special education status
- gender
- socioeconomic disadvantage status
- specific demographic data related to the eligibility criteria (for example, McKinney-Vento status or foster status)
- dosage/number of hours participating in the intervention
- intervention attendance
- tutor/student match rate (where applicable)

Contract and Fiscal Data: WestEd compiled and analyzed data from OBC contracts shared by participating districts. These data included the following:

- contract value
- base rate and contingent payment structures
- percentage of contingent value paid based on outcomes achieved
- specified outcomes and metrics
- district and provider responsibilities

Qualitative Data Sources

Interviews: WestEd conducted virtual interviews over Zoom with participating school district leaders, tutoring providers, EdTech providers, and school site leaders involved with OBC implementation. School leaders consisted of principals and "school champions"—school staff members designated as implementation leads for the OBC intervention. The site champion was often a teacher on special assignment or a similar site-level position focused on intervention coordination.

WestEd conducted interviews with district-level leaders from seven districts in the study sample,[3] 6 tutoring providers, 3 EdTech providers, and 12 school leaders. In total, 32 individuals participated in the interviews. Appendix B details the numbers of district leaders, school leaders, and providers interviewed within each district, as well as the number of school observations per district.

Interviews were semi-structured and typically lasted 45–60 minutes. WestEd researchers asked district leaders and providers about their interest in and perceptions of OBC broadly, the process of constructing OBC requests for proposals (RFPs), contract development, the district–provider relationship, implementation of the intervention, implementation support, and any observed outcomes. Interviews with school leaders focused primarily on implementation of the intervention, implementation support, and any observed outcomes, including shifts in school practices.

In-Person Observations: WestEd researchers joined staff from the COBC for multiple in-person observations throughout the evaluation period. These observations yielded qualitative data about many more districts beyond the eight that are the primary focus of this evaluation, as well as additional providers. The events where in-person observations took place included the following.

Partner Visits: WestEd staff joined staff from the COBC for two 2-day Partner Visits, each taking place at a school district in a different region of the United States. During each visit, WestEd staff, COBC staff, district leaders, and school leaders observed student participation in the OBC intervention. WestEd staff also observed COBC staff facilitation of numerous capacity-building activities with district leaders and school leaders. During one of the two Partner Visits, providers also participated in the classroom observations and capacity-building activities, as is COBC's expectation for providers. One of the districts included in the Partner Visits later declined to participate in the research due to concerns about staff capacity at the time, but we do draw from observations from that district's Partner Visit in this report.

3 Interviews were conducted with leaders from seven districts, and six of these districts were included in the study sample for the quantitative analysis.

Annual Site Visit: Each year, the COBC hosts a Site Visit that brings together the new cohort of OBC districts for a joint learning experience to support them in developing their OBC contracts and for planning their implementation of an intervention under an outcomes based contract. Participants in the Site Visit included the new cohort of districts, leaders from the host district, providers offering services in the host district, and researchers. Key activities of the Site Visits included a district-provider panel, shared learning sessions, and a coaching session, in addition to visits to the school sites themselves.

OBC Summer Institute: WestEd attended the OBC Summer Institute, an intensive training for OBC coaches. The training included a detailed overview of what an outcomes based contract is and its various mechanisms. COBC staff also ensured that the perspectives of district leaders implementing OBC were thoroughly embedded throughout the training.

Fall Convening and Provider Summit: WestEd attended the COBC's 2024 Fall Convening and Provider Summit and the 2025 Fall Convening (which combined the Fall Convening and Provider Summit into a single event). This 4-day event for district leaders, providers, researchers, and others interested in learning more about OBC featured presentations by participating districts and COBC staff that aimed to build attendees' understanding of OBC and the importance of strong partnerships between districts and providers.

State Cohort Discussions: The research team attended a 2-day state cohort workshop hosted by the COBC. Four districts, one charter school, and the state education agency were in attendance, as well as the intermediary organization sponsoring the cohort and providing support and relationship building across the many actors in the state cohort.

Document Reviews: WestEd's document reviews included the following materials, when available.

- OBC contracts from participating districts
- non-OBC contracts from participating districts (for comparison)
- OBC rate cards and RFP materials
- district OBC Team Onboarding Questionnaires
- district OBC Team Onboarding Meeting Notes
- district OBC Team Roles and Activities Charts
- the COBC's 2023/2024 District Outcomes Data
- COBC training materials and other publicly available documents describing OBC

Analytical Approaches

Descriptive Analysis of Quantitative Data: For each intervention in each participating district, WestEd analyzed demographic and prior performance profiles of students who were in participating schools, were eligible for interventions conducted under OBCs, and were actually served by those interventions. WestEd compared these to the profiles of students in non-participating schools, ineligible students, and eligible students who did not participate in the interventions. Where possible and appropriate for the particular intervention, WestEd made multiple comparisons for each intervention: participating schools versus non-participating schools, eligible students versus ineligible students, and served students versus students not served.

Impact Analysis: The evaluation team used a mix of research designs in assessing the impact of each intervention on student achievement. Where data and implementation made it possible, the team relied on a regression discontinuity design (RDD) to evaluate the impact of the intervention under an OBC on student achievement. RDD is a research method with strong causal validity that can be used when access or eligibility for a program is clearly defined by a cutoff—in these cases, the eligibility score thresholds that determine treatment eligibility status (e.g., the proficient/not proficient boundary from a prior spring math assessment). Conceptually, students that fall closely on either side of the eligibility threshold are not expected to differ meaningfully, save for access to the intervention. While RDD enables researchers to separate cause from correlation without the use of random assignment, the results are well defined only for those students near the performance threshold that determines tutoring eligibility.

Causal impact analyses using an RDD were possible for at least some subset of the students and interventions in three of six districts that provided quantitative data for analysis. For the other three districts, the evaluation team had to rely on correlational methods, including propensity-score matching and simple regressions that control for demographic characteristics and prior performance. While these methods are often used when natural experiments are not possible, controlling for baseline characteristics that happen to be available in administrative data sets does not credibly eliminate the threat of selection bias. Accordingly, the evaluation team cautions that these correlational analyses should not be interpreted as the causal impact of an intervention under an outcomes based contract.

Contract Analysis: To examine how OBC might impact district resource allocation, the evaluation team also analyzed information related to two districts' spending for three interventions. These districts and interventions were selected because payout data—data on how much they ultimately paid their providers and on which types of outcomes—were available, thereby enabling an analysis of actual spending rather than budgeted costs alone. To do this, the evaluation team reviewed the contract, the rate card, and the payout to

providers for the three interventions of focus (one EdTech and two tutoring interventions). A complementary analysis conducted by researchers at the National Student Support Accelerator (NSSA) at Stanford University examined a broader sample of 123 traditional contracts and outcomes based contracts to compare the extent to which the contracts incorporated elements of the OBC framework, such as clearly defined outcomes, contingent payment structures, and mutual accountability provisions (Lu, Gilbert, et al., 2025). While the NSSA analysis focused on contract content and structure, WestEd's analysis focused on district spending and the relationship between dollars paid and student outcome achievement.

Traditional contracts for similar educational interventions were also reviewed by the WestEd evaluation team; these served as a comparison point to identify the direct effect of OBC on spending. However, such materials were only available for the district that implemented an EdTech intervention, so this information is reported for that district only.

Qualitative Analysis: In addition to this summative final report, the evaluation team began developing qualitative findings throughout the duration of data collection and reported preliminary findings to the COBC team via formative quarterly reports. To identify these preliminary qualitative findings, evaluation team members reviewed observational notes, documents, and interview transcripts for common themes that addressed the study's research questions. The evaluation team used these preliminary findings, as well as discussions of these findings with the COBC team, to inform the final qualitative analysis.

Once all qualitative data had been collected, the evaluation team began its summative analysis by leveraging AI (Claude) to generate initial lists of themes and supporting evidence from all interviews conducted, separated by participant type (providers, districts, and schools) and guided by the research questions. Evaluation team members then manually checked all generated results for accuracy and revised as appropriate. Appendix C offers a more detailed summary of the protocol used for this analysis. The team then reviewed and incorporated other sources of qualitative data (i.e., documents and observational notes) for additional evidence aligned with each research question. The evaluation team also checked for alignment with the identified preliminary themes.

Finally, the evaluation team compiled individual summaries focusing on each district's implementation of a single intervention (so districts implementing two interventions under outcomes based contracts had two summaries each). The intervention summaries included the impact analysis results and qualitative data for all sources for each district and intervention. The evaluation team reviewed these individual summaries for additional evidence that related to any of the summative evaluation's key findings.

Findings

Research Question 1: How does OBC impact the bidding process?

OBC requests for proposals (RFPs) and contract development require elements that necessitate internal district collaboration. External guidance is particularly important during the bidding process for those new to OBC.

In interviews and during in-person observations, district leaders expressed the importance of internal district capacity for collaboratively developing OBC RFPs and contracts. As one example, they described needing to coordinate across district departments, such as legal, purchasing, and programmatic departments, for the creation of the RFP. In addition, district leaders described the need for internal collaboration so that they could be intentional and specific in selecting which student groups would participate in the OBC intervention and which outcome targets they aimed to achieve.

District leaders also underscored the importance of outside guidance on how to craft an OBC RFP and a contract. For example, one district leader described several of the OBC elements not seen in traditional contracts and how support from the COBC was essential in helping the district understand what to include:

> So, making sure that you have your metrics outlined, the target group outlined, how many kids you're servicing, when you're going to have the data ready for [district–provider progress monitoring conversations]—all of that. Those things are critical. Those things came with guidance from the [COBC at the] Southern Education Foundation. … We couldn't have done it without their work because we didn't know what to include in an RFP like this; it's different.

Contract development provides a key opportunity for building the district–provider partnership's foundation.

Interview participants, particularly providers, underscored that the contracting phase is a pivotal time for shaping the district–provider relationship. When both parties collaborated meaningfully during this stage, they were more likely to establish mutual understanding, trust, and alignment on shared goals. Providers noted that working through the terms of the contract and the rate card was not just a procedural task but also an opportunity to build shared ownership of the intervention.

Some providers also described limited opportunities for collaboration at the contracting stage as harmful to the partnership. These providers thought that their relationship with the district—and, ultimately, their ability to deliver on outcomes—was impacted by what they saw as the district's limited openness to negotiate or compromise on the details of the contract. This dynamic was most visible when districts issued RFPs because by that point, key contract terms (e.g., outcome targets, the rate card, the implementation calendar) had typically already been established. What remained for providers was the decision of whether or not to participate—accept the terms or do not bid. As one provider described it, "It wasn't really a contract [negotiation]. It was more of, 'We want to do OBC. Are you in?'" In several cases, providers described situations in which the contracting process felt rushed or inflexible. In one instance, the provider walked away from the contract because of these differences.

When both parties collaborated meaningfully during this stage, they were more likely to establish mutual understanding, trust, and alignment on shared goals.

As discussed further below, some providers also expressed that providers should have an opportunity to take a more active role in outlining their roles and responsibilities for implementation—including "what [providers] bring to the table"—within the contract.

In the districts included in this evaluation, some providers appeared better positioned to participate in and sustain OBC than others, with particular challenges for nonprofit providers and smaller providers.

While many providers were largely positive about OBC—identifying the benefits of getting to showcase the positive effects they were confident they provided and, similar to the district respondents, noting the benefits of closer collaboration—some expressed that OBC participation may be more feasible for some types of providers than others.

Interviews with district leaders offer some insight into the size and composition of the provider pool that OBC RFPs attract. Across districts that issued RFPs, district leaders reported that the number of applications received ranged from 9 to 14 providers. One district leader shared that OBC broadened the types of providers that responded to their RFP compared to the responses they have received for traditional contracts, attracting those they had never partnered with before. However, in at least one case, the bidding process narrowed the pool of interested providers: a district shared that the two strongest nonprofit applicants were ultimately filtered out because the outcome stakes were too high for their organizational risk tolerance.

Data from the current evaluation illustrate this dynamic: Across the 16 district–provider partnerships, only 9 unique providers are represented, with 4 providers serving multiple

districts within the cohorts and 2 providers alone accounting for 6 of the 16 district–provider partnerships. Rather than reflecting a diverse open market, this sample's pool seems to be concentrated among a small core of providers who work across multiple districts and cohorts. This pattern could reflect a higher tolerance for OBC among providers who are renegotiating a contract with a district they have worked in previously, or it could reflect that the bidding process results in filtering effects that constrain the pool of providers.[4] Qualitative evidence suggests that providers who successfully navigated these dynamics and entered into an outcomes based contract may share certain characteristics, such as confidence in their outcomes, tolerance for financial risk, and the organizational capacity to absorb the operational intensity of OBC.

Nonprofit and smaller providers: Nonprofit providers expressed the greatest concerns about OBC's financial risks, citing two related but distinct challenges. Smaller providers may lack the ability to distribute this risk across a broader portfolio of work, and nonprofits may lack the profit margins to absorb losses when outcomes are not fully met. As one nonprofit provider explained, nonprofits' typical business model has minimal margins because they are "not interested in maximizing prices. We're interested most in serving the most kids with the greatest amount of outcomes and decreasing our prices so that we can do that." But as a result of their low baseline pricing, the provider explained, they cannot afford to take on the same risk as for-profit providers that can dip into their margins.

For example, a nonprofit provider described how during their first OBC experience the contract was structured in a way that made it nearly impossible for them to cover their costs. Specifically, the provider had initially discussed entering into a traditional contract with the district, and the district's board of education had approved the provider's usual price, which barely covers their actual costs of providing services. According to the provider, the district's leaders then told the provider that they wanted to shift to an outcomes based contract, and they structured the funding so that the provider would only receive their usual pricing—and thus only be able to cover their costs—if they achieved 100% of the possible contracted outcomes. The COBC recommends that the full potential payout under an OBC be slightly higher than the usual pricing, so this district's structure is not representative of the COBC's model. The provider said that "there was no malicious intent" on the part of the district—suggesting that it may have been simply due to the district's lack of understanding of typical OBC practices—but that this represented a "lesson learned" for the provider.

The provider recommended that for nonprofit providers, the base price—not the maximum total payment—should be set at or near the actual cost of delivering services or else nonprofits will not routinely be able to cover their costs when operating under OBC contracts. At least one provider also noted that to manage this risk, nonprofits may have to raise their rates above their usual pricing, which is challenging for nonprofit organizations

4 Of the seven districts we interviewed, three districts participated in OBC with existing providers.

that—by design—structure pricing to maximize the number of students served (rather than turning a profit).

In addition, some tutoring providers noted that available student assessments are imperfect measures of student progress and, as a result, an outcomes based contract increases their risk. If the assessment data do not accurately reflect the student's growth or performance, then the provider will receive less compensation. If they are a large firm, then this increased risk for one program site may not substantially affect their bottom line, but for smaller programs or programs working only in one area, it can. This suggests that the OBC model may be more appealing to specific types of providers than to others, not only based on their ability to deliver quality programs but also based on their ability to absorb risk.

Other concerns about risk: Beyond nonprofits' structural barriers, some for-profit providers expressed hesitancy about taking on financial risk for implementation challenges that might be outside of their control. While some of the implementation challenges that the district can control—such as making sure students show up to tutoring sessions—can be built into the contracts as the district's responsibility, other challenges may be more difficult to account for in advance.

For example, one provider described a variety of internal school conditions, including scheduling conflicts due to field trips and other school events, temporary unavailability of the computer lab, and other access issues that disrupted implementation but were not necessarily accounted for within the contract. (Note: COBC's Standards of Excellence for OBC design recommend that dosage and access conditions be reflected in the contract.) "There's all these things that at the district level could impact our outcomes and then therefore our payment," the provider said. "So we do feel like we have an outsized share of the risk because we have to set up this program and assume all those fixed costs and setup costs."

Some of the challenges faced by the providers described in this section may stem from contract design and negotiation decisions as much as from structural barriers related to OBC itself. In the former case, thoughtful contract structuring could help broaden provider participation.

Research Question 2: How does OBC impact the relationship between districts and providers?

OBC meaningfully strengthens working relationships between districts and providers.

Both district leaders and providers expressed the view that OBC deepened the district–provider partnership. This finding is echoed in Digital Promise's 2026 research report on OBC (Noakes et al., 2026). Providers expressed that the process of working collaboratively

during contract development and throughout implementation promoted alignment on expectations, roles, and shared goals. On the district side, leaders observed positive shifts in provider contributions to the partnership, such as greater investment and attentiveness, as well as a deep understanding of the district's context. One district leader shared that the experience of OBC led the district to assess new providers based on their potential for strong partnership. This leader shared that the district now incorporates elements of partnership from OBC into the process for traditional contracts.

The interviews also highlighted that an outcomes based contract can be used to strengthen existing partnerships—both partnerships that are going well and partnerships that need some help getting back on track. As one provider pointed out, if a district has an underperforming provider under a traditional contract, the district could propose a change to the provider similar to this:

> We'd love to move forward with you in continuing this work, but we're seeing these things slip here. So, in the next contract we'd like to include some provisions around X, Y, and Z. I think that could be a really smart use of OBC and a good way to refocus a long-standing partnership.

This provider added that after OBC participation, if the district–provider relationship is strong and functioning well, the added structure of an outcomes based contract might not continue to be necessary. The provider indicated that this was the case with a district they had previously partnered with under an outcomes based contract. The year following OBC participation, the district and provider had returned to using a traditional contract, as they felt that they could continue to benefit from the norms, routines, and positive working relationships established during their OBC partnership.

OBC leads to enhanced communication and coordination between districts and providers.

Both district leaders and providers shared that the contract development stage and the first year of implementation require significantly more attention and support under an outcomes based contract than under a traditional contract. They described how implementation of an intervention under an outcomes based contract requires dedicated staff, frequent communication, and intensive coordination. These investments in staffing and communication are needed to coordinate implementation with integrity at the site level and monitor the progress of the intervention.

Participants often shared that, based on their experience, there needs to be at least one person on the district side and one person on the provider side who are responsible for managing the rollout of the intervention to sites. The COBC's model guides districts in identifying an individual from the district (the "project champion") as well as site-level leaders ("site champions") to lead this responsibility. Interview participants also described

district–provider relationships as particularly fruitful when the district team and the provider both had consistent, designated point people to build rapport and trust.

District leaders often shared how this frequent communication led to faster resolution of implementation issues compared to traditional contracts. For example, one district leader stated,

> I think we were successful because we stay connected all the time. All the time. The day one student said they couldn't understand their tutor, I was on the phone with [the provider]. Nothing waited. … We have a lot of district partners and that's not something I hear often enough. I feel like it's a quick turnaround time to resolve issues.

OBC prompts stronger accountability and enhanced service from providers.

During both in-person site visits and interviews, many district leaders and providers described how OBC prompted greater accountability from providers. District leaders, in particular, referred to this heightened accountability as a major asset of the OBC model. Some expressed that OBC has shifted the way they view providers—"as partners" in the work—and heightened their expectations of providers' responsibilities to the district.

These heightened expectations included adjustments to the interventions when requested by the district, and districts reported that providers promptly fulfilled these requests. Examples of adjustments made by various providers included changing an EdTech program's feature to resolve technical issues, changing tutoring content to improve alignment with district standards, implementing a parent communication feature into the tutoring program's online system, and developing weekly data reports for teachers that show individual students' progress.

Several district leaders described the contrast between their partnerships with OBC providers and the more transactional, less collaborative relationships with previous providers. For example, in describing how OBC's sustained level of strong communication transformed the district's relationships with its OBC providers, one district leader shared the following:

> One of the things that we noticed immediately was the partnership that it created between the school district and the vendor as our partner. We work with vendors all the time, but it's more: okay, they sell us something, they train us on it, and then they kind of go away. [But with our OBC partners], both companies assigned us our own special person to keep in contact with, and we met with them weekly.

According to several district leaders, this mindset shift around providers serving as responsive, invested partners has extended beyond the OBC intervention, and they indicated

that moving forward they intend to hold other providers to higher expectations as well. One stated,

> All of our providers, even now, the way we engage with them is a whole other level of expectation. … The way we have engaged with our providers has really changed because of our partnership, our engagement, and [the] outcomes based contract.

OBC prompts stronger accountability from districts to support effective implementation.

District leaders and providers also noted that districts engaged more deeply with program implementation under an outcomes based contract. One district leader described that the district accountability that originates from the contract is "unlike other contracts that we prepare—this one involves skin in the game on the part of the district." Specifically, the contract enumerates implementation responsibilities that the district must fulfill or else the district must pay the provider in full regardless of outcomes.

Providers expressed appreciation for this heightened accountability on the part of the district. One shared, "It's like a two-way accountability street now built into the contract, where there's criteria, parameters that the schools need to live up to … and so I think the accountability piece is huge." Another provider commented, "We've become joint problem solvers … it's not left up to the vendor only to solve the problem. I think our [district] partner is a true partner in trying to find the solution on their end, trying to build capacity with the staff."

However, districts' accountability for strong implementation may differ under the state cohort model, in which districts are not paying for the intervention and so they are not responsible for additional payments if they do not meet implementation integrity measures in the contract. As noted previously, under a typical OBC model, the contract explicitly states the provider's and the district's incentives for fulfilling their implementation commitments. When asked about incentives for implementation integrity under the state cohort model, one provider noted that districts' incentives had not been explicitly expressed. The provider did note that in both models, districts were motivated by the inherent incentive to increase student growth.

"All of our providers, even now, the way we engage with them is a whole other level of expectation. … The way we have engaged with our providers has really changed because of our partnership, our engagement, and [the] outcomes based contract."

Ultimately, OBC does increase accountability on the provider's end relative to traditional contracts, but this increased accountability does not mean that they bear the sole burden of

reaching the goals outlined in the contract. Rather, district leaders and providers tended to describe fulfillment of the "mutual accountability" concept that is central to the OBC model.

Although OBC heightens accountability for both districts and providers, both generally described districts as having a larger number of explicit responsibilities for supporting successful implementation than do providers.

Interviews with providers and district leaders revealed that many of the conditions for successful OBC are largely dependent on districts—as would be the case under a traditional contract. While providers are expected to deliver high-quality services, provide implementation data, and resolve service issues promptly, interviewees described them as still having fewer concrete responsibilities to support implementation.

Essential district responsibilities for successful implementation included the following:

- determining an approach to rolling out the intervention to sites
- having at least one dedicated staff member to manage rollout (especially in the first year of implementation)
- getting buy-in for the intervention from school leadership and teachers
- integrating the intervention seamlessly into the class or school schedule, including during after-school hours
- ensuring attendance/ensuring that students receive the required dosage for the intervention
- providing the technical hardware (e.g., laptops, headphones) and appropriate space
- supervising student participation in the intervention
- identifying technical issues or other implementation issues and raising them to the provider

Interviewees described comparatively fewer responsibilities for the provider. These included primarily the following:

- having at least one dedicated staff member to manage rollout (especially in the first year of implementation)
- providing the service (ideally with consistent staffing, in the case of tutoring)
- attending to technical issues or other service-related issues
- providing progress monitoring data

Some providers acknowledged this imbalance, noting that implementation leaders in districts face substantial pressure, and expressed a desire to better define in the contract how the provider could do more to support implementation. One tutoring provider reflected that "providers need to think about what resources they can bring to the table to make sure implementation runs smoothly, in addition to the district's work." Another provider suggested strategies for assigning greater implementation responsibilities to providers, such as stating in the rate card what kinds of support the provider would continue to provide (e.g., a project manager) under specific contingency planning scenarios (e.g., interruptions to the intervention).

Research Question 3: What impact has OBC participation had on students' academic performance, including spillover effects beyond the students targeted for intervention?

OBC interventions focused on students with the greatest need, the majority of whom went on to meet at least one of their targeted outcomes.

Across the six districts included in the quantitative sample, over 10,000 students (10,805) received an intervention implemented under an outcomes based contract. The number of treated students by district ranged from a low of 726 students in District B to a high of 3,278 in District A. The students reached by these interventions reflected populations that frequently experience the largest opportunity gaps in K–12 education: More than four in five were students of color (81%), more than half were socioeconomically disadvantaged (54%), and nearly half were English Learners (44%). Additionally, 14% of treated students were also receiving special education services. Eligibility criteria for all of the interventions were constructed to focus on students who were below proficiency in the prior year. Together, this descriptive analysis suggests that districts focused OBC interventions on the students most in need of additional support.

As shown in Table 1, the composition of treated students varied meaningfully across districts, reflecting differences in both local demographics and intervention design. The share of English Learners among treated students ranged from 20% in District A to 100% in District C—the latter reflecting an intervention designed specifically for English Learners. Special education enrollment among treated students ranged from 4% in District D to 26% in District A, and the share of socioeconomically disadvantaged students ranged from 64% in District A to 89% in District D. This variation demonstrates how districts tailored their eligibility criteria to the student populations they aimed to serve.

Table 1. Demographics and Performance of Treated Students

	District A	District B	District C	District D	District E	District F
Count of Treated Students	3,278	736	2,230	2,325	1,152	1,084
Percent of Students of Color	81%	*	99%†	93%	86%	97%
Percent Socioeconomically Disadvantaged	64%	*	88%†	89%	*	*
Percent Special Education	26%	*	12%†	4%	6%	18%
Percent English Learners	20%	*	100%†	25%	90%	53%
Percent of Students Meeting at Least 1 Outcome	70%	22%	32%	51%	75%	70%
Percent of Students Meeting all Outcomes	12%	N/A	2%	18%	3%	4%

*Demographic data unavailable for cells marked with an asterisk.

†District C had demographic data available for only one of their interventions. These demographics reflect only that intervention.

Districts' reports of the number of students attaining at least one outcome are indicative of the progress students made in districts using OBC. Overall, 55% of treated students met at least one of their targeted outcomes, and 10% met all of their targeted outcomes. Table 1 shows that attainment of at least one outcome ranged from 22% in District B to 75% in District E, and the share of students meeting every targeted outcome ranged from 2% in District C to 18% in District D. These rates should be interpreted in light of the fact that each district defined its own outcomes, which varied in number and ambition across districts (for example, District B defined a single outcome, whereas District E defined up to six).

This finding is consistent with a 2025 COBC report that analyzed 46 OBC contracts over 2025 and found that the roughly 63,300 students served achieved more than 21,000 outcomes—56% of those contracted (Center for Outcomes Based Contracting, 2025b).

Causal analyses examining academic performance for the subset of district–provider pairs in which quasi-experimental designs were possible are presented in the following section.

Assignment to an intervention under an outcomes based contract was found to have a positive, causal impact on student outcomes in all three of the interventions that allowed for causal analyses and that were delivered for the intended duration.

The evaluation team set out to conduct impact analyses for 10 interventions across six districts. To protect confidentiality, the districts are presented here as Districts A, B, C, D, E, and F. Some districts implemented multiple interventions:

- District A implemented two interventions under an outcomes based contract, referred to here as Interventions 1 and 2.
- District C implemented three interventions under OBC, referred to here as Interventions 1, 2, and 3.
- District E implemented the same type of intervention with two different providers, referred to here as Interventions 1 and 2

Across the 10 interventions analyzed, there were three instances in which it was possible to conduct a causal analysis and the intervention was delivered for its intended duration. In all three of these cases, the evaluation team found that assignment to an intervention under an OBC had a positive impact on student outcomes.

Those three interventions are summarized as follows:

- **District A, Intervention 1:** Eligibility for math tutoring showed strong gains. Although the results suggest a positive effect, the evaluation team cannot attribute gains directly to tutoring since many eligible students did not receive tutoring and some ineligible students did receive it. However, supplemental analysis suggests that at least some of those gains are due to receiving tutoring.
- **District A, Intervention 2:** AI-based English language arts (ELA) tutoring produced large, statistically significant gains in proficiency for eligible 2nd grade students—findings reinforced by a placebo check in non-participating schools.
- **District E, Intervention 1:** Literacy tutoring led to substantial increases in the likelihood that students met typical reading growth targets across both cohorts.

... there were three instances in which it was possible to conduct a causal analysis and the intervention was delivered for its intended duration. In all three of these cases, the evaluation team found that assignment to an intervention under an OBC had a positive impact on student outcomes.

In a fourth intervention, a delay in launching the intervention meant that students only received a fraction of the intended dosage. For the remaining six interventions analyzed for this study, the methods that districts used to assign students to the intervention prevented the opportunity to conduct causal analyses. Collectively, those seven interventions are as follows:

- **District B:** Tutoring produced null findings, which are likely explained by the truncated implementation window because the intervention was intended to run for 25 weeks but was implemented for only about 2 months. Some of these grade-level null results are imprecise, particularly for comparisons among students with higher prior scores.
- **District C (all interventions); District E, Intervention 2; and District F:** Many eligible students did not receive tutoring, weakening the alignment between eligibility and treatment that is needed to use an RDD. As a result, the findings cannot be interpreted as causal.
- **District D:** Students were considered eligible for the intervention if they scored anything less than a perfect score on the prior year's assessment. Because an RDD requires multiple score values on either side of the eligibility threshold, an RDD could not be used in this case.

Tables 2 and 3 include summaries of the impact analysis for the 10 interventions. As noted earlier, dosage and implementation integrity to the eligibility criteria set in the contract were associated with positive impact—for example, longer intervention windows and stronger correspondence between eligibility and assignment to the intervention. The underlying efficacy of the intervention itself is also likely a contributing factor.

It should also be noted that within OBC contracts, districts used independent standardized assessments (not provider-developed assessments) to measure student outcomes. However, the selected assessments varied across OBC contracts. As a result, districts across Table 1 use varying terminology to describe student performance.

Table 2. OBC Results Summary for Regression Discontinuity Design Analysis: Districts A, B, and E (Intervention 1)

	District A Intervention 1	District A Intervention 2	District B	District E Intervention 1
Results Summary	Eligible socioeconomically disadvantaged students in participating schools showed large,[a] significant gains on the state math assessment: a .39 standard deviation increase (p = .021; analytic sample = 418)). Supplemental analysis suggests that at least some of that gain is due to receiving tutoring.	Second graders showed large, significant proficiency gains on the state ELA assessment: a 28 percentage point increase in proficiency (p = .03; analytic sample = 351). A gain of .26 standard deviations on the underlying ELA scale score was also found, but it was less precise (p = .11; analytic sample = 476). No effects were found in 3rd and 4th grades.	Effect sizes on the state math assessment were near 0 across all grades, thresholds, and outcomes. Results are consistent with what researchers typically observe for a short implementation period.	The combined fall and spring cohort effect of receiving tutoring shows large, significant gains in the probability of achieving typical reading growth targets on the NWEA MAP: a 52 percentage point increase (p = .02; analytic sample = 537). Gains in the underlying conditional growth percentile measure were also found: 31 percentile points (p = .03; analytic sample = 454).

	District A Intervention 1	District A Intervention 2	District B	District E Intervention 1
Results Interpretation	The results suggest the intervention under an OBC was effective at improving outcomes, though more information on implementation is needed to determine causal impact.	The intervention under an OBC appears to be more effective with lower grades, since implementation integrity appears similar across grades.	There was no impact of the intervention under an OBC on outcomes, but the implementation window was likely too short to detect effects.	The intervention under an OBC significantly improved outcomes.
Count of Students Receiving Intervention	1,276 students	1,998 students	736 students	Cohort 1: 340 students Cohort 2: 271 students
Year	2023/24	2024/25	2023/24	2024/25
Intervention	Online Math Tutoring	ELA EdTech (AI Tutoring)	Online Math Tutoring	Online Reading Tutoring
Dosage/Duration	Intended: 48 sessions of 45 minutes each for 25 weeks Actual: Mean attendance rate: 61%	Intended: 30 minutes per week for 25 weeks Actual: 74% of students met this dosage requirement	Intended: ~3 times a week for 45 minutes each over about 2 months Actual: Mean number of sessions attended: 20 sessions	Intended: 12 weeks of high-dosage tutoring in fall (Cohort 1) and spring (Cohort 2)[b]
Defined Student Population (Eligibility Criteria)	• 7th–8th graders • 7 specific schools Score below proficiency on prior spring's state math assessment	• 2nd–4th graders • 22 specific schools Score below proficiency on prior spring's state ELA assessment	• 6th–8th graders • All middle schools in the district Score of 3 on prior spring's state math assessment	• 3rd–5th graders • 17 specific schools • English Learners, homeless youth, foster youth

	District A Intervention 1	District A Intervention 2	District B	District E Intervention 1
				• Below 40th percentile on district-selected reading assessment tool (spring for Cohort 1 and fall for Cohort 2)

[a] Interpreting the magnitude of an effect as "large" is context-dependent (Kraft, 2020) and somewhat subjective, with much of the field's guidance being unhelpful for the binary outcomes used in this evaluation. The evaluation team uses the adjective "large" to describe the effects observed here in order to express their practical importance. For example, the percentage point differences found here are larger than the 4th-grade reading proficiency rate differences between West Virginia (where scores are among the lowest in the nation) and Massachusetts (the state with the highest proficiency rate) in the 2022 NAEP (U.S. Department of Education, n.d.). In instances where it was possible to use standard deviations on a scale score as an outcome, the evaluation team also tended to find gains that were "large" with respect to a widely cited benchmark (Kraft, 2020). [b] Student-level dosage data were not provided by the district.

Table 3. OBC Results Summary for Quasi-Experimental Design Analyses: Districts C, D, E (Intervention 2), and F

	District C Intervention 1	District C Intervention 2	District C Intervention 3	District D	District E Intervention 2	District F
Method	Propensity score matching on prior-year test scores On two outcomes (of four) treated students outperformed matched peers	Propensity score matching on prior-year test scores ELA tutoring: lower outcomes than matched peers Math tutoring: no statistically detectable differences	Propensity score matching on prior-year test scores	Propensity score matching Treated students performed worse than matched peers on all end-of-year outcomes. They also scored substantially lower than matched peers on a pre-intervention diagnostic, indicating selection bias.	Few eligible students received tutoring, preventing a causal analysis of the intervention. Results are suggestive of positive impacts if a greater share of eligible students were tutored.	Tutoring assignment did not consistently follow the eligibility criteria set in the contract, preventing a causal analysis. Results from noncausal analyses suggest a mix of positive and negative impacts of the intervention.
Results Interpretation	Outcome differences likely reflect underlying group differences rather than tutoring effects.	Findings cannot be interpreted causally. Lower ELA outcomes likely reflect underlying group differences rather than negative tutoring effects.		Findings are not causal. Lower outcomes likely reflect underlying baseline differences rather than negative intervention effects.	Preliminary findings suggest the intervention had positive impacts on outcomes, but implementation integrity issues prevent a causal analysis.	The intervention showed mixed impacts on outcomes, and implementation integrity issues prevent a causal analysis. The district did not participate in an OBC cohort.

	District C Intervention 1	District C Intervention 2	District C Intervention 3	District D	District E Intervention 2	District F
Count of Students Receiving Intervention	100 students	234 students	1,896 students	2,325 students	Cohort 1: 170 students Cohort 2: 195 students	Reading Y1: 197 students Math Y1: 186 students Reading Y2: 185 students Math Y2: 218 students
Year	2023/24	2023/24	2024/25	2024/25	2024/25	2023/24 and 2024/25
Intervention	Online Early Literacy Tutoring	Online Math and Reading Tutoring	Online English Language Learner Tutoring	Literacy EdTech	Online Reading Tutoring	Online ELA and Math Tutoring
Dosage/Duration	Intended: 20 minutes of tutoring four times a week for 20 weeks Actual: Mean attendance rate: 75%	Intended: Four times a week for 24 weeks Actual: Mean percent of sessions: 83%, reading; 82%, math	Intended: 45 minutes each week for 28 weeks Actual: Average minutes: 1,008, or 36 minutes a week	Intended: 35 minutes each week for 21 weeks*	Intended: 12 weeks of high-dosage tutoring in fall (Cohort 1) and spring (Cohort 2)*	Intended: two times a week for 45 minutes over 24 weeks in 2023/24 and 20 weeks in 2024/25*

	District C Intervention 1	District C Intervention 2	District C Intervention 3	District D	District E Intervention 2	District F
Defined Student Population (Eligibility Criteria)	• 3rd graders • 3 specific schools • Below the state's grade-level benchmark on the prior spring's ELA test	• 3rd–5th graders • 3 specific schools • Below the state's grade-level benchmark on the prior spring's ELA assessment	• 1st–10th graders • Emergent Bilingual status • Beginning or intermediate on the state English language assessment	• 1st graders • Less than a perfect score on a kindergarten high-frequency word assessment	• 3rd–5th graders • 10 specific schools • English Learners, homeless youth, foster youth • Below 40th percentile on reading diagnostic (spring for Cohort 1 and fall for Cohort 2)	3rd–6th graders All elementary schools in the district Below 40th percentile on start-of-year diagnostic of tutoring subject

*Student-level dosage data were not provided by the district.

District A

District A had two OBC contracts: one for virtual math tutoring in the 2022/23 school year and another for English language arts (ELA) EdTech tutoring in 2024/25.

Intervention 1

Intervention 1 was math tutoring for 7th and 8th grade students. The tutoring intervention was delivered virtually by a live tutor on a one-to-one basis. Students in participating schools were eligible for the intervention if they scored below proficient on the prior spring's state math assessment. The goal was to have each child attend 48 sessions of 45 minutes each over 25 weeks, embedded in the school day, beginning in October and lasting through April. The instruction was to be aligned with the district's mathematics curricula for the relevant grade level.

Using an RDD, the research team found meaningful gains on math proficiency on end-of-year test scores for students who were eligible for the tutoring intervention—this is also known as the intent to treat (ITT) effect. In the end-of-year math scores, the ITT effect was a .39 standard deviation increase, which is typically considered large in education research. Proficiency rates went up by 29 percentage points, and student's probability of scoring at a Level 4 or higher increased by 23 percentage points for those who were eligible to receive math tutoring. Because of the discrepancy between which students were eligible for tutoring and which students actually received tutoring, it is difficult to link that ITT effect directly to the tutoring itself; it may instead reflect the additional focus on identified students in schools participating in OBC. However, supplemental but less rigorous analyses showed that students who actually received the tutoring benefited as well.

Although noncompliance with the treatment assignment made it impossible to establish a causal link between the tutoring and the observed math gains, qualitative data suggest that district leaders considered this intervention a success and felt that OBC directly contributed to the observed performance gains.

District leaders reported that the district increased the number of its students who were ready for Algebra 1 when enrolling in high school and decreased the number who would have been referred into the remedial math class, and they attributed this improvement in performance to the intervention. They indicated that this success is what prompted the district to expand its use of OBC to another intervention (Intervention 2) the following year.

District leaders explained that OBC strengthened their involvement with school-level implementation, such as by prompting district leaders to conduct site visits and data conversations with school staff. Furthermore, a school leader reported that both teachers

and students responded positively to the intervention and that they valued the individualized support it provided, which their typical classroom arrangements otherwise made difficult.

Intervention 2

Intervention 2 was an EdTech intervention that provided AI literacy tutoring (EdTech) to 2nd through 4th graders. The EdTech intervention was delivered virtually by an AI tutor on a one-to-one basis. Students at participating schools were eligible for the intervention if they scored below proficient on the prior spring's state assessment in ELA. The goal was to have each student use the platform for 30 minutes per week over 25 weeks, embedded in the school day, beginning in October and lasting through May.

Using an RDD, the research team found meaningful gains in ELA proficiency for 2nd grade students who received the tutoring intervention. Specifically, students who received tutoring were 28 percentage points more likely to score proficient on the end-of-year assessment compared to similar students who did not receive tutoring. The research team also found a .26 standard deviation gain in the underlying ELA scale score for 2nd graders, though the finding is statistically imprecise. Students in grades 3 and 4 do not appear to have benefited from the intervention, though the modest sample sizes for the individual grades mean that small to modest effects of the treatment can't be distinguished from zero.

The positive finding for ELA proficiency was further validated through a placebo check comparing students in schools that did not offer the tutoring intervention. Among 2nd graders who met the eligibility criteria but attended schools without access to the intervention, there was no statistically meaningful difference in the probability of scoring proficient on the end-of-year assessment. This further confirms that the effect observed above is attributable to the intervention itself.

Qualitative data suggest that implementation factors in District A may have shaped these results, including why some students benefited more than others. For example, schools varied in technological readiness; older devices and inconsistent IT infrastructure occasionally disrupted platform access, prompting the provider to train teachers to help with troubleshooting. Site-level facilitation also varied, and the provider reported that when schools had staff dedicated to monitoring and supporting usage, implementation was more consistent.

The stronger effects among younger students may also reflect developmental differences in how students engage with this type of intervention. Younger students may respond more readily to repetitive, scaffolded literacy practice in a structured digital format. Meanwhile, older students may need greater instructional integration or active facilitation to stay engaged, making their outcomes more sensitive to variation in implementation conditions.

District B

District B had one OBC intervention in the 2023/24 school year, which provided math tutoring to middle school students (6th–8th grade). The tutoring intervention was delivered virtually by a live tutor in small groups of three or four students. Students at all middle schools in the district were eligible for tutoring if they scored at Level 3 on the prior spring's state assessment in math. The goal was to have each student attend 48 sessions of 45 minutes each over 25 weeks, embedded in the school day, beginning in October and lasting through April. However, due to delays during the contracting process, the intervention did not begin until the spring semester, so students only received tutoring for about two months. The instruction was to be aligned with the district's mathematics curricula for the relevant grade level.

The intervention showed no detectable effects on student achievement, although the delay in the intervention's launch meant that it had been delivered for only about a third of the initially planned duration. The null effects, while small, are also somewhat imprecise, especially for contrasts involving students at Level 3 versus Level 4 in the prior year.

Additionally, qualitative data gathered from interviews with district and school leaders characterized District B's 2 months of implementation during the 2023/24 school year as a period in which district leaders, school leaders, and the provider were still working through early implementation challenges and operational adjustments—as is common during the first months of implementation of a new intervention. Early challenges included initial resistance among some students and teachers, curriculum alignment issues, some early issues with tutor no-shows, and—most of all—challenges connected to scheduling the tutoring after school rather than during the school day. Having learned from this major challenge during the initial 2023/24 implementation period, the district scheduled the tutoring within the school day in subsequent years and reported that implementation became more successful.

District C

District C had three OBC contracts: one for virtual reading tutoring in the 2023/24 school year (Intervention 1), another for virtual math and ELA tutoring in the 2023/24 school year (Intervention 2), and a final contract for an EdTech intervention for English language development in the 2024/25 school year (Intervention 3).

Intervention 1

The evaluation team had initially planned to use an RDD for the analysis by leveraging the eligibility cutoff on the prior year's spring ELA exam. However, because of both a small sample and because only about 60% of eligible students received tutoring, the team was

unable to use an RDD. As an alternative, the team used a propensity score match, comparing treated students to nontreated peers on prior-year test scores (the only baseline data we were able to obtain). This approach should not be interpreted as causal, as those matching variables still cannot account for why certain students received tutoring and others did not. For two out of four outcomes, there are no statistically detectable differences between the treated students' performance and that of their matched peers. For the other two, the treated students performed better than their matched peers. These differences in the outcomes are almost certainly due to underlying differences in the treatment and matched comparison groups rather than being an effect of the tutoring.

Qualitative data suggest implementation went smoothly overall, though challenges arose around outcome measurement. The provider described the district as a highly collaborative partner that drove strong attendance and school engagement. The provider shared that biweekly meetings with district leadership and weekly or biweekly check-ins with school leaders kept implementation on track and that district leadership intervened when schools fell short of expectations around attendance or student follow-up.

However, a mismatch between assessment tools created challenges. The provider used an assessment tool that targets foundational literacy skills, whereas the district prioritized a different assessment tool that emphasizes comprehension. As a result, contracting required explicit negotiation around goal setting given the lack of alignment and lack of baseline data corresponding to the provider's assessment tool.

Intervention 2

Similar to Intervention 1, the evaluation team had initially planned to use an RDD for the analysis by leveraging the eligibility cutoff on the prior year's spring ELA and math exams. However, because of both a small sample and because over half of eligible students did not receive tutoring, the team was unable to use an RDD. As an alternative, the team used a propensity score match, comparing treated students to nontreated peers on prior-year test scores (the only baseline data the team was able to obtain). This approach should not be interpreted as causal, as noted previously for Intervention A, since those matching variables still cannot account for why certain students received tutoring and others did not. On all four outcomes for the ELA tutoring sample, the students who received tutoring for ELA had lower outcomes than their statistically matched peers. On all four outcomes for the math tutoring sample, the outcome differences between students who received tutoring in math and their matched peers were not statistically distinguishable from zero. Given the inability to sufficiently control for selection bias in which students received tutoring (among those eligible), the differences in the outcomes for the ELA tutoring sample are almost certainly due to underlying differences in the treatment and matched comparison groups rather than negative effects of the tutoring.

In District C, Interventions 1 and 2, a low percentage of students who were eligible for tutoring actually received the treatment, preventing a causal analysis of the intervention's impact on student achievement.

Compliance with eligibility criteria aside, qualitative data suggest that implementation of this intervention reflected both successes and challenges. The provider indicated that a high level of collaboration between the district and provider supported strong attendance and school-level engagement. District leadership and the provider communicated regularly in both formal meetings and informal check-ins, leading to structural improvements in the provider's support throughout the contract. The provider also periodically joined professional learning community (PLC) sessions with grade-level teams to review data and build school-level buy-in. This layered collaboration structure was seen as a meaningful driver of strong attendance and implementation integrity.

However, the provider also noted implementation challenges. Teacher turnover in some classrooms left substitutes unfamiliar with the tutoring program. A midyear curriculum change further disrupted implementation: Tutors who had already built relationships with their student groups were no longer aligned with the new curriculum, requiring tutors to be reassigned and groups to be reorganized. In addition, some schools canceled tutoring sessions in the final month of the program without prior notice. Canceling 3 sessions over 4 weeks amounted to 12 lost sessions, meaningfully reducing the time available for students to achieve intended outcomes.

Intervention 3

During the 2024/25 school year, District C provided an EdTech English language development intervention to emerging bilingual students in the district. Students were eligible for the intervention if they were enrolled in grades 1–10 and scored at the beginning or intermediate level on the prior spring's state English language assessment. The intervention was designed to deliver 45-minute sessions per week over 28 weeks. Although 2,778 students met the eligibility threshold, only 1,896 received the intervention, a compliance rate of approximately 68%.

The research team had initially planned to use an RDD for the analysis by leveraging the eligibility cutoff on the prior spring English language assessment. However, because only approximately one third of barely eligible students received the treatment, and around 15% of barely ineligible students did as well, the jump in treatment participation at the cutoff was not large enough to support a well-powered RDD. As an alternative, the team used a propensity score match, comparing treated students who met the eligibility threshold to eligible students who did not receive the intervention; students were matched on prior ELA and English language assessment scores and demographic indicators. This approach should not be interpreted as causal, as those matching variables cannot fully account for why

certain eligible students received tutoring and others did not. On two of the three measured outcomes—one level of composite growth on the English language assessment and the end-of-year English language assessment reading domain scale score—there were no statistically detectable differences between treated students and their matched nontreated peers. On the third outcome, treated students were 1 percentage point more likely to achieve two levels of composite growth on the English language assessment, a difference that was marginally statistically distinguishable from zero ($p = 0.06$). Given the inability to sufficiently control for selection bias, this difference likely reflects underlying differences between the treatment and comparison groups rather than an effect of the intervention.

Qualitative data suggest that implementation reflected some notable successes alongside several challenges. On the positive side, educators' use of the tool evolved meaningfully over the course of the year: Teachers who initially treated the EdTech platform as a stand-alone computer program gradually learned to integrate it into their instruction and tailor its use to classroom needs. School leaders similarly reported using the platform's data to identify student learning gaps and design targeted minilessons aligned with intervention content.

In terms of challenges, leaders noted that the EdTech intervention was not well matched to the needs of all students served—particularly newcomer students, for whom portions of the program were described as developmentally misaligned or overstimulating. Educators also reported tension between dosage requirements and the individualized, in-person support they felt some students needed, with school champions at times feeling they had to prioritize time on the platform over other instructional supports. Additional challenges included a simultaneous districtwide rollout that complicated buy-in and coordination and a midyear transition in the district's implementation lead that disrupted oversight.

District D

District D had one OBC contract for the 2024/25 school year that delivered an EdTech literacy intervention to 1st grade students across the district. Students were eligible for the intervention if they did not earn a perfect score on the end-of-year Kindergarten High Frequency Words (KHFW) assessment in the spring of 2024. The goal was to have each student use the platform for 35 minutes per week over 21 weeks.

The research team had initially planned to use an RDD for the analysis by leveraging the KHFW eligibility cutoff. However, because eligibility was defined as scoring anything less than a perfect 40 out of 40, there were not enough score values above the threshold to support an RDD. As an alternative, the team used a propensity score match, comparing treated students to nontreated peers on prior-year diagnostic scale scores and accounting for demographic indicators, including race/ethnicity, English Learner status, and special

education status. This approach should not be interpreted as causal, as those matching variables still cannot account for the differences in scores on the KHFW assessment, indicating that substantial selection bias remains even after the propensity score match.

To illustrate the degree of selection bias, the team examined student scores on a beginning-of-year diagnostic assessment taken before the intervention could plausibly have had any effect. Treated students scored substantially lower than their propensity score–matched peers on that baseline measure, indicating that differences in scores are due to unobservable factors in how students were selected to receive the intervention rather than any impact of the intervention. Across end-of-year outcomes—including the diagnostic scale score, whether students scored "on grade level," met typical growth, or met stretch growth—the treated students performed worse than their matched peers. As noted earlier, these differences almost certainly reflect underlying baseline differences between the treated and comparison groups rather than effects of the intervention itself.

Qualitative data collected from both in-person observations and interviews suggest that aside from the eligibility criteria issue mentioned above, additional implementation factors would likely have made it difficult to accurately estimate the impact of OBC. In particular, although the district included all schools within the OBC contract, the degree to which schools actually participated in OBC varied. In addition, within schools that were fully engaged in OBC participation, the treatment of treated students and the comparison group was virtually identical. For example, students from both the treatment and comparison groups received the EdTech intervention under similar conditions, within the same classroom environments, and educational staff reported that they approached data collection and progress monitoring similarly for both groups.

District E

During the 2023/24 school year, two separate providers delivered virtual literacy tutoring for two cohorts of students (Semesters 1 and 2).

Intervention 1

Intervention 1 offered literacy tutoring during both cohorts. The tutoring intervention was delivered virtually. The goal was for each student to receive tutoring after school for 12 weeks in either the fall or the winter/spring. To be eligible for tutoring, students had to be English Learners, experiencing homelessness (as defined by McKinney-Vento status), or living in foster care. They also had to be in 3rd, 4th, or 5th grade and below the 40th percentile on the state standardized ELA test in the spring of the 2022/23 school year (the school year preceding the intervention).

Using an RDD, the evaluation team found large positive effects of the tutoring delivered by Intervention 1 on the probability that students in both cohorts will meet typical growth. In other words, students who received tutoring from Intervention 1 were more likely to demonstrate typical growth in reading achievement than those students who did not participate in tutoring. The gains from tutoring are similarly present in the underlying conditional growth percentile measure. Evidence also suggests improvements in the probability of meeting stretch growth in reading, but those results are not statistically significant and are therefore inconclusive.

Qualitative data offer insight into some of the enabling conditions in the district that may have supported these results. First, leaders reported deliberate efforts to build school-level ownership and support for the intervention. District leaders noted that implementation success depended on having dedicated staff at both the school level and the district level whose primary responsibility was to manage day-to-day implementation and maintain strong communication between the district and the schools. These structures may have helped sustain compliance with eligibility criteria and address implementation issues as they arose.

Second, district leaders described strengthening progress monitoring in the second year of implementation by more directly involving principals. In addition, the provider sent a weekly "consecutive no-show" list so district leaders could follow up with site leaders about students' circumstances and attendance barriers. Together, these practices may have supported stronger implementation integrity.

Intervention 2

Intervention 2 also offered virtual literacy tutoring during both cohorts. The goal was for each student to receive tutoring 3 days per week for 30 minutes per session. The contract stated that cohorts should cover 12 weeks in the fall and the winter/spring. To be eligible for tutoring, students had to be either be English Learners, experiencing homelessness (as defined by McKinney-Vento status), or living in foster care. They also had to be in 3rd, 4th, or 5th grade and below the 40th percentile on the state standardized ELA test in the spring of the 2022/23 school year (the school year preceding the intervention).

The majority of students who were eligible for the intervention based on the criteria set in the contract did not receive tutoring. This suggests that factors beyond the eligibility criteria influenced tutoring assignment, likely including school-level discretion. As a result, the RDD provides only imprecise impact estimates that cannot be distinguished from zero.

Compliance with eligibility criteria aside, qualitative data suggest that implementation of this intervention reflected both strengths and challenges. As a notable strength, leaders engaged in deliberate efforts to build school-level ownership and support for the intervention. District leaders noted that implementation success depended on having dedicated staff at both the school level and the district level whose primary responsibility

was to manage day-to-day implementation and maintain strong communication between the district and the schools. The provider similarly reported assigning a point person for the district to oversee the intervention, conduct regular site visits, and support implementation. District leaders also strengthened progress monitoring in the second year of implementation by more directly involving principals.

However, leaders and the provider also described some implementation challenges. They reported that scheduling the intervention after school reduced attendance, as some parents viewed it as "extracurricular" and lower priority. The virtual format presented additional barriers, with one district leader noting difficulty gaining buy-in even from school administrators due to misconceptions about virtual tutoring. The provider also reported that more students were swapped out mid-program than they normally would have anticipated, which may have compromised outcomes, as the new students participated in the program for fewer weeks.

District F

District F, which did not participate in an OBC cohort but received informal support from COBC staff, had an OBC contract with the same provider over 2 school years. In 2023/24 and 2024/25, the provider offered tutoring in reading and math. Across both years and for both subject areas, the eligibility criteria remained the same. Students in grades 3–6 who scored below the 40th percentile on the diagnostic test administered at the start of the given school year were eligible to participate in the intervention. The intervention was designed to have tutoring delivered twice a week for 45 minutes over a 24-week period in 2023/24 and over a 20-week period in 2024/25 during after-school programming. In some cases, tutoring was provided during the school day depending on teacher approval.

In District F, tutoring assignments did not consistently follow the eligibility criteria in the OBC contract—a fifth of eligible students as well as some ineligible students (based on the contract) received tutoring. This prevents a rigorous causal evaluation of the tutoring intervention in District F. Descriptive analyses show that the tutored students sometimes outperformed and sometimes underperformed their nontutored peers who had similar baseline performance. Those differences are almost certainly due to unobserved baseline differences in the two groups rather than effects of the tutoring.

Qualitative data gathered from interviews with district and school leaders suggest that flexibility in the timing of the intervention may have contributed to this noncompliance with the eligibility criteria. School leaders reported that the intervention was offered after school but that teachers could also offer it during the school day during small-group instruction time. One school leader noted that some teachers were reluctant to use small-group instruction time for the intervention because they were concerned it would disrupt regular

instruction. However, attendance and implementation integrity appeared stronger during the school day, whereas after-school sessions were associated with lower student engagement and more variable attendance.

In addition to the direct impact of the OBC intervention, full participation in OBC builds districts' and schools' organizational capacity, which may improve student outcomes beyond the OBC intervention.

Data collected through interviews and discussions with district, school, and COBC staff at the Partner Visits and Site Visits suggest that OBC may be impacting organizational capacity in addition to having any direct impacts on student outcomes. As described in more detail under Research Question 5, OBC participation improved district and school practices in numerous ways, including

- strengthened capacity for integrity of implementation, including by setting classroom-level implementation routines;
- stronger processes for continuous improvement; and
- enhanced use of data.

In addition, districts often reported that their involvement with OBC led to a "mindset shift" that involved a greater sense of self-efficacy, improved processes for performance monitoring and accountability, and improved coordination between departments. Some districts explicitly shared that OBC had changed their approach to non–outcomes based contracts. Other district leaders commented that they read and understood their district contracts for the first time.

This increase in organizational capacity has the potential to improve outcomes for all students, not just those who are participating in an intervention under OBC, and may improve districts' approach to implementing interventions with integrity in the future—including in-house interventions and interventions under traditional contracts.

Research Question 4: How does OBC impact district spending on interventions?

The first four findings in this section are derived from WestEd's analysis of OBC contracts, rate cards, and payout data for three interventions (one EdTech and two tutoring interventions) for which student-level payout data were available. The fifth finding is based on interview data.

District spending on an EdTech intervention under an outcomes based contract was similar to spending on comparable interventions under traditional contracts, suggesting that districts may receive greater intensity of services for a comparable cost.

The total amount paid under the EdTech outcomes based contract was $63,524.97, including a base payment of $37,772.50 and a contingent payment of $25,752.47, yielding an overall per-pupil cost of $23.38 across all 2,717 students who received the intervention. This per-pupil cost is similar to the median amount of $23.12 spent per pupil by the district across three other EdTech interventions under traditional contracts conducted during 2023/24 and 2024/25. However, qualitative findings suggest that implementation under an outcomes based contract, coupled with the structured, rigorous support provided by the COBC, provides a significantly greater intensity of services than that typically associated with a traditional contract.

Similarly to the EdTech intervention, spending on two high-dosage tutoring interventions under outcomes based contracts in one district was broadly consistent with expected spending for this intervention type. However, the absence of comparable traditional contract data in this district limits conclusions about the direct influence of OBC on district spending.

Actual payout data were available for two high-dosage tutoring outcomes based contracts in one district.[5] One served 222 students with a total payout of $177,690, yielding a per-pupil cost of $800.41. The second intervention served 325 students with a total payout of $258,557, yielding a per-pupil cost of $795.56. Comparable traditional contract data were not available for this district, precluding a direct comparison between OBC and traditional contract spending for this intervention type. However, research on high-dosage tutoring programs suggests that per-pupil costs for this intervention type typically range from $1,000 to over $3,000 per student depending on intensity, group size, and tutor type and whether the tutoring is virtual or in-person (National Student Support Accelerator, n.d.). The per-pupil costs observed in both tutoring outcomes based contracts fall somewhat below this range, situating these contracts at the lower end of costs for comparable tutoring interventions. Across the three interventions, quasi-experimental analyses found evidence of positive student outcomes for two—the EdTech intervention and one tutoring contract (the former)—providing relevant context for interpreting per-pupil costs in relation to value.

Across all three interventions, a substantial share of spending was directly tied to students meeting outcomes.

A key indicator of how OBC creates fiscal accountability is the degree to which spending is linked to demonstrated student results rather than service delivery alone. For the EdTech intervention, 1,484 of 2,717 enrolled students (55%) met at least one outcome and generated outcome-based contingent payments. For one tutoring intervention, 156 of 222 enrolled students (70%) generated contingent payments, and for the other tutoring intervention, 246 of 325 enrolled students (76%) generated contingent payments. Across all

5 Since the 2024/25 cohort year, the COBC's data collection has expanded to include more granular, student-level payout information, which would allow for these types of analysis to be conducted across substantially more interventions in the future.

three interventions, contingent payments represented approximately 35%–37% of total spending, suggesting that a meaningful and consistent share of district spending under outcomes based contracts was tied to student performance rather than guaranteed regardless of results.

The EdTech intervention offers the most granular view of what this contingent spending represents in practice. Of the $25,752.47 in contingent payments, 82% ($21,117.03) was generated by students who met at least one outcome, with the remaining 18% ($4,635.44) generated by students who did not meet outcome thresholds but triggered payment due to unmet dosage or data requirements. While the substantial majority of performance-linked spending was directly attributable to demonstrated student results, this 18% is indicative of the district's financial stake in meeting implementation requirements.

For the tutoring interventions, the relationship between outcome achievement and contingent payment ran in the opposite direction, though the difference was small. In both tutoring outcomes based contracts, slightly fewer students generated contingent payments than met at least one outcome—in one intervention, 163 students met at least one outcome while 156 generated a contingent payment, and in the other intervention, 253 students met at least one outcome while 246 generated a contingent payment—a gap of 7 students in each contract across two different providers. This suggests that meeting an outcome was necessary but not sufficient to trigger payment in these contracts and that the district received some student results it did not pay for. In OBC contracts, contingent payments are often conditioned on providers meeting minimum service or implementation requirements (e.g., tutoring match rates) in addition to student outcome attainment. When these thresholds are not met, districts may withhold payment even if students achieved measured outcomes. (Conversely, providers may also withhold services or require payment under certain conditions, even if outcomes aren't met, if the district does not uphold its contractual responsibilities.) While the gap is too small to draw strong conclusions, it is worth noting because it suggests the contingent payment figures for the tutoring interventions may slightly underestimate the true link between spending and student outcomes. Rate card data were not available for these contracts, so the specific conditions driving this pattern could not be determined.

Clear documentation of payout information can highlight important links between spending and student outcomes, promoting transparency and accountability in district resource allocation.

As these findings illustrate, the ability to identify what share of the total contract cost is attributable to students meeting outcomes is a powerful indicator of return on investment under an outcomes based contract. It also facilitates transparent and accountable use of funds to district interest holders and the broader community. The opportunity to measure and communicate this return on investment is greatest when districts and providers

document payout information in a way that makes these links explicit—specifying not only total amounts paid but also the number of students in each payment category and the outcomes to which payments were tied.

OBC participation can shift districts' mindsets to more critically examine how their resources are used and redirect investments toward higher impact interventions.

Interviews and onsite conversations with district leaders suggest that in some cases, OBC participation may influence districts' resource allocation decisions beyond the OBC intervention. For example, several leaders referenced a mindset shift in which they now expect to pay for outcomes rather than for services, and so even with non-OBC interventions, they expect providers to demonstrate evidence of effectiveness. For example, one district leader described a greater willingness to discontinue programs that don't measurably improve student outcomes: "If the data continue to be flat …, I don't care if the teachers like that [program]. Their kids aren't benefiting."

Another leader described how OBC's focus on integrity of implementation has shifted the district's thinking to first consider whether existing programs are being fully leveraged before investing in additional programs. When considering new purchases, the district now regularly assesses them, asking, "Is this competing with another platform? Do we really need this?" The leader continued,

> We've cut down platforms and providers now, because we were like, "This right here does this. If we use all of what we're supposed to use of this provider or this platform, we don't need this [new] platform right here."

Furthermore, one district leader shared that the decision to participate in OBC led the district to intentionally scale back other contracts so that it would have sufficient staff time to invest in OBC implementation. This included hosting principal focus groups to identify areas of program duplication and other places to cut. These intentional cuts, the district leader explained, were "the magic sauce to contribute the time" necessary for implementing an intervention with the level of integrity required by OBC. While only one district leader volunteered such an example, this instance suggests the possibility that even prior to OBC participation, simply starting an outcomes based contract may encourage some districts to critically examine how their resources are used and to redirect time and effort toward higher impact interventions.

This included hosting principal focus groups to identify areas of program duplication and other places to cut. These intentional cuts, the district leader explained, were "the magic sauce to contribute the time" necessary for implementing an intervention with the level of integrity required by OBC.

Research Question 5: Which components of OBC contribute to program effectiveness?

Part of the underlying theory of OBC is that it creates mutual accountability between the district and provider, which then leads to stronger implementation and, ultimately, improved student outcomes.

As noted earlier under Research Question 2, district leaders and providers described how the contract itself and the full OBC process fostered stronger accountability from both providers and districts—mutual accountability—to ensure effective implementation. Specific ways in which OBC contributed to stronger implementation included the following.

OBC shapes classroom-level practice by setting clear expectations in contracts and in communication between the district, the provider, and the sites.

District and school leaders reported that the outcomes based contract provided clear expectations for how the intervention would be implemented, including the dosage requirements and staff responsibilities around ensuring student attendance and tracking implementation data. To ensure that these expectations were met, district and school leaders communicated guidelines for what classroom-level implementation should specifically look like. These guidelines had a strong impact on classroom-level practice.

For example, during observations of two schools that were fully committed to OBC participation, the evaluation team observed how school leaders had established clear routines for implementing the school's EdTech intervention with integrity. At the classroom level, there was clear evidence of teachers and students embracing these routines, ensuring that students received the required intervention dosage and tracking students' progress. When the intervention period began, students knew to collect their laptops in order to begin using the EdTech intervention; projector screens with timers marked the duration of the intervention; students stayed on task during the intervention period; and teachers used sheets to keep track of students' weekly participation (dosage), which were visible to students as well.

The principal attributed classrooms' adherence to these routines to the clear expectations that school leadership set for how the intervention should be implemented, along with OBC's accountability and monitoring practices. The principal explained, "What gets monitored, gets done."

Another principal within that district, whose classrooms exhibited similar routines, noted that the school had been using this same EdTech tool for years, but prior to OBC, they had not received much training on it and had not been implementing it as intended. The principal indicated that as a result, they had not been seeing positive outcomes from it until that year, when OBC prompted them to begin implementing with a high level of integrity.

During interviews, school leaders—which in some cases included the staff who were directly responsible for implementing the intervention in the classroom—described similar adherence to specific implementation routines and strong, widespread attention to dosage requirements. In fact, schoolwide, classroom-level adherence to dosage requirements was the most frequent feature of OBC that school leaders described during interviews.

The principal attributed classrooms' adherence to these routines to the clear expectations that school leadership set for how the intervention should be implemented, along with OBC's accountability and monitoring practices. The principal explained, "What gets monitored, gets done."

Taken together, these classroom-level protocols and routines produced a level of implementation integrity that may not have been achievable without OBC. Leaders reported that this level of implementation integrity represented a meaningful departure from how they had implemented interventions previously and one they often credited with producing stronger outcomes.

Stronger communication between districts and providers enables joint problem-solving and continuous improvement of implementation.

As described earlier under Research Question 2, district leaders shared that OBC prompted stronger communication between districts and providers, as well as stronger accountability from providers, both of which led to faster resolution of implementation issues as they arose.

In addition, one joint responsibility included in OBC contracts is for district leaders and providers to proactively meet and review the intervention's implementation data and student progress data on a regular basis. Districts and providers described these meetings as productive spaces of joint problem-solving; as one provider described, they were "truly a collaboration around the needs of the schools and the district and the particular project at hand."

For example, a district leader said that when reviewing attendance data with the provider, they "saw attendance was not where it needed to be." As a result, the leader said, "[The provider] assisted us with coming up with incentives. How can we incentivize to get students in the seat to attend? What can we do about teachers to get them to have the buy-in?" Other district leaders described how the provider made adjustments to their service (most often, either technical adjustments to the platform or adjustments to their data reports) as a result of these meetings.

Some district leaders and providers reported that their frequent meetings also prompted them to reflect collaboratively on best practices for implementation. For example, one provider shared, "We need to learn together what the right dosage is, and OBC allows us to do these types of reflections; they are built in." Some district and school leaders likewise described ongoing, nuanced conversations about adjusting the dosage to find the optimal balance, given that an intervention's evidence base is based upon a minimum dosage level, yet dosage levels that are higher than necessary can present trade-offs (e.g., more screen time and less instructional time with the classroom teacher). Such reflective conversations suggest that within these schools, OBC meetings with providers have prompted sustained, intentional focus on continuously improving implementation.

OBC strengthened providers' capacity for using data to support successful implementation.

Several providers expressed that they valued how OBC improved their organization's capacity for data collection and analysis and that it shifted their data practices even with non-OBC contracts. For example, one tutoring provider described how OBC participation helped them realize the importance of sharing timely implementation data with their school partners so that school staff could help resolve implementation issues—such as students missing their tutoring sessions. "We figured out that every morning, these coordinators at the sites would need data that says, here are the kids who didn't show up yesterday," so

that they could follow up with those students. The provider noted: "I think OBC forced us to figure some of those things out that we hadn't been doing before. … [Now] that's standard practice for us, whether it's an OBC or not. Because it just means better implementation."

Along with supporting implementation adjustments, enhanced use of data can improve teachers' and students' motivation.

With data collection and monitoring playing a central role in the OBC model, district leaders, school leaders, and providers consistently reported a high level of data use, including for goal setting and progress monitoring, during implementation of interventions under outcomes based contracts. Some district and school leaders shared that they already had strong data practices in place prior to OBC participation, but many others stated that OBC participation strengthened their use of data.

In addition to the required data tracking and progress monitoring meetings with providers described earlier, some beneficial data practices prompted by OBC included the following.

Individual Student Goal Setting: Although this is not a required practice of OBC, one principal described how OBC's focus on individual student growth and progress monitoring had prompted the school to set individual student growth goals for each student (including students who are not part of the target population). The principal also shared that students are involved in setting their own goals collaboratively with educators and that students are enthusiastic about this process.

Motivation From Seeing Growth Data: Referring to the weekly student-level growth data reports that providers produce for teachers, district leaders reported that "seeing this weekly growth has been motivating for teachers and has played a powerful role in strengthening teacher buy-in" for implementing the intervention with integrity. The evaluation team also observed students tracking their own progress on tracker wall displays.

Interviewed students expressed that seeing their progress (e.g., via the EdTech tool and via classroom progress displays) was motivating to them and helped them recognize their growth. For example, interviewed students at this school consistently reported that they liked using the EdTech intervention, and when asked why, a student explained, "It makes me smarter. I get better at reading." Students seemed proud to share their scores and progress with teachers and interviewers, and school and district leaders shared how OBC's focus on these growth data has helped them build a culture that celebrates student growth.

… district leaders reported that "seeing this weekly growth has been motivating for teachers and has played a powerful role in strengthening teacher buy-in" for implementing the intervention with integrity.

Detailed OBC contracts support clarity of implementation requirements.

The WestEd team observed that the OBC contract template used by some districts included a comprehensive set of key details to support mutual accountability for OBC. Examples of such detailed included intervention format (e.g., 1:1 vs. small group, dosage/frequency); number of school sites; district responsibilities for progress-monitoring; grade levels served; number of students served; assessment instrument and timing; entity responsible for pre- and post-intervention assessment; specific contracted outcomes (e.g., growth in phonics, growth in comprehension, growth in phonemic awareness); and per-pupil payments for each of those outcomes.

Some contracts included core essential details (e.g., specific contracted outcomes, per-pupil payments for outcomes) but lacked other details, such as those pertaining to assessment and progress monitoring. For example, few of the contracts included details about district responsibilities for progress monitoring and the planned instruments and timing for pre- and post-intervention assessment. This is consistent with the findings of the National Student Support Accelerator's 2025 research, which found that 100% of outcomes based contracts in their sample included details on outcomes and contingent payments, but somewhat fewer (78.9%) specified mutual accountability mechanisms and only about half (52.6%) outlined continuous improvement processes (Lu, Gilbert, et al., 2025). This variation in contracts' level of detail indicated some variation in how closely participating districts adhered to the COBC's model during the contracts stage, which may have influenced the consistency of OBC implementation and its potential for strengthening mutual accountability within these district–provider partnerships.

In addition, a few providers described instances in which contracts did not specify how certain student circumstances would affect the outcome payments, leading to payment disputes. Two different providers shared the example of student mobility. One provider explained,

> What if a student started the program but left the program halfway through because they moved schools? … It was not detailed enough for: What happens to that student? Do we still get the outcome? … Do we still get the base payment?

Research Question 6: Under what conditions is OBC successful?

Successful implementation requires a baseline level of district capacity.

While full OBC participation demonstrated observable value in strengthening implementation of interventions—and also often strengthened district capacity—participation required a fairly substantial dedication of district staff time and a baseline level of organizational capacity. District-level organizational capacity includes staff from

procurement, evaluation, and measurement who can collaborate on the RFP contracting process; district staff to collaborate with school site staff on planning, implementing, and measuring the progress of the intervention; and data infrastructure to collect and use student progress data. The COBC has structures and supports in place to support districts in developing these capacities, such as a planning year, involvement in a district cohort before initial implementation of an intervention under an outcomes based contract, and dedicated coaches.

The procurement and contracting process, initial launch, and first year of implementation were described as particularly time intensive. Some aspects of this resource intensity pertained to OBC-specific processes and early implementation (e.g., learning how to develop OBC contracts and metrics for the first time). Others pertained to the inherent resource intensity involved in implementing interventions with high levels of integrity and focus (e.g., ensuring the required dosages, supporting school staff with implementation, collecting and monitoring data, participating in continuous improvement routines).

As noted under Research Question 2, interview participants also described district–provider relationships as particularly fruitful when the district team and the provider both had consistent, designated point people to build rapport and trust. The time required of district staff to effectively support implementation of the intervention requires additional district staff capacity.

Additionally, as described under Research Question 2, at least one district leader reported that they intentionally scaled back other contracts in order to have sufficient staff time to invest in OBC implementation. This included hosting principal focus groups to identify areas of program duplication and other places to cut. According to the district leader, these intentional cuts were critical to ensuring that staff had the necessary time available to implement the intervention with the integrity required by an outcomes based contract.

Adherence to key implementation requirements within the OBC contract, such as ensuring the intended dosage, duration, and assignment to eligible students, was strongly associated with the observation of positive outcomes.

OBC contracts specify a number of implementation requirements that are central to the theory of the model—including which students should receive the intervention, the required dosage, and the intended duration. The quantitative findings from this evaluation suggest that these implementation requirements are not merely procedural specifications but rather are meaningful conditions for successful outcomes: Positive causal impacts were observed exclusively in interventions where these requirements were substantially met.

As noted earlier in the impact analysis findings, positive impacts were observed in all three of the interventions that were implemented for the intended duration and in which causal impact analyses were possible. In the remaining interventions, where the intervention's

duration was truncated or the intervention was not delivered primarily to eligible students, positive impacts were not found—in the latter case, because causal impact analyses were not possible. Nevertheless, the impact analyses' results suggest that alignment with dosage, duration, and assignment to eligible students are likely important conditions for creating strong positive outcomes for students.

School-level capacity and interest are important factors in OBC success.

District leaders noted that school sites' existing capacity for implementation, including the strength of their school leadership and their willingness to participate, played a key role in determining which schools experienced greater success with implementation of an intervention under an OBC contract. Similar to the district level, schools with greater organizational capacity were better equipped to participate in the implementation planning, ongoing collaboration with the district and with providers, data analysis, and continuous improvement processes involved with OBC, according to district leaders.

In terms of school-level commitment to OBC, districts' varying approaches to school recruitment for OBC offered insight into this variation. At least two districts factored schools' eagerness to participate into their selection of early-adopter schools. A leader from one of these districts explained their rationale: "Schools that truly wanted to participate [were] the ones selected, because they have to be all-in with it or else you're not going to get the usage that you need from the product." Leaders in another district noted that they only selected schools that they believed had sufficient capacity for participation.

By contrast, most districts in this study required their selected schools to participate. Within these districts, discussions with district leaders and interviews with school leaders indicated that school leadership's support for full OBC implementation varied considerably. Some were fully invested and were implementing with integrity, and others were not.

Including school leaders in collaborative problem-solving and continuous improvement routines plays an important role in achieving school-level buy-in and successful implementation.

While some school leaders described high levels of staff buy-in and viewed the OBC model as highly valuable, this experience was not universal. An important determining factor appeared to be the extent to which district leaders included school leaders—including site champions or other key staff involved in leading implementation—in collaborative problem-solving conversations and continuous improvement routines.

Some school leaders praised their inclusion in ongoing meetings with district leaders and providers to monitor implementation and problem-solve as needed. "The district and the provider both did an awesome job," one school leader shared, "because … there were regular check-ins where everybody came to the table to be able to discuss, how things are going? What's working well? What do we need to tweak, what do we need to change?" As a result of this ongoing, collaborative relationship, the leader felt that "anytime there was something outside of those regularly scheduled check-ins, we were able to reach out to them" for support.

In another district, school leaders described layers of support from the district, including regular collaborative check-ins as well as site visits from district program specialists who offered implementation support. One school leader reported feeling that there "was tons of support throughout the whole process. … I never felt like I was on an island by myself."

An important determining factor appeared to be the extent to which district leaders included school leaders—including site champions or other key staff involved in leading implementation—in collaborative problem-solving conversations and continuous improvement routines.

However, in two other districts, school leaders described that the district's level of support under the OBC contract was minimal and was not noticeably different from what was offered for other, non-OBC interventions. Moreover, one of these school leaders shared that because there was not much communication about OBC's value-add, school staff were not enthusiastic about the additional requirements (e.g., dosage requirements, data collection). The school leader noted, "It was more like compliance: 'Do it, and we're going to be monitoring you.' I don't think that's received very well."

Communication to school leaders that empowers them to make necessary site-level implementation adjustments to interventions under an outcomes based contract is important for continuous improvement.

As is typical with implementation of a new program or intervention, district and school leaders consistently described the need for implementation adjustments after the initial launch. Most frequently, the adjustments involved shifting the intervention's timing in order to improve student participation and/or accommodate scheduling challenges. Most district and school leaders described sufficient site-level autonomy for school leaders to make such implementation adjustments as needed. However, some school leaders and staff, particularly in one district, reported perceiving that the OBC contract prevented them from

making adjustments—including adjustments that would have made implementation more effective.

In particular, the school champions from these schools perceived that the OBC contract's dosage requirement was a burden that they, as the staff responsible for implementing the intervention, were required to figure out how to implement within the difficult constraints of the existing school schedule and context. Both educators described having a very limited window in which to work with students on any interventions, and once the OBC contract began, they were required to use this window for the EdTech intervention—even if the student seemed to require something different at that moment.

One school champion noted that this time constraint "takes away my ability to teach in different ways … because I need to meet the requirements that were given to me based on the [OBC] contract." Similarly, the school champion from the other school shared, "The amount of time that we truly have with the students in the classroom does not line up with what is being required." They continued, "This [time] is supposed to be for me really homing in on things that we need to work on. But I'm having to make sure that they get on this [device] to get their minutes" instead.

The school champions from this district also noted that the intervention did not always seem appropriate for some of the students assigned to receive it. Specifically, they explained that due to both its content level and its virtual nature, the EdTech intervention was not a good fit for some of their newcomer English Learner students. However, as with the scheduling challenge, these school leaders did not perceive an opportunity to discuss adjustments with district leadership or the provider, as they believed that OBC required all initially assigned students to continue with the intervention for the full duration of the contract, without the flexibility to swap out students.

The evaluation team's conversations with the COBC, district leaders, and providers, as well as a review of contracts, made it clear that OBC does provide an opportunity for implementation adjustments. Indeed, under the COBC's Standards of Excellence, the *continuous improvement* standard calls for providers, district leaders, and school leaders and staff to collaboratively work through challenges, including making adjustments that will improve implementation. However, without full inclusion of school leadership and appropriate staff as part of this process, school staff in charge of implementation may not be aware of the flexibility and problem-solving opportunities that exist within an OBC context.

External support and guidance played a key role in building districts' knowledge of and capacity for OBC participation.

District leaders consistently reported that support from the COBC, including coaching, thought partnership, implementation guidance materials (e.g., sample contracts), and

facilitated partner meetings had a major impact on their ability to implement OBC successfully. Specific types of support described as valuable included the following:

- **training and professional development:** educating district teams on OBC principles and processes
- **RFP and contract development guidance:** reviewing drafts, providing feedback, and helping districts define measurables and outcomes
- **rate card development support:** guiding districts through the complex process of structuring performance-based payment
- **sample contracts and templates:** providing examples from other districts
- **cohort meetings for peer learning:** creating spaces for districts to troubleshoot with each other and share learning
- **site visits to other districts:** arranging opportunities to see OBC implementation in action
- **facilitation of conversations with partners:** facilitating collaborative meetings with providers, including mediating problem-solving conversations as challenges arose
- **thought partnership:** asking thought-provoking questions that helped districts think through decisions rather than providing prescriptive answers
- **regular check-ins:** monitoring progress throughout implementation and offering support proactively
- **ongoing coaching and availability:** assigning coaches who were highly responsive to questions and available for calls

District leaders described this external support as most essential during the early stages of the OBC process, including RFP development, contract development, and establishment of collaborative routines for implementation. District leaders consistently emphasized the value of receiving personalized support; some also noted the value of learning from peers through cohorts and conferences, but one district leader reported feeling that these were less valuable than the hands-on, individualized support that they received.

Based on interviews and site observations, it also appears that school leaders' level of involvement in collaborative activities with the COBC team varied by district and school. District-level leaders consistently described a high level of engagement with the COBC's support, and during site visits, school leaders from schools that appeared to be implementing OBC with a high level of integrity were also engaged with the COBC's support. For example, these site leaders routinely participated in COBC-facilitated partner meetings to review implementation progress. However, just as district and school leaders described variation in schools' level of collaboration with district leaders, the school leaders who were

less engaged in these collaborative activities likewise had less opportunity to benefit from the COBC's support.

Study Limitations

Several limitations affect the interpretation of findings from this evaluation, including the following.

Comparison Group Challenges: In allocating the OBC intervention at the student level, multiple school districts chose eligibility criteria that may have also assigned students to additional support beyond the intervention implemented under an OBC contract. In any districts where this may have been true, comparisons between treated and untreated students may have conflated the OBC effect with the potential effects of the other supports (be they positive or negative).

District Discretion in Selecting Participating Schools and Students: Districts exercised discretion in selecting which schools to invite to participate in OBC, and they allowed for school-level discretion in selecting which students from the eligible pool would receive the intervention. It is worth noting that some degree of local discretion is consistent with the OBC model, which is designed to be tailored to local context. Nonetheless, this local discretion may have introduced some selection bias into the sample of participating schools or the treatment group of students. In some cases, discretion in which students were assigned to the intervention was different enough from what was outlined in the contract terms that it prevented the research team from conducting a causal analysis. This represents an important limitation of this evaluation but not necessarily a deviation from the intended OBC model.

Commingling of Treatment Conditions: In multiple school districts, OBC and non-OBC students received similar treatment (e.g., receiving the same EdTech intervention within the same classroom environment). This makes it difficult to disentangle the impact of OBC from the impact of the intervention, since participating in an OBC may create spillover effects. For example, due to collaborative OBC conversations with the district, the EdTech provider may have made changes to its product that affect all students using the tool in the same classroom, school, or even district (both OBC and non-OBC students). From a research perspective, an implication of spillover effects for this evaluation is that they might make it more difficult to isolate the impact of OBC. That said, practically, spillover effects that improve outcomes for students beyond those targeted by the intervention may represent an additional benefit of OBC participation. This tension between the research limitations of spillover effects and their potential practical benefits will be of greater salience as OBC expands and affects a larger number of providers.

Difficulty Isolating OBC Effects From That of the COBC's Capacity-Building Activities: Capacity-building technical assistance on its own, like that provided by the COBC, may independently impact the effectiveness of an intervention by strengthening a district's systems and leadership; as a result of this capacity-building, improvement could then have occurred regardless of whether an outcomes based contract was in place. The COBC also provides rigorous guidance on the implementation of OBC (such as around the Standards of Excellence), which may have produced stronger results than districts would achieve implementing an outcomes based contract without such support. This study did not include a comparison group of districts that were implementing OBC without support from the COBC. As a result, it may be difficult to disentangle the potential effects of the COBC's support from the OBC model in isolation.

Variation in Participation Timing: Districts, schools, and providers were at different stages of OBC implementation during the evaluation period, with some in the first year of implementation, others several years into implementation, and others no longer participating. This variation affected the types of insights each could provide and, in the case of retrospective districts, the level of detail they could accurately recall about their experiences. In addition, the COBC has worked to refine its support over the years, so the support received by earlier cohorts differed somewhat from the support received by later cohorts.

Generalizability of Sample and Limited Sample Size: The districts, schools, and providers willing to participate in OBC may differ systematically from those who do not participate. This limits the generalizability of findings. Additionally, not all districts in the study sample provided data sets for inclusion in the impact analysis, and the number of school leaders from each district that opted to participate in research interviews, despite extensive encouragement to participate, was fairly small.

Note on Sample Size: The WestEd team anticipated a higher level of responsiveness from district and school leaders to their requests to participate in the evaluation based on initial conversations with leaders during the OBC Convening and Site Visits. During these discussions, district leaders expressed interest in the evaluation and a strong belief in the OBC model. Yet, overall responsiveness to data requests and the requests for interviews was slow and some districts opted out of participation, ultimately affecting the sample sizes of both interview participants and quantitative data. Together, the WestEd and COBC teams made extensive efforts to engage districts in evaluation study activities, including through warm introductions from the COBC team, frequent follow-ups from the WestEd team, and follow-up support from the COBC team. However, after initially agreeing to participate, some district and school leaders dropped out of communication when the evaluation team followed up about next steps.

The difficulty of engaging district and school leaders in the evaluation was particularly unexpected given that these same leaders demonstrated a deep level of investment in OBC participation (e.g., during in-person site visits). However, there is evidence suggesting that, generally, their low level of responsiveness stemmed not from unwillingness to participate but from lack of time. This was indicated both by their initial agreement to participate and, in the case of at least one district, their explicit explanations that their staff lacked the capacity to participate in the evaluation at this time. Part of this may have been due to the challenging context in which districts and schools had been operating throughout 2025 and the beginning of 2026, including due to federal policy shifts that created funding uncertainty and affected their operations.

Implications

Consider how the COBC can build district capacity before launching an OBC contract.

Successful implementation of an intervention under OBC requires dedicated project champions at the district and school levels, cross-departmental collaboration (e.g., between procurement, evaluation, and instructional leadership), and sufficient staff bandwidth to sustain intensive communication between the district, provider, and schools. OBC should not be layered onto already-stretched school and district teams without deliberate resource planning—including, when necessary, scaling back lower priority contracts to create the time and focus that OBC requires.

External capacity-building support, such as that provided by the COBC, should also continue to be available and prioritized for districts that are new to OBC or that may need additional capacity development support. Districts, in collaboration with COBC staff, should conduct an honest assessment of their internal capacity prior to launching OBC to identify areas needing further development during the planning year. The COBC can also support districts that are new to OBC in navigating key trade-offs, such as aligning interventions with existing district priorities and maintaining focus on a limited number of priorities.

The COBC should examine whether the state cohort model, intended to improve the feasibility of scaling OBC, affects districts' level of accountability and the local conditions that enable OBC.

The COBC is expanding the OBC model to state-based cohorts as one approach to addressing concerns about scaling the OBC model, given its time intensity and complexity.

The issue of scalability of the OBC model was raised during the Research Collaborative and in prior research on OBC (Lu, Groom-Thomas, et al., 2025). The time intensity and complexity of OBC may make scalability a challenge for COBC staff capacity and for provider, district, and school capacity. COBC staff noted during the Research Collaborative that state-level cohorts are one approach to improving the scalability of the OBC model. Examples of potential benefits to a state-level strategy include a stronger market signal (indicating to vendors that OBC is of interest to multiple districts within their region), greater ability to customize OBC support to state-specific contexts (e.g., aligning OBC outcomes with the state's accountability system), and greater accountability and sustainability through support from both the state education agency and a local intermediary (typically a nonprofit organization that supports the state cohort). More broadly, the state cohort model may strengthen the enabling conditions for successful OBC implementation, such as by providing districts with closer technical assistance, with peer learning opportunities, and with more-tailored support from intermediaries familiar with local policy environments.

However, the COBC will want to continue to study whether the incentives embedded within the OBC district–provider mutual accountability model are maintained in a state cohort model. For example, when districts enter into a contract directly with a provider, they have a financial incentive to meet the implementation integrity measures included in the contract. Under the state model, funding for the intervention comes from the state, rather than the district, so this financial incentive shifts. It will be important, therefore, to study whether the state cohort model leads to shifts in implementation as a result.

Attracting providers that offer effective interventions is critical to the long-term effectiveness of OBC.

The COBC will want to continue to attend to and address barriers that may make it difficult for some providers—for example, nonprofits or smaller providers—to enter into outcomes based contracts. Part of the long-term success of OBC hinges on its ability to attract strong providers because the effectiveness of the intervention itself is a key part of what makes OBC successful in improving student outcomes. Even the most effective implementation under an outcomes based contract is unlikely to improve student outcomes if the intervention itself is not effective.

Ensure that the continuous improvement component of the OBC model consistently reaches the school level.

The experiences of school-level implementers—site champions, teachers, and principals—are a critical factor in OBC success. Districts should ensure that school leaders are included in collaborative progress-monitoring meetings with providers and that they have a genuine problem-solving partner when implementation challenges arise. The COBC's continuous improvement standard should thus consistently be operationalized at the school level as

well as the district level, with clear guidance for districts on how to structure this school-level engagement and support.

Encourage districts and schools to place greater emphasis on ensuring that those students identified for the intervention receive the intervention as intended.

Given that having a clearly defined target population is central to the OBC model, consider placing greater emphasis on fidelity of student assignment to the intervention. One approach to strengthening this alignment is to encourage districts to define the target population in a way that matches the number of targeted students to the number of seats available for the intervention. Aligning the target population to the available capacity from the outset might reduce the need for downstream discretion over the final participant list. Another approach is to provide more information to schools about which students should be targeted for the intervention so that they can help ensure those are the students that receive the intervention. Although some district- and school-level discretion over the final participant list is to be expected and may support stronger implementation integrity, these considerations can help ensure that OBC interventions are provided to the students for whom they are intended.

Invest in rigorous future research to build the OBC evidence base, and strengthen district involvement in contributing to such research.

Although this evaluation's impact analysis leveraged a quasi-experimental design, there were nonetheless limitations. Notably, the quantitative analysis primarily focuses on districts' early implementation of an intervention under OBC. The COBC has continued to evolve the model and strengthen their supports for districts and providers since this early implementation. The field would greatly benefit from additional years of research to establish clearer causal evidence of OBC's impact on student outcomes and explore OBC's applicability to other intervention types and its scalability through state cohort models.

In addition, although districts or schools may understandably want to exercise a great deal of local discretion in which students receive the interventions, defining eligibility criteria that has strong alignment with the targeted population outlined in the contract and the number of seats available allows for stronger quasi-experimental research to measure the impact of the intervention under OBC. This is a tension between local discretion and evidence building that may well be present in future local education agencies using outcomes based contracts.

In the time period since this evaluation's cohorts launched, the COBC's data collection of district spending on interventions under OBC has evolved to include more granular, student-level information. Previously, spending data were collected at the district or contract level, limiting the ability to directly connect resource allocation decisions to individual student outcomes. With more granular spending information, researchers and practitioners can

examine not only whether an intervention produced gains on average but also for which students, in which contexts, and at what cost. These data will be useful for directly tying spending to outcomes and helping the field understand how district resource allocation changes under OBC. Over time, this data infrastructure may also support cost-effectiveness analyses that compare interventions under outcomes based contracts to traditional contracts to strengthen the evidentiary foundation for scaling the model.

A stronger OBC evidence base benefits the entire field—including districts, providers, and researchers—by clarifying what works, for whom, and under what conditions. Early and ongoing research partnerships with OBC districts and providers might include greater emphasis on the value of their input into research through interviews and data sharing for building the OBC evidence base. As an important step, the COBC has invited districts to participate in the COBC Research Collaborative. The COBC may also want to explore use of state data-sharing agreements to ensure access to data.

Conclusion

This evaluation provides early evidence that outcomes based contracting (OBC) has the potential to reshape how school districts procure, implement, and monitor educational interventions. Across participating districts, OBC strengthened accountability systems, deepened district–provider collaboration, improved implementation integrity, and, in several cases, contributed to measurable gains in student outcomes. Importantly, the findings suggest that OBC's value has the potential to extend beyond any single intervention by helping districts build stronger systems for continuous improvement, data use, and performance monitoring.

These analyses also demonstrate the value of the OBC model for educational research by enabling multiple rigorous causal analyses across varied district contexts and multiple intervention types within a single, coordinated effort. Clearly defining target populations and linking payment to measurable outcomes creates the conditions needed for credible,

scalable, and cumulative evidence generation—while also revealing the implementation factors necessary for both effective interventions and valid research.

The evaluation highlights several conditions that appear critical for successful implementation, including clearly defined target populations, adherence to dosage and duration requirements, meaningful school-level engagement, strong district and school capacity, and sustained collaboration between districts and providers. In the interventions where implementation aligned closely with the contract's intended design and rigorous causal analyses were possible, students experienced positive and statistically significant academic gains. These findings suggest that implementation quality—and the systems that support it—are central to OBC's effectiveness.

The rapid growth of OBC over the past several years also suggests increasing interest in this model among districts and providers nationwide. Between 2022 and 2025, the number of completed OBC contracts grew from 5 to 46, while participation expanded from a small number of early adopters to 87 school systems nationwide (Center for Outcomes Based Contracting, 2025a). This growth reflects increasing momentum around approaches that tie educational investments more directly to measurable student outcomes. The rapid expansion of OBC also has the potential to stretch the capacity of COBC staff, whose support appears to be central to strong initial implementation of OBC by school districts.

The findings further suggest that OBC's greatest long-term impact may lie in its ability to change organizational behavior. District leaders consistently described shifts in mindset toward paying for measurable results rather than services alone, using data more strategically, and holding providers—and themselves—to clearer expectations for implementation and outcomes. In several districts, these changes began influencing practices beyond the OBC interventions themselves.

Although additional rigorous research is still needed, particularly across larger samples and longer implementation periods, this evaluation indicates that OBC offers a promising pathway for improving procurement and strengthening the connection between educational investments and student outcomes. As the application of OBC continues to expand, ongoing attention to implementation quality, provider sustainability, district capacity-building, and evidence generation will be essential to ensuring that OBC fulfills its potential to improve outcomes for students at scale.

References

Barshay, J. (2024, May 27). PROOF POINTS: We have tried paying teachers based on how much students learn. Now schools are expanding that idea to contractors and vendors. *Hechinger Report*. https://hechingerreport.org/proof-points-outcomes-based-contracting/

Center for Outcomes Based Contracting. (2025a). *Standards of Excellence for outcomes based contracting™: High-dosage tutoring*. https://obc.southerneducation.org/standards-of-excellence-high-dosage-tutoring/

Center for Outcomes Based Contracting. (2025b). *2025 impact report*. https://obc.southerneducation.org/resource/2025-impact-report/

Iovan, S., Lantz, P. M., & Shapiro, S. (2018). "Pay for success" projects: Financing interventions that address social determinants of health in 20 countries. *American Journal of Public Health*, *108*(11), 1473–1477. https://doi.org/10.2105/AJPH.2018.304651

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253.

LeMahieu, P. (2011). *What we need in education is more integrity (and less fidelity) of implementation*. Carnegie Foundation for the Advancement of Teaching. https://www.carnegiefoundation.org/blog/what-we-need-in-education-is-more-integrity-and-less-fidelity-of-implementation/

Lu, A., Gilbert, E., Jocelyn, B., Waymack, N., & Loeb, S. (2025). *What's in a contract? How outcomes-based contracting reshapes school district–vendor relationships*. SCALE Initiative Accelerator for Learning, Stanford University. https://nssa.stanford.edu/sites/default/files/How%20Outcomes-Based%20Contracting%20Reshapes%20School-District%20Vendor%20Relationships.pdf

Lu, A., Groom-Thomas, L., & Loeb, S. (2025). *Exploring the potentials of outcomes-based contracting: Findings from initial implementations* (Working paper no. 25-1145.) Annenberg Institute's EdWorkingPapers at Brown University. https://doi.org/10.26300/8r8r-7f03

National Center for Education Statistics. (2024). *Public school expenditures: Condition of education*. Institute of Education Sciences, U.S. Department of Education. https://nces.ed.gov/programs/coe/indicator/cmb

National Student Support Accelerator. (n.d.). *District playbook for high-impact tutoring*. Stanford University. https://nssa.stanford.edu/district-playbook/section-2/developing-budget

Noakes, S., Shell, A., & Vollavanh, A. (2026). *Instructional coherence and shared responsibilities: OBC's impact on EdTech procurement*. Digital Promise. https://digitalpromise.dspacedirect.org/server/api/core/bitstreams/ed67f4a8-8b5e-41e9-a8bd-f4e6f6bcfaa3/content

Rangan, V. K., & Chase, L. A. (2015). The payoff of pay-for-success. *Stanford Social Innovation Review*, *13*(4), 28–38. https://doi.org/10.48558/RDHQ-5W31

U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics. (n.d.). *National Assessment of Educational Progress (NAEP), 2022 Reading Assessment. State achievement-level results*. https://www.nationsreportcard.gov/reading/states/achievement/?grade=4

Waymack, N., Groom-Thomas, L., Loeb, S., & Bendheim, K. (2024). *Outcomes based contracting in tutoring: Insights and recommendations.* National Student Support Accelerator, Stanford University. https://studentsupportaccelerator.org/briefs/outcomes-based-contracting-in-tutoring

Appendix A. Standards of Excellence for OBC

Full descriptions of the Standards of Excellence for OBC are as follows (Center for Outcomes Based Contracting, 2025a):

- **Clearly Defined Population:** The contract defines the population served, specifying the content area, grade level, and like-performing student group (defined by one or more quantitative criteria).
- **Clearly Defined Outcomes and Metrics:** The contract includes clear outcome definitions—determined by the district—with associated measures and achievement goals specific to the target population.
- **Contingent Outcomes Payments:** The contract articulates a financial structure including payments that are earned only as the agreed-upon outcomes are achieved.
- **Mutual Accountability:** The contract includes agreed-upon commitments made between a district and provider, ensuring each party is responsible for the achievement of student outcomes.
- **Continuous improvement:** The contract establishes a systematic approach to data collection, analysis, and collaboration that supports informed, transparent decision-making between districts and providers.

Appendix B. Interview and Observational Data Sources, by District

District	Interviewed District Leaders	Interviewed School Leaders	Interviewed Providers	Observed Schools
District A	2	2	1	6
District B	1	4	N/A	N/A
District C	1	2	2	N/A
District D	1	1	1	2
District E	4	1	3	N/A
District F	1	2	1	N/A
District G*	N/A	N/A	N/A	4
District H*	1	N/A	1	N/A

*Districts G and H declined to participate in the impact analysis.

Note. Table includes only recorded virtual interviews, not data collected from in-person discussions during site visits or in-person OBC convenings.

Appendix C. Protocol for Using AI (Claude) to Support Interview Analysis

In addition to the iterative, formative analysis of interview and observational data conducted by the evaluation team to inform the quarterly reports, the evaluation team employed AI-assisted analysis (via a Claude Enterprise license) to generate an initial summative analysis of interview data for the final report. This approach strengthened the comprehensiveness and analytical rigor of the summative analysis, particularly given the volume and breadth of interview data collected across multiple role types. The AI analysis protocol proceeded as follows:

1. To ground the analysis in the appropriate evaluative context, the evaluation team provided Claude with a brief overview of the OBC evaluation, including the guiding research questions.
2. The evaluation team uploaded transcripts from all district leader, school leader, and provider interviews.
3. The evaluation team directed Claude to perform the following actions:
    - review each set of transcripts in full, organized by role type
    - identify and code recurring concepts based on frequency, synthesizing codes into overarching themes. For each theme, Claude generated a structured table enumerating all associated codes and their supporting interview evidence. All evidence was required to consist of verbatim direct quotes—no paraphrasing—with the corresponding interview source clearly attributed.
    - map identified themes to the research question each addressed
4. To capture key findings that may not have emerged through thematic analysis alone, the evaluation team directed Claude to conduct a parallel analysis focused specifically on strengths and challenges of the OBC model.
5. The evaluation team implemented a multistep quality assurance process to verify the accuracy of results, including

- prompting Claude to self-review its outputs and confirm that all quoted material constituted verbatim direct quotes rather than paraphrased text or composite quotes;
- probing Claude on the context of flagged quotes and removing any that did not substantively support the identified code or theme; and
- directly cross-referencing a representative sample of quotes against the source interview transcripts, including all quotes and examples cited in this report.

The evaluation team then conducted a manual cross-role analysis to identify patterns across role types, consolidating closely related themes into key findings. This process included a review of quarterly report findings to assess alignment and identify any analytical gaps. Key findings were further corroborated through triangulation with additional data sources, including document review and observational data.

Appendix D. Impact Analysis Results

In accordance with research agreements with the districts involved in this study, districts must remain anonymous. To avoid accidentally revealing the identity of the district, we do not name the specific standardized assessments that are used as eligibility criteria or outcome measures. For the same reason, we do not name the intervention providers.

District A

District Overview

District A is a large urban district serving a diverse student population across dozens of elementary, middle, and high schools. During the 2023/24 and 2024/25 school years, District A entered into outcomes based contracting (OBC) agreements with two providers to address academic gaps in mathematics and literacy across multiple grade levels.

In total, approximately 3,300 students received services under OBC contracts in District A during this period. Intervention 1 provided one-to-one virtual math tutoring to 7th and 8th grade students at seven selected schools. Intervention 2 provided an AI tutoring intervention for early literacy to students in 2nd through 4th grades at 20 selected schools.

A meaningful share of treated students made progress toward their targeted outcomes. Approximately 70% of treated students in District A met at least one of their targeted outcomes, and 12% met all of their targeted outcomes.

District A: Intervention 1

Intervention Overview

During the 2023/24 school year, District A provided Intervention 1 to 7th and 8th grade students at seven selected schools. The intervention was designed to deliver tutoring on a one-to-one basis, virtually with a human tutor, for 48 sessions of 45 minutes each over approximately 25 weeks of the school year. This amounted to a planned total of roughly 36 hours of instruction per student. The intervention was scheduled to begin in October and run through April, embedded within the school day. Instruction was aligned with the district's mathematics curricula for the relevant grade level.

The contract between the district and the tutoring provider followed an OBC structure. For each student served by the provider, the district paid a base payment of $680, with additional outcomes-based payments tied to performance on the state standardized math assessment with growth targets based on each student's prior achievement level:

- **$530** if a student achieved their midyear growth target on the state math assessment
- **$700** if a student achieved their end-of-year growth target on the state math assessment
- **$420** if a student scored proficient (Level 3 of 5) on the state math assessment at the end of the year

The maximum per-student payment under this contract was therefore $2,330.

Eligibility Criteria

The criteria for this intervention were a mix of prior performance and attendance, demographic characteristics, and school-level capacity. To receive tutoring, a student had to have

- been enrolled in 7th or 8th grade at the time of the intervention;
- scored below proficiency on the state standardized math assessment in the spring of the 2022/23 school year;
- maintained a daily attendance rate of at least 80% in the prior school year; and
- been enrolled in a particular set of math courses typical for students at their prior level of performance (i.e., not advanced math).

The district selected seven participating schools based on district judgments about need and capacity. These seven schools contained students who had lower average math scores than other middle schools in the district. They also had higher poverty rates and had larger shares of Black and Hispanic students, English Learners, and students with disabilities. Table D1 provides the comparison between the seven selected schools and the other schools that serve 7th and 8th graders in the district.

Table D1. Performance and Demographics of Students in Selected and Non-Selected Schools

	Students in Selected Schools *n* = 3,796 students	Students in Non-Selected Schools *n* = 15,984 students	*P*-Value
End-of-Year State Math Assessment Score for Eligibility (in standardized units)	.14	.48	<0.001
Socioeconomically Disadvantaged	2,288 (60%)	7,995 (50%)	<0.001
English Learners	733 (19%)	1,708 (11%)	<0.001
Students With Disabilities	672 (18%)	2,698 (17%)	0.23
Black/African American	1,673 (44%)	6,485 (41%)	<0.001
Hispanic	866 (23%)	2,829 (18%)	<0.001
Met Eligibility Threshold	1,705 (45%)	5,475 (34%)	<0.001
Received Tutoring	1,276 (34%)	4 (<1%)	<0.001

Note. Only students in the seven selected schools were eligible to receive tutoring, so the *p*-value for Received Tutoring is omitted.

Although 1,705 students in the selected schools met the eligibility threshold, each school site was allocated a particular number of tutoring slots. When the number of allocated slots

was smaller than the number of eligible children, schools could give preference to students who were Black, Hispanic, or from lower socioeconomic families. The final group of students selected for tutoring also had to be allocated such that at least half of the group at each site had scored in the lowest range on last year's math assessment (Level 1 of 5).

Of the 1,705 students who met the eligibility threshold, 936 ultimately received tutoring, a compliance rate of approximately 55%. An additional 340 students received tutoring despite not meeting the eligibility criteria, bringing the total number of treated students to 1,276. These compliance patterns are detailed in the next section and shape the impact analysis that follows.

Descriptive Statistics

This section describes the analytic samples used in the impact analysis. Of the nearly 20,000 7th and 8th grade students in District A during the 2023/24 school year, approximately 3,800 attended one of the seven selected schools. Within those schools, 1,705 students met the eligibility threshold and 1,276 ultimately received tutoring. The remaining tables describe how these groups compared on prior spring assessments and how the treated group's baseline characteristics differed from the comparison group used in the impact analysis.

Within the seven selected schools, students who met the eligibility threshold scored substantially lower than students in the same schools who were not eligible by their prior performance. The eligible students were also more likely to be from lower socioeconomic status families, have a disability, be Black or Hispanic, or be English Learners. Table D2 provides a comparison between the students whose prior state standardized assessment scores made them eligible for tutoring and their peers in the same seven schools.

Table D2 also reveals that tutoring assignments did not perfectly follow the academic eligibility rule. Of the 1,276 students who received tutoring, 936 met the academic threshold and 340 did not. Of the 340 students treated despite not meeting the threshold, about half (190) were missing a prior spring assessment score (thus the threshold could not be evaluated), and the other half (150) scored above grade level on the prior spring math assessment and received tutoring anyway.

Table D2. Performance and Demographics by Tutoring Eligibility

	Ineligible Students *n* = 2,091 students	Eligible Students *n* = 1,705 students	*P*-Value
End-of-Year State Math Assessment Score for Eligibility (in standardized units)	.81	−.36	<0.001
Socioeconomically Disadvantaged	1,202 (57%)	1,086 (64%)	<0.001
English Learner	357(17%)	376 (22%)	0.003
Students With Disabilities	292 (14%)	380 (22%)	<0.001
Black/African American	851 (41%)	822 (48%)	<0.001
Hispanic	449 (21%)	417 (24%)	0.029
Received Tutoring	340 (16%)	936 (55%)	<0.001

Outcomes Analysis

We assess the impact of the tutoring intervention using three post-treatment outcomes, all measured at the end of the 2023/24 school year:

- the end-of-year state math assessment scale score, standardized to mean 0 and standard deviation 1 across the analytic sample
- a binary indicator for whether the student scored at least a Level 3 (proficient) on the end-of-year state math assessment
- a binary indicator for whether the student scored at least a Level 4 on the end-of-year state math assessment

To assess the impact of the math tutoring on student outcomes, we take advantage of the eligibility criterion of lower performance on the prior year's state math assessment, which allows the use of a regression discontinuity design. RDD is a research method with strong causal validity that can be used when access or eligibility for a program is clearly defined by a cutoff—in this case, the prior-year state math assessment score thresholds that determine

tutoring eligibility status by grade. Conceptually, students who fall closely on either side of the eligibility threshold are not expected to differ meaningfully, save for access to the tutoring. While RDD allows researchers to separate cause from correlation without the use of random assignment, the results are well defined only for those students near the performance threshold that determines tutoring eligibility.

At the time of this analysis, we lacked data on several eligibility criteria—attendance and course enrollment. Consequently, we provide here an evaluation of the impact of being eligible for the tutoring based on prior assessment performance and socioeconomically disadvantaged status. The results yield the impact of the intent to treat (ITT) on student achievement, often referred to as the "policy effect" because it captures the effect on student outcomes of the district's eligibility policy, without regard to school site decisions about compliance with that policy.

Table D3 provides the results from the ITT analysis for socioeconomically disadvantaged students in the schools that participated in the OBC intervention. The effect was a .39 standard deviation increase in the end-of-year state math assessment scores, which is typically considered large in education research. Proficiency rates went up by 29 percentage points, and the probability of scoring at a Level 4 or higher increased by 23 percentage points. The latter is notable for not being an outcome on which payment was contingent.

Table D3. ITT Effect in Participating and Non-Participating Schools

	Participating Schools			Non-Participating Schools		
Outcome	**Effect**	***P*-Value**	**95% Confidence Interval**	**Effect**	***P*-Value**	**95% Confidence Interval**
End-of-Year State Math Assessment Score	.39	.021	(.06, .72)	.09	.241	(−.06, .26)
Achievement Level 3 (Proficiency)	.29	.031	(.03, .56)	.02	.755	(−.11, .15)
Achievement Level 4	.23	.057	(−.01, .47)	.002	.963	(−.10, .11)

To support the evidence that these effects were due to the OBC process (including the tutoring), we provide the same effect estimates for students in schools that did not participate in the OBC process (or tutoring). Table D3 contains these findings—no effect can be found on any of the three outcomes. The effect sizes are substantively small and cannot be distinguished from random chance.

While the preceding analyses provide strong evidence that identification of students as eligible for tutoring (the ITT) was only present in schools participating in the OBC process, noncompliance with treatment assignment frustrates our ability to link the effect to the tutoring itself. Many students who were eligible based on their prior-year state math assessment scores did not receive tutoring. Table D4 contains the cross-tabulations of eligibility (based on assessment score) and actual tutoring received.

Table D4. Eligibility Versus Treatment

	Ineligible Students *n* = 2,091 students	Eligible Students *n* = 1,705 students
Received Tutoring ***n* = 1,276**	340	936
Did Not Receive Tutoring ***n* = 2,520**	1,751	769

To supplement the ITT results, we also include comparisons between students who both were eligible and received treatment to their peers in non-participating schools whose prior performance would have made them eligible had their schools participated in the tutoring intervention. We also control for the prior spring math score (the main eligibility criterion). To align this supplemental analysis with the ITT results, we focus on socioeconomically disadvantaged students and those whose prior score was in the Level 2 range (which is near the eligibility threshold). This analysis cannot completely separate the influence of tutoring participation from the influence of everything else that affects a student's outcomes. The potential for selection bias may also be higher in this context because of the unknown choices involved in determining which students actually received tutoring.

With those caveats in mind, the results in Table D5 suggest that the treated students outperformed their peers in non-participating schools. The results for the state math assessment scale score and the probability of achieving a Level 4 are both significant at the 10% level, while the result for the probability of achieving a Level 3 is just outside of that range. Taken together, the ITT analysis and the supplemental analysis point to a meaningful

gain in math achievement for socioeconomically disadvantaged students whose scores from the prior year made them just eligible for tutoring.

Table D5. Comparisons Between Tutored Students and Their Peers in Non-Participating Schools

Outcome	Effect	*P*-Value	95% Confidence Interval
End-of-Year State Math Assessment Score	.19	.08*	(−.03, .40)
Achievement Level 3 (Proficiency)	.10	.12	(−.03, .23)
Achievement Level 4	.08	.10*	(−.02, .18)

*Significant at the .10 level

Implications

Although the noncompliance with treatment assignment made it impossible to establish a causal link between the tutoring and the observed math gains, qualitative data suggest that district leaders considered this intervention a success and felt that OBC directly contributed to the observed performance gains. The evaluation team interviewed two district leaders and two school leaders involved with implementing this intervention.

District leaders reported that the district increased its number of students who were ready for Algebra 1 when enrolling in high school and decreased the number who would have been referred into the remedial math class, and they attributed this improvement in performance to the intervention. The district leaders noted that this success is what prompted the district to expand its use of OBC to another intervention (Intervention 2) the following year.

District leaders reported that OBC strengthened their involvement with school-level implementation by, for example, prompting district leaders to conduct site visits and have data conversations with school staff. They reported that this heightened level of collaboration contributed to "closer relationships between the district and teachers," creating "long-lasting" positive impacts in their ability to work together to support students. District leaders also reported a high level of support from the tutoring provider.

A school leader confirmed that the intervention was well received by students and teachers at their school. The school leader reported that teachers appreciated the opportunity for students to receive individualized support, particularly since these teachers' classroom

arrangements and schedules otherwise did not provide much opportunity for small-group or individualized support. The school leader reported that students, likewise, felt positive toward the tutoring and appreciated the opportunity to ask questions and receive one-on-one support without the pressure of being observed by their peers.

The main challenge described by both school leaders was the initial logistical challenge of scheduling the intervention at a time during the school day that presented the least disruption to normal instructional time. After observing that the initial timing presented issues, both schools adjusted their scheduling to resolve this challenge.

District A: Intervention 2

Intervention Overview

During the 2024/25 school year, District A provided Intervention 2 to 2nd through 4th grade students at 20 selected schools. The intervention was designed to deliver AI tutoring on a one-to-one basis for 30 minutes per week over 25 weeks of the school year. This amounted to a planned total of roughly 12.5 hours of instruction per student. The intervention was scheduled to begin in October and run through May, embedded within the school day.

The contract between the district and the EdTech provider followed an OBC structure. For each student served by the provider, the district paid a base payment of $14.50, with additional outcomes-based payments tied to performance on various assessments:

- **$3.45** if a student achieved their midyear Oral Reading Fluency (ORF) target, defined as reaching the 50th percentile or higher on at least one of two district ORF assessments
- **$6.17** if a student achieved their end-of-year ORF target
- **$13.18** if a student achieved their end-of-year growth target on the state English language arts (ELA) assessment, where growth targets were based on each student's prior achievement level
- **$11.60** if a student scored proficient (Level 3 of 5) on the state ELA assessment at the end of the year

The maximum per-student payment under this contract was therefore $48.90.

Eligibility Criteria

The eligibility criteria for receiving the intervention were based on a student's prior state assessment performance. To receive the AI tutoring, a student had to have

- been enrolled in 2nd, 3rd, or 4th grade at the time of the intervention and
- scored below proficiency on the state standardized ELA assessment in the spring of the 2023/24 school year.

Twenty-two specific schools were named in the contract for participation, but only 20 were ultimately selected (about 1 out of every 5 schools in the district). These 20 schools were identified for participation based on district judgments about need and capacity. These schools served students who had lower average ELA scores than other elementary schools in the district. They also had higher poverty rates and had similar or larger shares of Black and Hispanic students, English Learners, and students with disabilities. Table D6 provides the comparison between the 20 selected schools and the other schools that serve 2nd through 4th graders in the district.

Table D6. Performance and Demographics by School Participation in OBC

	Participating Schools *n* = 3,458 students	Non-Participating Schools *n* = 28,123 students	*P*-Value
End-of-Year ELA Score for Eligibility (in Standardized Units)	−.22	.03	<0.001
Socioeconomically Disadvantaged	1,977 (57%)	14,270 (51%)	<0.001
English Learner	442 (13%)	3,864 (14%)	0.12
Students With Disabilities	788 (23%)	4,646 (17%)	<.0001
Black/African American	1,554 (45%)	10,251 (37%)	<0.001
Hispanic	696 (20%)	5,601 (20%)	0.77

Descriptive Statistics

This section describes the analytic samples used in the impact analysis. Of the approximately 32,000 2nd through 4th grade students in District A during the 2024/25 school year, 3,458 attended 1 of the 20 selected schools. Within those schools, 1,944 students met the eligibility threshold and 1,998 received the EdTech intervention. The remaining tables describe how these groups compared on prior spring assessments and how the treated group's baseline characteristics differed from the comparison group used in the impact analysis.

Within the 20 selected schools, students who met the eligibility threshold scored substantially lower than students in the same schools who were not eligible by their prior performance. The eligible students were also more likely to be from socioeconomically disadvantaged–status families; to be Black, Hispanic, or English Learners; or to have a disability. Table D7 provides a comparison between the students whose assessment test scores made them eligible for the EdTech intervention and their peers in the same 20 schools.

Table D7 also reveals that AI tutoring assignments did not perfectly follow the academic eligibility rule. Of the 1,998 students who received the EdTech intervention, 1,838 met the academic threshold and 160 did not. Of the 160 students treated despite not meeting the threshold, 97 were missing a prior spring assessment score (thus the threshold could not be evaluated), and 63 scored above grade level on the prior spring ELA assessment but received the intervention anyway.

Table D7. Performance and Demographics by Eligibility for the Intervention

	Ineligible Students *n* = 1,514	Eligible Students *n* = 1,944 students	*P*-Value
End-of-Year State ELA Assessment Score for Eligibility (in standardized units)	.68	−.88	<0.001
Socioeconomically Disadvantaged	703 (46%)	1,274 (66%)	<0.001
English Learner	79 (5%)	363 (19%)	<0.001
Students With Disabilities	207 (14%)	581 (30%)	<0.001

	Ineligible Students *n* = 1,514	Eligible Students *n* = 1,944 students	*P*-Value
Black/African American	541 (36%)	1,013 (52%)	<0.001
Hispanic	244 (16%)	452 (23%)	<0.001
Received EdTech Intervention	160 (11%)	1,838 (95%)	<0.001

Outcomes Analysis

We use four binary outcomes for this assessment of the EdTech intervention. The first two are derived from standardized assessments of oral reading fluency—one given in the middle of the school year and the other given near the end of the school year. These are transformed into binary indicators for being in at least the 50th percentile for grade and time of year (e.g., winter of 3rd grade year, spring of 2nd grade year). The third outcome is a binary indicator of proficiency status on the end-of-year state ELA assessment. This binary indicator was derived from two separate ELA assessments—one for students in 2nd grade and another for students in 3rd and 4th grades—each of which have a defined proficiency threshold. The fourth outcome is a binary indicator for reaching at least Level 4 on these end-of-year ELA exams. These outcomes are linked directly to the outcomes based contract for EdTech in the district.

To assess the impact of the OBC intervention on student outcomes, we take advantage of the eligibility criterion of lower performance on the prior year's ELA assessment, which allows the use of an RDD. RDD is a research method with strong causal validity that can be used when access or eligibility for a program is clearly defined by a cutoff—in this case, the prior-year ELA score thresholds that determine intervention eligibility status by grade. Conceptually, students who fall closely on either side of the eligibility threshold are not expected to differ meaningfully, save for access to the EdTech intervention. While RDD enables researchers to separate cause from correlation without the use of random assignment, the results are well defined only for those students near the performance threshold that determines eligibility for the intervention.

We calculate results both for the impact of being eligible for the intervention and the impact of actually receiving it. When focusing on eligibility, the results yield the impact of the ITT on student achievement, often referred to as the "policy effect" because it captures the effect on student outcomes of the district's eligibility policy, without regard to school site decisions about compliance with the policy. When focusing on receiving treatment, the results yield what is often referred to as the effect on "compliers": those students who comply with their

treatment assignment status. For this analysis, that distinction is academic given that most students "complied" with their treatment status—that is, they received the intervention if their scores were in the eligibility range and did not receive treatment if their scores were outside the eligibility range.

Tables D8–D10 provide the results from the ITT analysis for students in the schools that participated in the OBC intervention. As far as proficiency status, students in grades 3 and 4 appear not to have benefited from the intervention, but the students in 2nd grade saw a substantial increase in their likelihood of being proficient: 28 percentage points higher than without the treatment. The complier effect for those students is exactly the same to the second decimal place: a 28-percentage point increase in proficiency. The intervention appears not to have improved winter or spring oral reading fluency exams across any grade level. The effect sizes are all small and not statistically distinguishable from zero.

Table D8. ITT Effect of EdTech Intervention on Proficiency

	Effect	*P*-Value	*N* Below Threshold	*N* Above Threshold
Grade 2	0.28*	0.03	151	200
Grade 3	0.03	0.80	270	302
Grade 4	−0.18	0.13	214	234

*Significant at the .05 level

Table D9. ITT Effect of EdTech Intervention on Spring Oral Reading Fluency

	Effect	*P*-Value	*N* Below Threshold	*N* Above Threshold
Grade 2	0.00	0.99	204	259
Grade 3	−0.02	0.55	229	257
Grade 4	0.03	0.54	278	287

Table D10. ITT Effect of EdTech Intervention on Winter Oral Reading Fluency

	Effect	*P*-Value	*N* Below Threshold	*N* Above Threshold
Grade 2	−0.04	0.39	150	196
Grade 3	0.01	0.83	281	315
Grade 4	0.02	0.70	325	341

Because the threshold for treatment eligibility aligns with achievement level (by design), a concern may arise that the apparent effect of the intervention on 2nd graders is merely the effect of parents' or teachers' differential effort to support students who score at Level 2 rather than Level 3. To provide supporting evidence that the observed effect is due to the intervention, we provide the ITT effect for students in schools that did not participate in the OBC EdTech intervention. Table D11 shows the null effects at all grade levels, including 2nd grade, of this "placebo" intervention. These null effects at non-participating schools, where the parent or teacher effect would presumably still be present, is further evidence that the effect of the intervention on 2nd graders is real.

Table D11. Effect on Proficiency in Participating and Non-Participating Schools

	Participating Schools			Non-Participating Schools		
	Effect	(95% Confidence Interval)	*P*-Value	Effect	(95% Confidence Interval)	*P*-Value
Grade 2	0.28*	(.03, .53)	0.03	.01	(−.08, .10)	.86
Grade 3	0.03	(−.18, .23)	0.8	.02	(−.07, .11)	.65
Grade 4	−0.18	(−.40, .05)	0.13	−.06	(−.15, .03)	.19

*Significant at the .05 level

Finally, to ensure that the positive result for proficiency in 2nd grade is not driven by the binary nature of the outcome, we evaluate the impact of the treatment on the underlying scale score from which proficiency was calculated. The result, in Table D12 below, is a quarter of a standard deviation increase in the scale score. The *p*-value is just barely too

high to meet the 10% confidence level. Despite the imprecision, the results suggest true improvement in the underlying scale score.

Table D12. Effect of EdTech Intervention on ELA Scale Score

	Effect	*P*-Value	*N* Below Threshold	*N* Above Threshold
Grade 2	0.26	0.11	206	270

Implications

While many of the outcomes showed no intervention-related improvement, it is notable that the youngest students made substantial improvement in their ELA proficiency as a result of the intervention. The evaluation team interviewed two district leaders and the provider for this intervention, and qualitative data suggest that implementation factors at District A may have shaped these results.

One challenge was that schools varied in their technological readiness, with older devices and inconsistent IT infrastructure creating occasional disruptions that affected student access to the EdTech platform. As the provider noted, there were "so many permutations of devices," from 2012-era iPads to newer Chromebooks, which complicated troubleshooting technical challenges and could have contributed to uneven implementation. To address these challenges, the provider increased teacher training, which, according to the provider, helped staff "troubleshoot issues independently and better align their classroom setup with the format of the intervention." As reported by a school leader, teachers who engaged more deeply with the platform reported greater success, were able to adapt their instruction to maintain continuity despite technical difficulties, and expressed seeing value in the intervention.

On-the-ground, site-level facilitation also may have contributed to successful implementation in the sites that were able to provide it. The provider noted that districts cannot simply "put the technology in front of the kids and hope that it does some good," emphasizing that "the only way to get good dosages … [is] to have someone who's engaging at the ground level." This suggests that variation in staffing capacity and the presence (or lack) of a dedicated individual to monitor and support usage may have influenced the consistency with which students engaged with the platform.

Furthermore, the stronger effects observed among younger students may reflect developmental differences in how students engage with structured, skills-based interventions, especially in a virtual format. Younger students may be more responsive to

repetitive, scaffolded practice targeting foundational literacy skills, particularly when supported by consistent usage. In contrast, older students may require more integration with broader instructional contexts or more active facilitation to sustain engagement—and therefore may have been less engaged with a digital intervention unless an in-person staff member was also present to monitor usage. As a result, older students' outcomes may have been more sensitive to variation in implementation conditions. Taken together, these findings suggest that both developmental responsiveness and the need for sustained, hands-on implementation support among older students may have contributed to the stronger observed gains among younger students.

District B

District Overview

District B is a large urban district serving a diverse student population. During the 2023/24 school year, District B entered into an OBC agreement with a tutoring provider to address academic gaps in mathematics in middle school grades. Due to delays in the contracting process, the intervention launched in the spring semester rather than October as planned, resulting in students receiving approximately 2 months of tutoring—roughly one third of the intended duration. By the end of the school year, 21% of students who received the tutoring intervention in District B met the contract's proficiency target, scoring a Level 4 or above on the end-of-year standardized math assessment.

Intervention Overview

During the 2023/24 school year, District B provided a mathematics tutoring intervention to 6th, 7th, and 8th grade students. The intervention was designed to deliver math tutoring virtually with a human tutor in small groups of three to four students for 48 sessions of 45 minutes each over approximately 25 weeks of the school year. The intervention was scheduled to begin in October and run through April, embedded within the school day, with instruction aligned with the district's mathematics curriculum for the relevant grade level. As noted above, implementation delays pushed the start date to the spring semester, reducing students' exposure to roughly 2 months of tutoring.

The contract between the district and the tutoring provider followed an OBC structure. For each student served by the provider, the district paid a base payment of $450. The outcomes-based payment was $500 for each student who achieved proficiency (Performance Level 4 or above) on the end-of-year standardized math assessment. The maximum per-student payment under this contract was therefore $950.

Eligibility Criteria

The criteria for this intervention were straightforward: All middle school students who scored a Level 3 on their state standardized math assessment in the spring of the 2022/23 school year (the school year preceding this intervention) were eligible. Some students were excluded from tutoring by the district based on prior-year attendance and/or behavioral concerns.

Of the 972 students who met the eligibility threshold, 736 ultimately received tutoring, a compliance rate of approximately 76%. Students without a previous state standardized math

assessment score or with a score above proficient did not receive the tutoring intervention, making this intervention an appropriate candidate for an RDD, which exploits the score-based eligibility cutoff to estimate causal effects (see the Outcomes Analysis section below).

Descriptive Statistics

Students qualified for OBC tutoring based on a prior-year state math assessment score of 3, placing them in the middle of the performance scale. Students scoring below 3 were excluded because they were already receiving more intensive intervention support targeting their remediation needs. Students scoring above 3 were excluded because they had demonstrated proficiency and did not require additional support.

Table D13 compares prior end-of-year state math assessment performance (math performance used to determine eligibility) across four groups: students excluded for low prior-year scores (Levels 1 and 2), students who participated in tutoring (Level 3), students who were eligible but did not receive tutoring (Level 3), and students excluded for high prior-year scores (Levels 4 and 5). Standardized prior end-of-year scores follow a logical pattern across groups, ranging from −0.93 among excluded low-scorers to 1.35 among excluded high-scorers, with tutored and untutored eligible students scoring 0.04 and 0.02, respectively. Differences across groups are statistically significant ($p < 0.001$). Notably, tutored and untutored eligible students scored nearly identically (0.04 and 0.02, respectively), suggesting that assessment scores were not used to sort students into tutoring among students within the Level 3 range.

Table D13. Performance by Student Eligibility and Tutoring Participation in OBC

	Excluded prior score = 1 or 2	Tutored prior score = 3	Not Tutored prior score = 3	Excluded prior score = 4 or 5
Count of Students	1,051	736	236	733
Prior-Year State Math Assessment Score (in standardized units)	−0.93	0.04	0.02	1.35

Note. Differences in the prior year's end-of-year math state assessment scores across groups are statistically significant ($p < 0.001$).

Outcomes Analysis

We assess two outcomes, both derived from the end-of-year state math assessment: whether a student continued to score at least a Level 3 (passing) and whether a student scored at least a Level 4 (proficient). The Level 4 outcome is the one linked to the OBC contract. To account for the fact that students in the lower end of the Level 3 range (at the 2/3 boundary) may have a difficult time reaching Level 4 the following year, we add the binary Level 3 indicator to evaluate the intervention's impact on maintaining at least a Level 3. Conversations with the district confirm that part of the impetus for providing tutoring to students scoring at Level 3 is to ensure they do not slide back into Level 2 in subsequent years.

To assess the impact of the OBC tutoring on student outcomes, we take advantage of the eligibility criterion of lower performance on the prior year's state standardized math assessment, which allows the use of a regression discontinuity design. RDD is a research method with strong causal validity that can be used when access or eligibility for a program is clearly defined by a cutoff—in this case, the prior year's state math assessment score thresholds that determine tutoring eligibility status by grade. Conceptually, students who fall closely on either side of the eligibility threshold are not expected to differ meaningfully, save for access to the tutoring. While the RDD enables researchers to separate cause from correlation without the use of random assignment, the results are well defined only for those students near the performance threshold that determines tutoring eligibility.

For this intervention, there are two thresholds of eligibility that we use to evaluate the impact of tutoring on student achievement—the threshold that separates Level 2 from Level 3 (the "low" threshold) and the threshold that separates Level 3 from Level 4 (the "high" threshold).

We calculate results both for the impact of being eligible for the intervention and the impact of actually receiving it. When focusing on eligibility, the results yield the impact of the intent to treat on student achievement, often referred to as the "policy effect" because it captures the effect on student outcomes of the district's eligibility policy, without regard to school site decisions about compliance with that policy. When focusing on receiving treatment, the results yield what is often referred to as the effect on "compliers": those students who comply with their treatment assignment status. For this analysis, that distinction is important because approximately 25% of students who were eligible for tutoring did not receive it.

Across all grade levels and both eligibility boundaries, effect sizes are small and none are statistically distinguishable from zero, indicating no detectable impact of the tutoring intervention on either outcome. Tables D14 and D15 contain the ITT results for both outcomes, growth to Level 4 and maintenance of Level 3. Tables D16 and D17 contain the same results but for students who were compliers or those who actually received tutoring.

Table D14. ITT Effect of Tutoring on Growth to Level 4

	Boundary	Effect	*P*-Value	*N* Below Threshold	*N* Above Threshold	Confidence Interval
Grade 6	High	0.01	0.92	247	126	(−.27, .30)
Grade 6	Low	0.02	0.68	194	187	(−.08, .12)
Grade 7	High	−0.05	0.74	205	111	(−.35, .25)
Grade 7	Low	0.01	0.89	326	259	(−.12, .14)
Grade 8	High	0.05	0.68	225	128	(−.18, .28)
Grade 8	Low	0.02	0.72	171	191	(−.10, .14)

Note. "High" boundary compares tutored students to peers who scored at Level 4. "Low" boundary compares tutored students to peers who scored at Level 2.

Table D15. ITT Effect of Tutoring on Maintaining a Level 3

	Boundary	Effect	*P*-Value	*N* Below Threshold	*N* Above Threshold	Confidence Interval
Grade 6	High	0.06	0.56	280	142	(−.26, .14)
Grade 6	Low	−0.04	0.78	135	120	(−.33, .25)
Grade 7	High	0.05	0.34	176	98	(−.05, .16)
Grade 7	Low	−0.05	0.59	290	231	(−.25, .14)
Grade 8	High	0.08	0.61	156	111	(−.23, .39)
Grade 8	Low	0.02	0.92	171	191	(−.28, .32)

Note. "High" boundary compares tutored students to peers who scored at Level 4. "Low" boundary compares tutored students to peers who scored at Level 2.

Table D16. Effect of Tutoring on Compliers' Growth to Level 4

	Boundary	Effect	*P*-Value	*N* Below Threshold	*N* Above Threshold	Confidence Interval
Grade 6	High	−0.04	0.83	280	142	(−.44, .35)
Grade 6	Low	0.03	0.67	194	187	(−.11, .17)
Grade 7	High	0.04	0.80	232	123	(−.30, .39)
Grade 7	Low	0.03	0.78	376	278	(−.15, .20)
Grade 8	High	−0.07	0.63	225	128	(−.37, .22)
Grade 8	Low	0.04	0.61	171	191	(−.10, .18)

Note. "High" boundary compares tutored students to peers who scored at Level 4. "Low" boundary compares tutored students to peers who scored at Level 2.

Table D17. Effect of Tutoring on Compliers' Maintaining Level 3

	Boundary	Effect	*P*-Value	*N* Below Threshold	*N* Above Threshold	Confidence Interval
Grade 6	High	0.07	0.66	311	142	(−.22, .36)
Grade 6	Low	−0.06	0.79	135	120	(−.47, .35)
Grade 7	High	−0.06	0.36	176	98	(−.19, .07)
Grade 7	Low	−0.08	0.56	326	259	(−.35, .19)
Grade 8	High	−0.08	0.67	179	111	(−.45, .29)
Grade 8	Low	0.04	0.85	130	146	(−.34, .41)

Note. "High" boundary compares tutored students to peers who scored at Level 4. "Low" boundary compares tutored students to peers who scored at Level 2.

Implications

For the high boundary, which compares treated students to peers who receive no special services based on their assessment scores, the results indicate that the tutoring intervention did not enhance subsequent achievement. It is worth noting that there is some evidence of "missing" eligibility tests for rising 6th graders on the Level 4 side of the high boundary. This would be consistent with higher performing 5th graders being slightly more likely to leave the public school system as they transition to middle school. That would be expected to inflate the treatment effect, which is already not distinguishable from zero. We find no similar evidence of missing eligibility tests for other grades or at the low boundary.

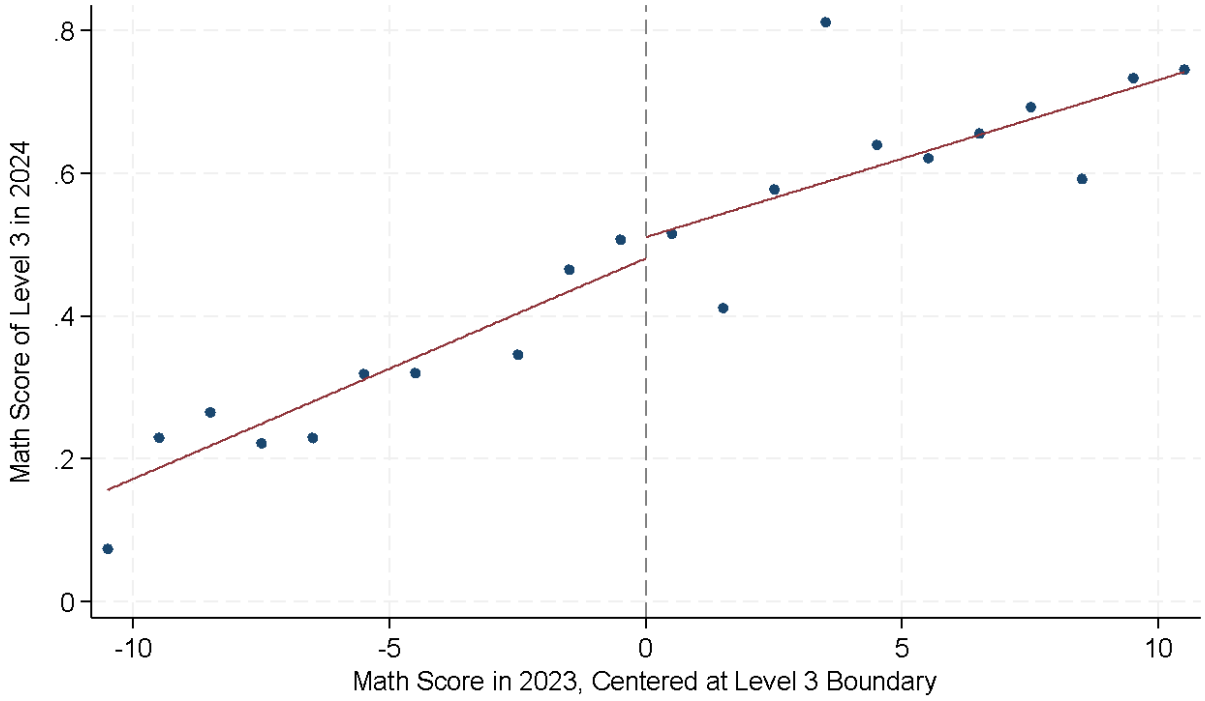
The interpretation of the null results at the low boundary is more complicated. These analyses compare treated students to their peers who receive more intensive interventions—that is, all of the district's efforts that are explicitly and implicitly targeted at the lowest performing middle school students. A null effect of this tutoring intervention at the low boundary means that the tutoring was as effective as the intensive support given to low-scoring students, the magnitude of which we are unable to estimate. It could be a large, small, or null effect. Figure D1 below makes this clear. Among students who scored near the boundary between Level 2 (assigned to intensive intervention) and Level 3 (assigned to tutoring) in their spring 2023 state standardized math assessment, roughly 50% scored at least a Level 3 the following spring regardless of which treatment regime they experienced.

For the OBC model, these results suggest that the OBC component was not enough to make a short burst of tutoring more successful in raising assessment score performance than the set of interventions received by the low-boundary comparison group or the regular instructional conditions of the high-boundary comparison group. However, this may not be surprising given that the evidence base for high-dosage tutoring is based on longer durations; research has shown that effective programs last a minimum of 10 weeks, although many last for the full school year. As noted earlier, District B had planned for its intervention to last nearly 6 months, from October through April, but implementation was stalled by delays in the contracting process.

Additionally, qualitative data gathered from interviews with district and school leaders characterized District B's 2 months of implementation during the 2023/24 school year as a period in which district leaders, school leaders, and the provider were still working through early implementation challenges and operational adjustments, as is common during the first months of a new intervention. Early challenges included curriculum alignment issues, initial resistance among some students and teachers, and occasional tutor absences, most of which were resolved within the first weeks of implementation. The most significant and persistent challenge, however, was scheduling: Tutoring was provided at the very end of the school day, and so middle school students were tired and unfocused, and some left school before their sessions began. Having identified this as a critical implementation barrier, the

district embedded tutoring within the school day in subsequent years and reported meaningfully improved implementation as a result.

Figure D1. Equal Performance of Students in Tutoring Versus Intensive Interventions



District C

District Overview

District C is a large urban charter network of elementary, middle, and high schools. During the 2023/24 and 2024/25 school years, District C entered into OBC agreements with three providers to address academic gaps in mathematics, literacy/ELA, and English language development across multiple grade levels.

In total, approximately 2,200 students received services under OBC contracts in District C during this period. Intervention 1 provided one-to-one ELA tutoring to 3rd grade students at three selected schools. Intervention 2 provided tutoring in both reading and math to students in 3rd through 5th grades at four selected schools. Intervention 3 provided an EdTech-based English language development intervention to 1st through 10th grade emergent bilingual students across all schools in the district.

Approximately 32% of treated students in District C met at least one of their targeted outcomes.

District C: Intervention 1

Intervention Overview

During the 2023/24 school year, District C provided Intervention 1 to 3rd grade students at three selected schools. The intervention was designed to deliver ELA tutoring four times a week for 20 minutes per session over approximately 20 weeks of the school year, for a planned total of roughly 27 hours of instruction per student.

The contract between the district and the tutoring provider followed an OBC structure. For each student served by the provider, the district paid a base payment of $1,050, with additional outcomes-based payments tied to end-of-year assessment results. One measure was students' conditional growth percentile (CGP), which measures a student's academic growth over a year relative to peers with similar prior performance, where a CGP of 50 indicates growth equal to that of the typical student starting from the same baseline. Additional outcomes-based payments were structured as follows:

- **$350** if a student achieved a CGP of at least 60 on the end-of-year ELA assessment
- **$580** if the student achieved a CGP of at least 66, for $930 in CGP-based payments
- **$700** if the student scored proficient on the state standardized ELA assessment

These CGP payments are cumulative: A student reaching a CGP of 66 or higher resulted in the additional payment to the provider of both the $350 and the $580. Including the base payment of $1,050, the maximum per-student payment was therefore $2,680.

Eligibility Criteria

For this tutoring intervention, eligibility was based on three criteria. To receive tutoring, a student had to

- be enrolled in 3rd grade,
- be attending one of three selected schools, and
- be scoring below the state's grade-level benchmark on the prior spring's ELA assessment.

The district selected the three participating schools based on internal judgments about academic needs across their campuses. Comparing the selected schools to non-selected schools in the district on prior spring assessments suggests that this selection aligned with measurable ELA gaps: Students in the selected schools scored significantly lower in ELA, though math differences were not statistically significant (see Table D18).

Although 165 students in the selected schools met the academic threshold, only 86 of them received tutoring, a compliance rate of approximately 52%. An additional 14 students were treated despite not meeting the threshold, for 100 treated in total. These patterns are detailed in the next section and shape the impact analysis that follows.

Table D18. Performance of Students in Selected and Non-Selected Schools

	Students in Non-Selected Schools *n* = 1,374 students	Students in Selected Schools *n* = 280 students	*P*-Value
Mean State ELA Assessment Score (Prior Spring)	.03	−.16	.01

	Students in Non-Selected Schools *n* = 1,374 students	Students in Selected Schools *n* = 280 students	*P*-Value
Mean State Math Assessment Score (Prior Spring)	.01	−.04	.46
Met Academic Eligibility Threshold	687	165	.006
Received Tutoring	0	100	–

Note. Met Academic Eligibility Threshold reflects students who scored below grade level on the prior spring state standardized assessment, regardless of school. Only students in the three selected schools were eligible to receive tutoring, so the *p*-value for Received Tutoring is omitted. Demographic data for this intervention were not available.

Descriptive Statistics

This section describes the analytic samples used in the impact analysis. Of the 1,654 3rd grade students in District C during the 2023/24 school year, 280 attended one of the three selected schools. Within those schools, 165 students met the academic eligibility threshold and 100 ultimately received tutoring. The remaining tables describe how these groups compared on prior spring assessments and how the treated group's baseline characteristics differed from the comparison group used in the impact analysis.

Within the three selected schools, students who met the academic threshold scored substantially lower than their schoolmates on both subjects assessed in the prior spring (Table D19). The ELA gap is expected—eligibility was defined by ELA performance below grade level—but the comparable math gap (a roughly 1.0 SD difference, despite math not being an eligibility criterion) suggests that academically eligible students faced broader academic challenges than literacy alone.

Table D19 also reveals that the tutoring assignment did not perfectly follow the academic eligibility rule. Of the 100 students who received tutoring, 86 met the academic threshold and 14 did not. Of the 14 students treated despite not meeting the threshold, 9 were missing a prior spring assessment score (thus the threshold could not be evaluated), and 5 scored above grade level on the prior spring ELA assessment and received tutoring anyway.

Table D19. Performance by Tutoring Eligibility (Students in Eligible Schools Only)

	Did Not Meet Academic Threshold *n* = 115 students	Met Academic Threshold *n* = 165 students	*P*-Value
Mean State ELA Assessment Score (Prior Spring)	.93	−.58	<.001
Mean State Math Assessment Score (Prior Spring)	.67	−.33	<.001
Received Tutoring	14	86	<.001

To examine the baseline characteristics of treated students relative to a comparison group, Table D20 compares the 86 treated students who met the academic threshold against the 687 academically eligible students in non-selected schools. We use this cross-school comparison group rather than the within-school alternative (the 79 academically eligible students in selected schools who were not treated) in an attempt to assemble a comparison group that was unexposed to the treatment.

Treated students had significantly lower prior state standardized ELA assessment scores than the comparison group, whereas prior state standardized math assessment scores were similar. This baseline imbalance is consistent with targeting students with the greatest literacy needs, but it also indicates that treated and comparison students differed at baseline in ways that complicate causal inference, a challenge the outcomes analysis addresses through propensity score matching.

Table D20. Baseline Performance of Treated Versus Academically Eligible Comparison

	Treated *n* = 86 students	Academically Eligible, Non-Selected Schools *n* = 687 students	*P*-Value
Mean State ELA Assessment Score (Prior Spring)	−.78	−.63	.038

	Treated *n* = 86 students	Academically Eligible, Non-Selected Schools *n* = 687 students	*P*-Value
Mean State Math Assessment Score (Prior Spring)	−.50	−.48	.83

Outcomes Analysis

We assess the impact of the tutoring intervention using four post-treatment outcomes, all measured at the end of the 2023/24 school year:

- a binary indicator for whether the student achieved a CGP between 60 and 65 on the end-of-year ELA assessment
- a binary indicator for whether the student achieved a CGP between 66 and 99 on the end-of-year ELA assessment
- a binary indicator for whether the student achieved proficiency on the end-of-year state standardized ELA assessment
- the end-of-year ELA scale score, standardized to mean 0 and standard deviation 1 across the analytic sample

The three binary outcomes correspond directly to the payment thresholds in the OBC contract; the standardized scale score provides a continuous outcome that captures performance variation within and across these thresholds. In addition to these four outcomes, we report results for a pre-treatment placebo outcome—the prior spring's (2022/23) CGP in math—to assess residual selection bias in the matched comparison. This measure was recorded before the intervention began and therefore cannot reflect any tutoring effect.

We had planned to take advantage of the eligibility criterion of lower performance on an ELA assessment, which would typically allow the use of a regression discontinuity design. The RDD is a research method with strong causal validity that can be used when access or eligibility for a program is clearly defined by a cutoff—in this case, the prior spring's state standardized ELA assessment threshold (below grade level) that determines intervention eligibility status. In this case, two features of the implementation precluded an RDD.

First, the sample of academically eligible students (*n* = 165) was too small to support a well-powered RDD. Second, compliance with the academic eligibility rule was imperfect in both directions: Of the 165 academically eligible students, only 86 (52%) received tutoring, and of the 100 students who received tutoring, 14 (14%) had not met the academic threshold.

Even a fuzzy RDD designed to accommodate imperfect compliance would have been underpowered given the combination of a small sample and bidirectional noncompliance. We instead rely on a propensity score match between the students who received treatment and those who did not. The match is based on the scale scores for end-of-year state standardized assessments in both ELA and math, along with CGP in ELA (which is not a linear transformation of the actual ELA assessment). We caution that the propensity score match should not be considered causal and will fail to completely control for selection bias between which students are treated and which students are not. For example, we are unable to add any information about why an eligible student participated or did not participate.

Table D21 below provides comparisons of the treated students with their propensity score–weighted peers in the control group. To illustrate why these results are not causal, we include the prior spring's CGP in math, taken before the treatment could possibly have had an impact, as the first outcome. As can be seen, it appears that the treated students scored, on average, nearly 8 percentile points lower on the prior spring's CGP in math than their propensity score matched peers. Because this effect cannot possibly be related to the treatment of tutoring under an OBC, it illustrates the high degree of selection bias that plagues the comparison of performance between these two groups. For the remainder of the analyses, we include the prior year's CGP in math in the set of matching variables.

As can be seen in Table D21 below, for two out of four outcomes, there are no statistically detectable differences between the treated students' performance and that of their matched peers. For standardized ELA performance, the treated group performance was a quarter standard deviation higher than the matched comparison group's performance. For 66%–99% of CGP met, the treated students performed 19 percentage points better than their matched peers. Given the "effect" of treatment on the prior spring's CGP in math, the differences in the outcomes are almost certainly due to underlying differences in the treatment and matched comparison groups rather than an effect of the tutoring.

Table D21. Differences Between Treated Students and Matched Comparison Group

	Difference	Standard Error	*P*-Value	95% Confidence Interval
MAP CGP, Spring 2023 (Placebo Outcome)	−7.95	4.62	.08	(−17.0, 1.1)
60%–65% CGP Met (End-of-Year ELA)	−.01	.03	.72	(−.08, .05)

	Difference	Standard Error	*P*-Value	95% Confidence Interval
66%–99% CGP Met (End-of-Year ELA)	.19	.07	.01	(.04, .34)
Proficiency on End-of-Year State ELA Assessment	.02	.05	.63	(−.07, .12)
End-of-Year State ELA Assessment Score	.25	.12	.04	(.01, .49)

Note. Sample size is between 761 and 767 (86 treated students and their matched peers). The first row is a pre-treatment placebo outcome included to assess residual selection bias. The remaining four rows are the impact outcomes (three contracted payment thresholds plus the standardized scale score).

Implications

In District C, Intervention 1, a low percentage of students who were eligible for tutoring actually received the treatment. Such noncompliance makes it difficult to conduct a causal analysis of the intervention's impact on student achievement. The small sample size further complicated the analysis. However, if compliance with the eligibility score had been higher in all three of District C's interventions, a stronger causal analysis could have been performed by statistically merging all three of the intervention's effects. Beyond forestalling rigorous impact analysis, noncompliance with treatment eligibility also creates misalignment between tutoring implementation and one of the central tenets of the OBC model wherein a defined group of students is identified for support based on empirical achievement data.

Qualitative data suggest that implementation of this intervention went smoothly, although there were some challenges related to the selection of an outcome measure for the contract. The evaluation team interviewed the provider and one leader from this district; the district leader was not connected to this intervention, so only qualitative data from the provider are included here.

The provider described that a high level of collaboration between the district and provider supported strong attendance and school-level engagement. From the provider's perspective, the district served as an engaged, highly collaborative partner throughout implementation. The provider met biweekly with district leadership and either weekly or biweekly with school leaders, and the provider noted that when schools were not meeting expectations around attendance or follow-through, district leadership intervened. They observed, "If things weren't going quite right at a school, we could lean on [our district contact] to put pressure on the school" to fully uphold their implementation responsibilities.

However, misalignment between provider and district assessment tools created challenges with measuring outcomes. The provider used one common assessment tool as their primary literacy assessment, whereas the district prioritized another. Because one of these assessment tools was more comprehension-focused and the other emphasized foundational skills, there was no straightforward alignment between what the provider was teaching and the measures the district used to evaluate outcomes. This required explicit negotiation during contracting around how to set goals using the district's selected assessment tool when there were no baseline data. The provider noted that this experience pushed them to think more carefully about which assessment measures to use for older elementary students in the future and how to balance instruction focusing on foundational skills versus comprehension when serving those grade levels.

District C: Intervention 2

Intervention Overview

During the 2023/24 school year, District C provided Intervention 2 (tutoring) for both reading and math. Reading tutoring was provided to 3rd and 4th grade students at three selected schools, and math tutoring was provided to 3rd, 4th, and 5th grade students at three selected schools. The intervention was designed to deliver tutoring four times a week over approximately 24 weeks of the school year.

The contract between the district and the tutoring provider followed an OBC structure. For each student served by the provider, the district paid a base payment of $1,050, with additional outcomes-based payments tied to end-of-year assessment results. One metric was the student's CGP, which measures a student's academic growth over a year relative to peers with similar prior performance, where a CGP of 50 indicates growth equal to that of the typical student starting from the same baseline. The additional outcomes-based payments were as follows:

- **$350** if a student achieved a CGP of at least 60 on the tutored subject end-of-year assessment
- **an additional $580** (for $930 total in CGP-based payments) if the student achieved a CGP of at least 66
- **$700** if the student scored proficient on the tutored subject's state standardized assessment

These CGP payments are cumulative: A student reaching a CGP of 66 or higher resulted in the additional payment to the provider of both the $350 and the $580. Including the base payment of $1,050, the maximum per-student payment was therefore $2,680.

Eligibility Criteria

For this tutoring intervention, eligibility was based on three criteria.

To receive reading tutoring, a student had to

- be enrolled in 3rd or 4th grade,
- be attending a selected school, and
- be scoring below the state's grade-level benchmark on the prior spring's state standardized ELA assessment.

To receive math tutoring, a student had to

- be enrolled in 3rd, 4th, or 5th grade;
- be attending a selected school; and
- be scoring below the state's grade-level benchmark on the prior spring's state standardized math assessment.

A total of four schools participated in the intervention: One school provided only reading tutoring, one provided only math tutoring, and two provided both reading and math tutoring. The district selected these schools based on internal judgments about academic needs across their campuses. Within the participating schools, the specific grade levels served by each intervention varied—not every school offered tutoring to every grade in the eligible range. Table D22 below summarizes the intervention coverage at each participating school.

In practice, a student's eligibility depended not only on grade and prior assessment performance but also on which grade levels their specific school served for each intervention. For example, a 3rd grader at School D was not eligible for reading tutoring even though the criteria above include 3rd graders because School D offered reading tutoring only to 4th graders.

Table D22. Intervention Coverage by School and Grade Level

School	Reading Tutoring Offered to	Math Tutoring Offered to
School A	Not offered	4th and 5th graders
School B	3rd graders	3rd and 4th graders
School C	3rd and 4th graders	4th graders
School D	4th graders	Not offered

The selected schools had significantly lower prior spring ELA and math scores than non-selected schools, consistent with the district's targeting of schools with broader academic needs. Tables D23 and D24 below describe these differences separately for the reading and math intervention samples.

Table D23. Performance of Students in Selected and Non-Selected Schools—Reading Tutoring Sample (Grades 3–4)

	Students in Non-Selected Schools *n* = 2,861 students	Students in Selected Schools *n* = 511 students	*P*-Value
Mean State ELA Assessment Score (Prior Spring)	.03	−.16	<.001
Mean State Math Assessment Score (Prior Spring)	.03	−.19	<.001
Met Reading Eligibility Threshold	1,363 (57%)	276 (67%)	<.001
Received Tutoring	0 (0%)	107 (21%)	–

Note. Sample includes all students in grades 3 and 4. Met Reading Eligibility Threshold reflects students who scored below grade level on the prior spring's state standardized ELA assessment, regardless of school. The percentages reflect the proportion of students with no missing eligibility information, not the total number of students. Only students in selected schools were eligible to receive reading tutoring, so the *p*-value comparison for tutoring receipt is omitted. Demographic data for this intervention were not available.

Table D24. Performance of Students in Selected and Non-Selected Schools—Math Tutoring Sample (Grades 3–5)

	Students in Non-Selected Schools *n* = 4,393 students	Students in Selected Schools *n* = 672 students	*P*-Value
Mean State Math Assessment Score (Prior Spring)	.02	−.13	<.001
Mean State ELA Assessment Score (Prior Spring)	.02	−.14	<.001
Met Math Eligibility Threshold	2,156 (59%)	368 (68%)	<.001
Received Tutoring	0 (0%)	127 (19%)	–

Note. Sample includes all students in grades 3 and 4. Met Reading Eligibility Threshold reflects students who scored below grade level on the prior spring's math assessment, regardless of school. The percentages reflect the proportion of students with no missing eligibility information, not the total number of students. Only students in selected schools were eligible to receive math tutoring, so the *p*-value comparison for tutoring receipt is omitted. Demographic data for this intervention were not available.

Descriptive Statistics

This section describes the analytic samples used in the impact analysis. The two interventions are described in parallel: first within-selected-school comparisons (Tables D25 and D26), then comparisons between treated students and the academically eligible comparison group used in the impact analysis (Tables D27 and D28).

Within Selected Schools

Within the participating schools, students who met the academic threshold for either intervention scored substantially lower than their schoolmates on prior spring assessments in both the eligibility-defining subject and the other subject (Tables D25 and D26). For reading, academically eligible students scored 1.4 standard deviations lower in ELA and roughly 0.9 standard deviations lower in math than their ineligible school peers. For math, academically eligible students scored 1.3 standard deviations lower in math and 1.0 standard deviations lower in ELA. The within-subject gap is expected—eligibility was defined by performance in that subject—but the parallel gap in the other subject suggests that academically eligible students faced broader academic challenges than any single subject.

Tables D25 and D26 also reveal that tutoring assignments did not perfectly follow the academic eligibility rules. Of the 107 students who received reading tutoring, 91 met the reading academic threshold, 4 scored above the threshold and received tutoring anyway, and 12 were missing a prior spring ELA score (thus the threshold could not be evaluated). Of the 127 students who received math tutoring, 112 met the math academic threshold, 4 scored above the threshold and received tutoring anyway, and 11 were missing a prior spring state standardized math score.

Table D25. Performance of Students in Selected Schools—Reading Tutoring Eligibility (Grades 3–4)

	Did Not Meet Academic Threshold *n* = 135 students	Met Academic Threshold *n* = 276 students	*P*-Value
Mean State ELA Assessment Score (Prior Spring)	.79	−.63	<.001
Mean State Math Assessment Score (Prior Spring)	.43	−.49	<.001
Received Tutoring	4 (3%)	91 (33%)	<.001

Note. Students who met the academic threshold include only students in the selected grades in the selected schools. See Table D22 for a breakdown of which grades each school provided the tutoring for.

Table D26. Performance of Students in Selected Schools—Math Tutoring Eligibility (Grades 3–5)

	Did Not Meet Academic Threshold *n* = 177 students	Met Academic Threshold *n* = 368 students	*P*-Value
Mean State Math Assessment Score (Prior Spring)	.75	−.56	<.001

	Did Not Meet Academic Threshold *n* = 177 students	Met Academic Threshold *n* = 368 students	*P*-Value
Mean State ELA Assessment Score (Prior Spring)	.54	−.46	<.001
Received Tutoring	4 (2%)	112 (30%)	<.001

Note. Students who met the academic threshold only include students in the selected grades in the selected schools. See Table D22 for a breakdown of which grades each school provided the tutoring for.

Treated Students Versus Comparison Group

To examine the baseline characteristics of treated students relative to a comparison group, Tables D27 and D28 compare treated students who met the academic threshold against academically eligible students in non-selected schools. We use this cross-school comparison group rather than the within-school alternative—the academically eligible students in selected schools who were not treated—in an attempt to minimize the impact of selection bias in the analysis.

For the reading intervention, treated students had significantly lower prior state ELA assessment scores than the comparison group and also somewhat lower prior state math assessment scores. For the math intervention, treated students and the comparison group looked similar at baseline on both subjects, with no statistically significant differences. The reading baseline imbalance is consistent with targeting students with the greatest literacy needs within the eligible pool, but it also indicates that treated and comparison reading students differed at baseline in ways that complicate causal inference—a challenge the outcomes analysis addresses through propensity score matching. The relative balance in the math comparison provides somewhat firmer ground for the math impact analysis, though selection on unobservable factors remains a concern.

Table D27. Baseline Performance of Treated Reading Students Versus Academically Eligible Comparison Group

	Treated *n* = 91 students	Academically Eligible, Non-Selected Schools *n* = 1,363 students	*P*-Value
Mean State ELA Assessment Score (Prior Spring)	−.87	−.65	.002
Mean State Math Assessment Score (Prior Spring)	−.68	−.50	.047

Note. The comparison group is restricted to students in non-selected schools (grades 3–4) who scored below the grade-level benchmark on the prior spring's state ELA assessment.

Table D28. Baseline Performance of Treated Math Students Versus Academically Eligible Comparison Group

	Treated *n* = 112 students	Academically Eligible, Non-Selected Schools *n* = 2,156 students	*P*-Value
Mean State Math Assessment Score (Prior Spring)	−.64	−.57	.36
Mean State ELA Assessment Score (Prior Spring)	−.49	−.46	.66

Note. The comparison group is restricted to students in non-selected schools (grades 3–5) who scored below the grade-level benchmark on the prior spring's state math assessment.

Outcomes Analysis

We assess the impact of the tutoring intervention using four post-treatment outcomes, all measured at the end of the 2023/24 school year on the tutored subject:

- a binary indicator for whether the student achieved a CGP between 60 and 65 on the end-of-year state assessment of the tutored subject

- a binary indicator for whether the student achieved a CGP between 66 and 99 on the end-of-year state assessment on the tutored subject
- a binary indicator for whether the student achieved proficiency on the end-of-year state assessment on the tutored subject
- the end-of-year scale score for the tutored subject, standardized to mean 0 and standard deviation 1 across the analytic sample

The three binary outcomes correspond directly to the payment thresholds in the OBC contract; the standardized scale score provides a continuous outcome that captures performance variation within and across these thresholds. In addition to these four outcomes, we report results for a pre-treatment placebo outcome—the prior spring's (2022/23) CGP in math or ELA—to assess residual selection bias in the matched comparison. This measure was recorded before the intervention began and therefore cannot reflect any tutoring effect.

We had planned to take advantage of the eligibility criterion of lower performance on an assessment of ELA and math, which would typically allow the use of a regression discontinuity design. RDD is a research method with strong causal validity that can be used when access or eligibility for a program is clearly defined by a cutoff—in this case, the prior spring assessment for the tutored subject threshold (below grade level) that determines intervention eligibility status. However, because of both a small sample and imperfect compliance with the assessment score eligibility criterion, we were unable to use an RDD. The number of eligible students for this intervention was 276 for ELA and 368 for math, which is near the lower boundary for a well-powered RDD. Moreover, around half of the eligible students did not participate in the tutoring, and a handful of ineligible students did participate in the tutoring, which would make even a large sample RDD less precise.

We instead rely on a propensity score match between the students who received treatment (and were eligible for it) and those who were in non-participating schools but had prior performance that would have made them eligible. The match is based on the scale scores for end-of-year state standardized assessments in both ELA and math, along with CGPs in ELA and math (which are not linear transformations of the actual ELA and math tests). We caution that the propensity score match should not be considered causal and will fail to completely control for selection bias between which students are treated and which students are not. For example, we are unable to add any information about why an eligible student participated or did not participate.

As can be seen in Table D29 below, on all four outcomes, the students who received tutoring for ELA had lower outcomes than their statistically matched peers. Two of these are all distinguishable at conventional levels of statistical certainty. In Table D30, the outcome differences between students who received tutoring in math and their matched peers were not statistically distinguishable from zero. Given the inability to sufficiently control for

selection bias in which students received tutoring (among those eligible), the differences in the outcomes for the ELA tutoring sample are almost certainly due to underlying differences in the treatment and matched comparison groups rather than to negative effects of the tutoring.

Table D29. Differences Between Treated Reading Students and Matched Comparison Group

Outcome	Difference	Standard Error	*P*-Value	95% Confidence Interval
End-of-Year State ELA Assessment Score	−0.03	0.10	.79	(−.22, .17)
60%–65% CGP Met (End-of-Year ELA)	−0.07	0.03	.04	(−.13, −.01)
66%–99% CGP Met (End-of-Year ELA)	−0.06	0.06	.38	(−.18, .07)
Proficiency on End- of-Year State ELA Assessment	−0.08	0.03	.01	(−.14, −.02)

Note. Sample size is 1,442 (91 treated, 1,351 matched comparison). Row 1 is the standardized end-of-year scale score (a continuous analytic outcome); the remaining three rows are the contracted payment-threshold outcomes; no separate placebo is shown.

Table D30. Differences Between Treated Math Students and Matched Comparison Group

Outcome	Difference	Standard Error	*P*-Value	95% Confidence Interval
End-of-Year State Math Assessment Score	0.04	0.09	.63	(−.13, .22)
60%–65% CGP Met (End-of-Year Math)	0.02	0.03	.52	(−.04, .08)
66%–99% CGP Met (End-of-Year Math)	−0.07	0.07	.36	(−.21, .08)

Outcome	Difference	Standard Error	*P*-Value	95% Confidence Interval
Proficiency on End- of-Year State Math Assessment	0.01	0.04	.81	(−.07, .09)

Note. Sample size is 2,226 (112 treated, 2,114 matched comparison). Row 1 is the standardized end-of-year scale score (a continuous analytic outcome); the remaining three rows are the contracted payment-threshold outcomes; no separate placebo is shown.

Implications

In District C, Intervention 2, a low percentage of students who were eligible for tutoring actually received the treatment. Such noncompliance makes it difficult to conduct a causal analysis of the intervention's impact on student achievement. The modest sample size further complicated the analysis. However, if compliance with the eligibility score had been higher in all three of District C's interventions, a stronger causal analysis could have been performed by statistically merging all three of the intervention's effects. Beyond forestalling rigorous impact analysis, noncompliance with treatment eligibility also creates misalignment between tutoring implementation and one of the central tenets of the OBC model wherein a defined group of students is identified for support based on empirical achievement data.

Compliance with eligibility criteria aside, qualitative data suggest that implementation of this intervention experienced both successes and challenges. The evaluation team interviewed the provider and one leader from this district, although the district leader was not connected to this intervention, so only qualitative data from the provider are included here.

The provider reported that a high level of collaboration between the district and provider supported strong attendance and school-level engagement. From the provider's perspective, the district served as an engaged, highly collaborative partner throughout implementation. District leadership and the provider communicated regularly in both formal meetings and informal check-ins, leading to structural improvements in the provider's support throughout the contract. For example, as a result of these conversations, the provider developed a structured practice of distributing daily school-level program attendance data each morning, which likely contributed to the program's high attendance rate. The attendance reports were sent to a central office point of contact, who then served as an intermediary with individual schools, flagging attendance issues and prompting schools to take action.

The provider also periodically joined professional learning community (PLC) sessions with grade-level teams to review data and build school-level buy-in, though these sessions were occasionally canceled. This layered collaboration structure was seen as a meaningful driver of strong attendance and implementation fidelity.

However, the provider also described implementation challenges. One challenge was that the district experienced teacher turnover in some classrooms where tutoring was taking place, resulting in substitutes who were unfamiliar with the tutoring program. Additionally, a midyear curriculum change meant that tutors who had already established relationships with student groups were no longer aligned with the new curriculum, requiring tutor turnover and regrouping. The district and the provider agreed upon an approach to adjusting the outcome metrics as a result, but the change created unexpected volatility mid-program.

In addition, some schools canceled tutoring sessions in the final month of the program without prior notice. The provider noted that canceling 3 sessions over 4 weeks amounted to 12 lost sessions, meaningfully reducing the time available for students to achieve intended outcomes.

District C: Intervention 3

Intervention Overview

During the 2024/25 school year, District C provided Intervention 3 to 1st through 10th grade emergent bilingual students at all schools across their district. The EdTech intervention was designed to deliver one 45-minute session a week for 28 weeks.

The contract between the district and the EdTech provider followed an OBC structure. For each student served by the provider, the district paid a base payment of $23.40 with additional outcomes based payments tied to performance on the English language assessment:

- **$22.75** if a student achieved one level of growth
- **$97.50** if a student achieved two levels of growth

The maximum per-student payment under this contract was **$143.65.**

Eligibility Criteria

For this EdTech intervention, eligibility was based on three criteria. To receive the intervention, a student had to

- be emergent bilingual,
- be enrolled in 1st through 10th grade, and
- be scoring beginning or intermediate on the state English language assessment.

In schools where more students qualified than EdTech intervention spots were available, schools used discretion to decide which students received the intervention. The district reported that schools considered things like state assessment results, attendance, and special education status when deciding which students would get the intervention.

Descriptive Statistics

This section describes the analytic samples used in the impact analysis. Of the 6,585 emergent bilingual students in 1st through 10th grade in District C during the 2024/25 school year, 2,778 met the eligibility threshold and 1,829 ultimately received the intervention. The remaining tables describe how these groups compared on prior spring English language assessments and how the treated group's baseline characteristics differed from the comparison group used in the impact analysis.

Students who met the eligibility threshold scored substantially lower than their school peers on both the English language assessment and the state ELA assessment in the prior spring (Table D31). The English language assessment gap is expected; eligibility was defined by English language assessment performance intermediate or below. The gap in state reading scores is notable as well; though the state reading assessment is administered to all students and was not an eligibility criterion, eligible students still performed meaningfully lower than their ineligible peers.

Table D31. Performance and Demographics by Intervention Eligibility

	Did Not Meet Eligibility Threshold *n* = 3,807 students	Met Eligibility Threshold *n* = 2,778 students	*P*-Value
English Language Assessment Mean Standardized Score	.62	−.77	<.001
English Language Assessment Mean Composite Score	2.97	1.80	<.001
State Assessment Mean Reading Score	202.5	178.2	<.001
Socioeconomically Disadvantaged	3,199 (87%)	2,478 (90%)	<.001

	Did Not Meet Eligibility Threshold *n* = 3,807 students	Met Eligibility Threshold *n* = 2,778 students	*P*-Value
Hispanic	3,430 (90%)	2,645 (95%)	<.001
Students Receiving Special Education Services	246 (6.5%)	453 (16%)	<.001
Treated	420 (11%)	1,409 (51%)	–

Note. Each percent above reflects the percent of students who had valid data for each demographic.

To examine the baseline characteristics of treated students relative to a comparison group, Table D32 compares the 1,409 treated students who met the academic threshold against the 1,369 academically eligible students who did not receive the intervention. Treated students had significantly lower prior English language assessment scores than the comparison group, while socioeconomic status was similar. This baseline imbalance is consistent with targeting students with the greatest literacy needs, but it also indicates that treated and comparison students differed at baseline in ways that complicate causal inference, a challenge the outcomes analysis addresses through propensity score matching.

Table D32. Baseline Performance of Treated Versus Eligible Comparison

	Treated *n* = 1,409 students	Eligible but Not Treated *n* = 1,369 students	*P*-Value
English Language Assessment Mean Standardized Score	−.90	−.63	<.001
English Language Assessment Mean Composite Score	1.70	1.90	<.001
State Assessment Mean Reading Score	173.9	183.5	<.001
Socioeconomically Disadvantaged	1,269 (91%)	1,209 (89%)	.17

	Treated *n* = 1,409 students	Eligible but Not Treated *n* = 1,369 students	*P*-Value
Hispanic	1,349 (96%)	1,296 (95%)	.025
Students Receiving Special Education Services	208 (15%)	245 (18%)	<.001

Outcomes Analysis

We assessed the impact of the EdTech intervention using three post-treatment outcomes, all measured at the end of the 2024/25 school year on English language development:

- a binary indicator for whether the student achieved one level of composite growth on the state English language assessment
- a binary indicator for whether the student achieved two levels of composite growth on the state English language assessment
- the end-of-year English language assessment reading domain scale score, standardized to mean 0 and standard deviation 1 across the analytic sample

The two binary outcomes correspond directly to the payment thresholds in the OBC contract; the standardized English language assessment reading domain scale score provides a continuous outcome that captures performance variation within and across these thresholds. In addition to these three outcomes, we report results for a pre-treatment placebo outcome—the prior spring's (2023/24) English language assessment reading domain scale score (standardized)—to assess residual selection bias in the matched comparison. This measure was recorded before the intervention began and therefore cannot reflect any intervention effect.

We had planned to take advantage of the eligibility criterion of lower performance on the English language assessment, which would typically allow the use of a regression discontinuity design. RDD is a research method with strong causal validity that can be used when access or eligibility for a program is clearly defined by a cutoff—in this case, the prior spring's English language assessment threshold (beginning or intermediate) that determines intervention eligibility status. However, because of imperfect compliance with the assessment score eligibility criterion, particularly near the eligibility threshold, we were unable to use an RDD. Approximately one third of students who were just barely eligible actually received treatment, whereas around 15% of students who were just barely ineligible received the treatment. That jump in treatment participation at the assignment threshold is not large enough to support a well-powered RDD.

We instead rely on a propensity score match between the students who received treatment (and were eligible for it) and those who did not receive treatment but were nonetheless eligible. The match is based on the scale scores for the end-of-year state assessment in ELA and the English language assessment composite score, as well as binary indicators for Hispanic, Asian, Black, special education status, and socioeconomic disadvantage. We caution that the propensity score match should not be considered causal and will fail to completely control for selection bias between which students are treated and which students are not. For example, we are unable to add any information about why an eligible student participated or did not participate.

As can be seen in Table D33 below, the placebo test confirms that the treated students had prior-year English language assessment reading domain scale scores that were, on average, similar to that of their matched peers. This provides reassurance that the two groups did not have dramatically different performance profiles at the outset of the intervention. Two of the three outcomes—one level of composite growth on the state English language assessment and the standardized English language assessment reading domain scale score—showed no statistically significant difference between the treated students and their matched peers. On the next outcome, the treated students did slightly better than their matched peers: They were 1 percentage point more likely to achieve two levels of composite growth on the state English language assessment. This difference is statistically distinguishable from zero near conventional levels ($p = 0.06$). Given the inability to sufficiently control for selection bias in which students received the EdTech intervention among those eligible, this difference could reflect underlying differences between the treatment and matched comparison groups rather than a true effect of the intervention.

Table D33. Differences Between Treated Students and Matched Comparison Group

Outcome	Difference	Standard Error	*P*-Value	95% Confidence Interval
Prior Spring English Language Assessment Standardized Score	−0.033	0.034	.33	(−.10, .03)
One Level of Composite Growth on English Language Assessment	0.001	0.031	.98	(−.06, .06)
Two Levels of Composite Growth on English Language Assessment	0.010	0.005	.06	(.00, .02)

Outcome	Difference	Standard Error	*P*-Value	95% Confidence Interval
Current Spring English Language Assessment Standardized Score	−0.024	0.040	.51	(−.10, .05)

Note. Sample size is 1,815 (1,027 treated who had sufficient background data and 788 matched comparison). The first row is a pre-treatment placebo outcome included to assess residual selection bias. The remaining three rows are the impact outcomes (two contracted payment thresholds plus the standardized scale score).

Implications

In District C, Intervention 3, a low percentage of students who were eligible for the EdTech intervention actually received the treatment (especially near the eligibility threshold). Such noncompliance makes it difficult to conduct a causal analysis of the intervention's impact on student achievement. Beyond forestalling rigorous impact analysis, noncompliance with treatment eligibility also creates misalignment between intervention implementation and one of the central tenets of the OBC model wherein a defined group of students is identified for support based on empirical achievement data.

Compliance with eligibility criteria aside, qualitative data suggest that implementation of this intervention experienced some successes and several notable challenges. The evaluation team interviewed two leaders at the school level (educators who served as school champions) and one leader from the district.

The most notable success was that schools' use of the EdTech intervention evolved over time, with educators gradually improving integration and alignment of the intervention with their in-person instruction. At first, teachers treated the intervention primarily as a stand-alone computer program, but over time, they began to incorporate the intervention into instruction and tailor its use to classroom needs. For example, one district leader noted that early feedback reflected teachers' desire to remain actively engaged in instruction: "Teachers still want to be teachers, [but] this was a computer platform that they thought was replacing them." Later in the year, district staff conducted learning walks and observed more engaging lessons in which teachers had "figured out how to make this their own," adding that "the fact that they are owning this instruction now is really impactful for kids."

School leaders similarly described teaching brief "mini lessons" aligned with intervention content to support students in meeting learning goals. These educators reported using the platform's data to identify areas in which students needed additional help and to plan targeted instruction. One school leader noted that they designed lessons to address gaps when a student needed more support and the platform alone was not sufficient.

In terms of challenges, one notable difficulty was that although the selection of the EdTech intervention was informed by perceived alignment between the intervention and the state English language proficiency exam for emergent bilingual students, leaders observed gaps in the ability of the intervention to meet the needs of some students who were selected to receive it.

School leaders noted that some student groups, especially newcomer students, needed more differentiated instruction than they felt the EdTech program provided. School leaders described portions of the program (e.g., reading activities) as developmentally mismatched for some students, and they reported that a virtual program was sometimes overstimulating for newcomer students who had limited prior experience with computers. One leader shared that this mismatch sometimes required providing additional, off-platform support to help students access the instructional content, reducing the amount of time students spent with the program.

Another notable challenge was that educators reported tension between the dosage requirements, the limited amount of instructional time available, and students' other immediate needs for individualized instruction. School champions reported feeling pressure to meet the required number of minutes on the platform. These leaders described challenges with meeting dosage expectations (i.e., ensuring students completed a specified number of minutes in the program each week), and they reported difficulty balancing sufficient time on the platform with the in-person instructional support that they perceived was necessary to meet students' individual needs.

These school leaders noted that in some cases they felt that they had to prioritize students' minutes on the intervention over individualized, in-person support that would have more closely aligned with the student's immediate learning needs. One leader observed that the "goals and outcomes of the intervention were not realistic with the amount of time we actually have with students in the classroom."

Another challenge was that the district rolled out the intervention to all school sites in the district simultaneously, which may have impacted implementation success. Leaders noted that this districtwide launch made it challenging to coordinate staffing the intervention adequately across campuses, build teacher buy-in, and provide clear, timely guidance to school staff about classroom-level implementation. In the district's view, the scale of the rollout also contributed to communication challenges between the district and teachers.

In addition, midway through the first year of implementation, the district's implementation lead transitioned into a new role. The district leader reported that this transition affected their ability to coordinate and oversee implementation, noting that it made it difficult to "really lean into opportunities where I think I would have otherwise" without the role change.

District D

District Overview

District D is a large urban district serving a diverse student population. During the 2024/25 school year, District D entered into an OBC agreement with an EdTech provider to address early literacy gaps among the district's 1st grade students. Approximately 51% of treated students met at least one of their targeted literacy outcomes, and 18% met all of them.

Intervention Overview

During the 2024/25 school year, District D provided an EdTech intervention to all 1st grade students across the district who had not mastered the end-of-year Kindergarten High Frequency Words (KHFW) assessment in spring 2024. The intervention was designed to deliver personalized early literacy practice through an EdTech platform for 35 minutes per week over 21 weeks of the school year. This amounted to a planned total of roughly 12 hours of instruction per student.

For each student assigned to the provider, the district paid a base payment of $23, with additional outcomes-based payments tied to performance on the end-of-year state ELA assessment:

- **$11** if a student was performing at or above grade level
- **$3** if a student met their typical growth target
- **$19** if a student met their stretch growth target

Because the stretch growth target subsumes the typical growth target, a student meeting stretch growth earns both growth payments; the maximum per-student payment under this contract was therefore $56.

Eligibility Criteria

The eligibility criteria for this intervention were straightforward. To be eligible, students had to be enrolled in 1st grade and must not have achieved a perfect score (out of 40) on the end-of-year KHFW assessment in spring 2024. Students enrolled in dual language programs were excluded.

Of the 1,709 students who met the eligibility threshold, 1,693 ultimately received the intervention—a compliance rate of approximately 99%. An additional 632 students received the intervention despite not meeting the eligibility threshold, bringing the total number of

treated students to 2,325. The presence of these students is discussed further in the Descriptive Statistics and Outcomes Analysis sections.

Descriptive Statistics

As expected given the eligibility criteria, students who qualified for the intervention had lower prior KHFW scores than peers in the same schools who did not qualify. Table D34 provides a comparison of performance and demographic characteristics between eligible and ineligible students in the same schools.

Table D34. Performance and Demographics by Eligibility

	Ineligible Students *n* = 2,597 students	Eligible Students *n* = 1,709 students	*P*-Value
Mean KHFW Score (Standardized)	0.50	−0.48	<0.001
English Learner	391 (15%)	382 (22%)	<0.001
Special Education	86 (3%)	69 (4%)	0.21
Hispanic	1,628 (63%)	1,214 (71%)	<0.001
Treated	632 (24%)	1,693 (99%)	<0.001

Note. A total of 632 ineligible students received the intervention and are reflected in the ineligible student counts.

As shown in Table D34, eligible and ineligible students differed significantly on prior KHFW performance and several demographic characteristics. Eligible students had higher shares of English Learners (22% vs. 15%) and Hispanic students (71% vs. 63%) compared to their ineligible peers in the same schools, suggesting that the students most in need of early literacy support were disproportionately drawn from historically underserved populations. Rates of special education enrollment were similar across both groups (4% vs. 3%), and this difference was not statistically significant.

Outcomes Analysis

We use four outcomes for this assessment of the EdTech intervention. The first three are derived from end-of-year state standardized ELA assessments provided by the EdTech

platform: one indicating the student is on grade level, one indicating that typical growth has been met, and another indicating that stretch growth has been met. The fourth outcome is the scale score for the end-of-year diagnostic test, on which the prior three binary indicators are based. We standardize the scale score to have a mean of 0 and standard deviation of 1. The three binary outcomes are linked directly to the outcomes based contract for EdTech in the district.

For clarity, three distinct measurement points are referenced throughout: the KHFW assessment, administered at the end of kindergarten in spring 2024 and used for eligibility; the beginning-of-year 1st grade diagnostic, administered in fall 2024 before the intervention began; and the end-of-year 1st grade diagnostic, administered in spring 2025.

We had planned to take advantage of the eligibility criterion of lower performance on the KHFW test, which would typically allow the use of an RDD, a method with strong causal validity that applies when eligibility is defined by a score cutoff. Here, however, eligibility required any score below a perfect 40, so there are no students above the threshold to serve as a comparison group, making an RDD impossible.

Instead, we rely on a propensity score match between the students who received treatment and those who did not. The match is based on the scale score for the end-of-year state standardized assessment in the prior spring and a set of demographic indicators: Asian, Black, Hispanic, White, English Learner status, and special education status. We caution that the propensity score match should not be considered causal and will fail to completely control for selection bias between which students are treated and which students are not. For example, we are unable to add the actual eligibility criterion to the list of control variables because there is zero overlap in performance between the treated and comparison groups of students.

Table D35 below provides comparisons of the treated students to their propensity score–weighted peers in the control group. To illustrate why these results are not causal, we include the beginning-of-year diagnostic, taken before the treatment could plausibly have had an impact, as the first outcome. As can be seen, it appears that the treated students did substantially worse on the diagnostic than their propensity score–matched peers. Because this effect cannot plausibly be related to the EdTech intervention treatment under the outcomes based contract, it illustrates the high degree of selection bias that influences the comparison of performance between these two groups. For the remainder of the analyses, we include the beginning-of-year diagnostic in the set of matching variables.

As can be seen, the treated students did worse on all outcome measures than their propensity score–matched peers. This is not surprising, as the treated students performed worse on a test of high-frequency word count in a way that we were unable to control for in the matching procedure. These figures are not the true impact of the program and merely reflect underlying differences in the students who were assigned to the treatment and

control conditions. Given these limitations, the findings presented here should be interpreted as descriptive rather than causal and contextualized alongside the qualitative implementation data presented in the Implications section.

Table D35. Differences Between Treated Students and Matched Comparison Group

	Impact	Standard Error	*P*-Value	Confidence Interval
Beginning-of-Year Diagnostic	−0.36	0.04	<0.001	(−.43, −.29)
End-of-Year Diagnostic	−0.31	0.04	<0.001	(−.40, −.23)
On Grade Level	−0.20	0.03	<0.001	(−.27, −.14)
Met Typical Growth	−0.24	0.03	<0.001	(−.30, −.18)
Met Stretch Growth	−0.16	0.03	<0.001	(−.22, −.10)

Note. The Beginning-of-Year Diagnostic row reflects a match on the pre-treatment diagnostic score and demographic indicators only. It is a pre-treatment measure included as a placebo: Absent selection bias, treated and matched-comparison students should not differ on it, so the observed gap reflects residual selection bias rather than any program effect. All subsequent rows are outcome measures from a propensity match. Sample size is between 3,680 and 3,699, depending on how many matched peers were selected.

Implications

These results further reinforce the difficulty in drawing causal inferences about the interventions given under an outcomes based contract. The word count test that District D used was sensible in separating students who need additional academic help from those who do not, but it unfortunately precludes the use of a strong causal design. In the future, districts might consider employing an eligibility criterion that allows for such designs. In the case of District D, that would have been a cut point on the scale score with sufficient observations on the eligible and ineligible sides.

Qualitative data collected from in-person observations and interviews with district leaders, school leaders, and the provider suggest that beyond the eligibility criteria constraints described in the Outcomes Analysis section, additional implementation factors would likely have made it difficult to accurately estimate the impact of OBC. In particular, although the district included all schools within the OBC contract, the degree to which schools actually participated in OBC varied; in addition, within schools that were fully

engaged in OBC participation, the treatment of treated students and the comparison group was virtually identical.

The evaluation team observed that some schools were implementing the EdTech intervention with a high degree of fidelity, down to consistent classroom-level routines. The team visited two schools that were identified as fully committed to OBC participation, and the team found robust evidence of OBC practices that led to strengthened implementation. These practices included consistent and well-integrated classroom-level implementation routines; school staff, district staff, and providers using data to monitor progress and make implementation refinements; and collaborative problem-solving as challenges arose.

For example, in these schools, teachers and students consistently followed established procedures to ensure that students received the intended dosage and that progress was regularly tracked. Students knew to collect their laptops and begin the EdTech intervention at the expected time, projector screens displayed timers to mark the intervention duration, students stayed on task, and teachers used visible tracking sheets to monitor students' weekly participation.

School leaders and teachers reported that the visibility of real-time data on academic progress supported motivation and buy-in among both teachers and students for high-fidelity implementation. OBC also appeared to prompt practices that deepened student engagement with the intervention and with their own learning progress through collaborative goal setting and frequent feedback. Across observations and interviews, these student-level practices—routines, goal setting, and frequent feedback—were consistently described as contributing to a culture that emphasizes and celebrates student growth.

However, these schools provided similar treatment to students who were targeted for the OBC intervention and those who were not, making it difficult to disentangle the effect of OBC on student outcomes. All students received the EdTech intervention under similar conditions, within the same classroom environments. Furthermore, district and school leaders noted that they approached the data collection and progress monitoring similarly for both OBC and non-OBC students—for example, by setting individual goals for each student. They shared that this was because having all students succeed is their first priority, so they wanted to provide all students with access to supportive implementation practices. While such decisions to prioritize students' needs over the design of the contract are understandable, these uncontrolled factors complicate the impact analysis's ability to estimate OBC's effect on student outcomes.

In addition, district leaders reported that school sites' autonomy led to wide variation in the degree to which schools participated in the intervention and OBC-related processes. In District D, all schools are included in the OBC contract. However, district leaders noted that school leaders have expressed varying degrees of commitment and are implementing the EdTech intervention with varying degrees of fidelity—so varied that they described some

schools as opting in to OBC participation and implementing with high levels of fidelity, whereas other schools opted to first observe their peers' participation before potentially committing to participate in a future year (despite all schools technically being considered as part of the current OBC treatment pool).

School staff reported receiving little guidance and few expectations about the intervention and how to implement it with fidelity. One school leader reported receiving little direction from the district on implementation of the EdTech intervention and did not know how to access support from the provider. The provider echoed these concerns, noting that some principals and teachers were not aware that the intervention operated under an outcomes based contract or what the usage expectations were. The provider viewed this information as important for building staff buy-in and helping staff understand the importance of monitoring implementation.

The school leader reported that staff from their school faced challenges in monitoring student engagement. Their teachers often did not know how to access or interpret monitoring reports, and it was difficult to verify whether students were actively using the platform or spending computer time on other activities (e.g., YouTube).

The provider also suggested that professional learning offerings may have limited staff understanding and engagement. Training on the EdTech platform and data reports were offered only in person and were optional, which the provider believed contributed to low participation and weaker familiarity with the platform and reports.

The gap in communication, support, and district-led expectations described by the school leader and provider conflicts with district leaders' accounts of high levels of OBC collaboration and support offered to schools. However, given that district leaders reported that they respected the decision of schools that opted to be less involved in OBC participation, it may be that schools the district perceived as opting in to full OBC participation received substantially higher levels of communication and support than those that did not.

Finally, one school leader directly observed that it was easier to engage younger students on the platform than older students. In the leader's view, middle school students showed mixed engagement: some participated actively, while others used computer time to navigate to other websites. By contrast, the leader reported that elementary students were generally more consistently engaged, in part because they were less computer literate and therefore less likely to switch to other tasks. The leader also suggested that this pattern was reflected in outcomes, noting that 1st graders were making "the most progress in 15 years."

Similarly, the two high-fidelity schools that were observed during the site visit were both elementary schools and were implementing the intervention with K–1st graders (though the OBC contract was only for 1st graders).

District E

District Overview

District E is a large urban district serving a diverse student population across dozens of elementary, middle, and high schools. During the 2024/25 school year, the district entered into OBC agreements with two literacy tutoring providers across two cohorts (Semesters 1 and 2). In total, approximately 1,000 students received services under OBC contracts in District E during this period. Twenty-seven of the district's 36 elementary schools participated in the OBC intervention. Seventeen schools were served by Intervention 1 and 10 by Intervention 2.

A majority share of treated students made progress toward their targeted outcomes. Approximately 75% of treated students in District E met at least one of their targeted outcomes, though only about 3% met all of the targeted outcomes.

Shared Intervention Design, Payment, and Eligibility Criteria

Both interventions were delivered virtually. The contracts stated that the goal was that each student receive tutoring 3 days a week for 30 minutes a session for 12 weeks in either the fall (Cohort 1) or winter/spring (Cohort 2).

For each student assigned to either provider, the district paid a base payment of $500. The outcomes-based payment structure, shown in Table D36, was identical across both providers and cohorts.

Table D36. Outcomes and Payment Structure by Cohort

	Cohort 1 Metric	Cohort 2 Metric	Payment
Outcome 1: Typical Growth	51–65 CGP on the NWEA MAP from fall to winter	51–65 CGP on the NWEA MAP from winter to spring	$300
Outcome 2: Stretch Growth	66+ CGP on the NWEA MAP from fall to winter	66+ CGP on the NWEA MAP from winter to spring	$250
Outcome 3: Phonics Proficiency	Proficiency on a standardized phonic skills test (end of fall)	Proficiency on a standardized phonic skills test (end of spring)	$125

	Cohort 1 Metric	Cohort 2 Metric	Payment
Outcome 4: Reading Fluency	Fall words-per-minute reading fluency defined by grade (with 95%–100% accuracy)	Winter/spring words-per-minute reading fluency defined by grade (with 95%–100% accuracy)	$167

Note. These outcomes and associated payments were the same for both providers.

Given this payment structure, the maximum amount either provider could have earned in either cohort was $842.

Both interventions used eligibility criteria that were determined in the contracting phase. To be eligible for tutoring, students had to meet all of the following:

- be an English Learner, a student experiencing homelessness (McKinney-Vento), or student in foster care
- be enrolled in 3rd–5th grade
- be below the 40th percentile on the state standardized ELA assessment in spring 2023/24 (the school year preceding the intervention)

District E: Intervention 1

Design, eligibility, and payment structure are shared across both interventions; see the District Overview above for more details.

Descriptive Statistics

The students whose state standardized assessment score performance qualified them for the intervention had (tautologically) lower prior ELA assessment scores than students in the same schools who were not eligible based on prior assessment performance. Also tautologically, given the explicit preference for students in specific student groups, the students were highly likely to be English Learners. Table D37 provides a comparison between the students whose prior assessment scores made them eligible for the intervention and their peers in the same schools for Cohort 1 (fall).

Table D37. Performance and Demographics by Eligibility for Cohort 1

	Ineligible Students *n* = 380	Eligible Students *n* = 680 students	*P*-Value
End-of-Year State ELA Assessment Percentile for Eligibility	53	17	<.001
Students in Foster Care	7 (2%)	8 (1%)	.380
English Learner	374 (98%)	677 (100%)	.053
Treated	3 (1%)	337 (50%)	<.001

Table D38 provides the same comparison for students in Cohort 2.

Table D38. Performance and Demographics by Eligibility for Cohort 2

	Ineligible Students *n* = 378	Eligible Students *n* = 682 students	*P*-Value
End-of-Year State ELA Assessment Percentile for Eligibility	55	17	<.001
Students in Foster Care	9 (2%)	6 (1%)	.047
English Learner	371 (98%)	680 (100%)	.008
Treated	0 (0%)	271 (40%)	<.001

Outcomes Analysis

We use four binary outcomes for this assessment of the tutoring intervention. The first two (for Cohort 1) are derived from growth between fall and winter ELA assessments—one indicating that typical growth has been met and another indicating that stretch growth has been met. The second two (for Cohort 2) are derived from growth between winter and spring

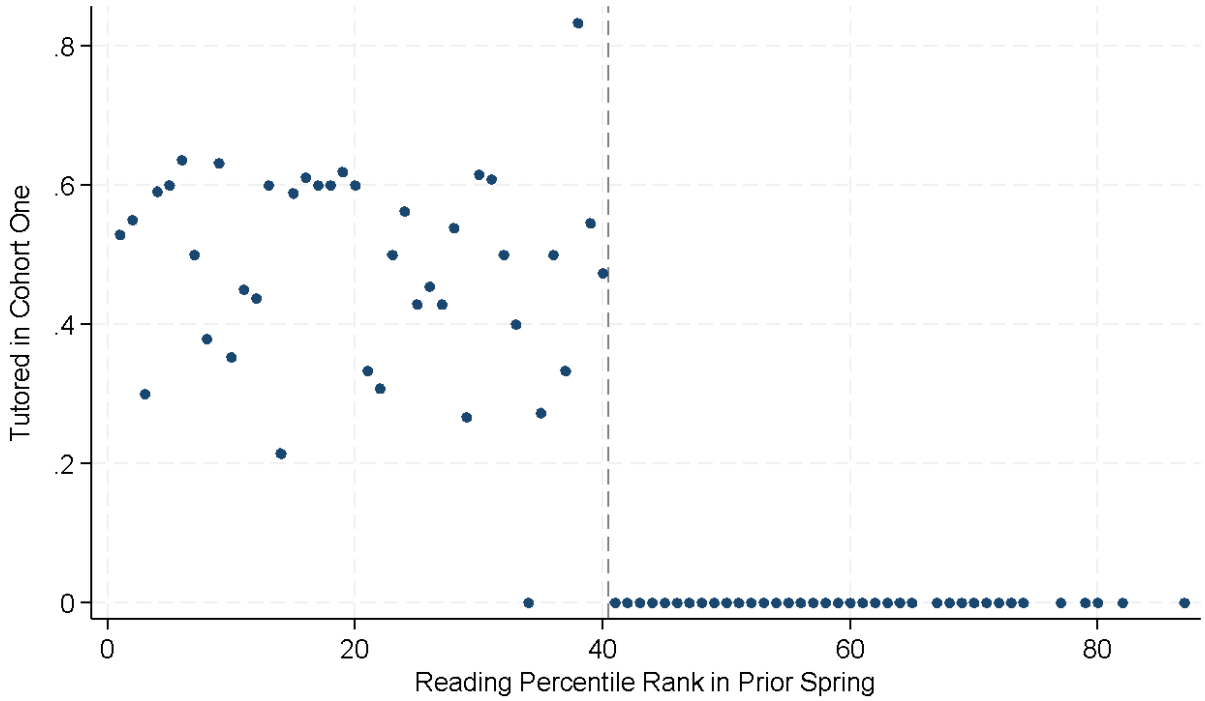
ELA assessments and are similarly coded as indicating that typical or stretch growth has been met. These outcomes are linked directly to the outcomes based contract for tutoring in the district.

To assess the impact of the OBC intervention on student outcomes, we take advantage of the eligibility criterion of lower performance on the prior year's state ELA assessment, which allows the use of a regression discontinuity design. RDD is a research method with strong causal validity that can be used when access or eligibility for a program is clearly defined by a cutoff—in this case, the prior spring's state assessment threshold (40th percentile) that determines intervention eligibility.

Conceptually, students who fall closely on either side of the eligibility threshold are not expected to differ meaningfully, save for access to the tutoring intervention. While an RDD allows researchers to separate cause from correlation without the use of random assignment, the results are well defined only for those students near the performance threshold that determines eligibility for the intervention.

As shown in Figure D2 below, about half of students with lower performance did not receive tutoring. Accordingly, we calculate results both for the impact of being eligible for the intervention and the impact of actually receiving it. When eligibility is the focus, the results yield the impact of the intent to treat on student achievement, often referred to as the "policy effect" because it captures the effect on student outcomes of the district's eligibility policy without regard to school site decisions about compliance with that policy. When focusing on those students who received the treatment, the results yield what is often referred to as the effect on "compliers": those students who comply with their treatment assignment status. For this analysis, that distinction is important. Given that only half of eligible students "complied" with their treatment status, the analysis of the effect on compliers is subject to greater statistical uncertainty than it would have been had 70% or 80% of students complied.

Figure D2. Moderate Compliance With Eligibility Criterion, Cohort 1



Note. A similar pattern is observed for Cohort 2.

Tables D39 and D40 provide the results from the ITT analysis for students in the schools that participated in the intervention with Provider 1. Assignment to the intervention improves the probability of meeting typical growth by 25 percentage points and 31 percentage points for the winter and spring, respectively. Both of these results are distinguishable from zero at or near traditional levels of confidence. Assignment to the intervention appears not to have improved stretch growth in the winter or spring, with effect sizes of 11 and 12 percentage points, respectively, neither of which are statistically distinguishable from zero.

Table D39. ITT Effect of Tutoring for Cohort 1 (Winter)

Outcome	Effect	Standard Error	95% Confidence Interval	*P*-Value	Bandwidth	*N* Below	*N* Above
Stretch Growth	0.11	0.13	(−.15, .37)	.41	11.02	129	93
Typical Growth	0.25	0.15	(−.04, .54)	.09**	11.55	144	104

* $p < .05$. ** $p < .10$

Table D40. ITT Effect of Tutoring for Cohort 2 (Spring)

Outcome	Effect	Standard Error	95% Confidence Interval	*P*-Value	Bandwidth	*N* Below	*N* Above
Stretch Growth	0.12	0.15	(−.16, .41)	.40	12.15	158	148
Typical Growth	0.31	0.15	(.01, .60)	.04*	11.03	148	141

* $p < .05$. ** $p < .10$

Tables D41 and D42 contain the effect of the intervention on compliers in both cohorts—those students who comply with their treatment status. As should be expected, the effect sizes are approximately double the effects in Tables D39 and D40 because the effect on compliers is the effect of eligibility divided by the compliance rate (i.e., effects from Tables D39 and D40 divided by 0.5). Also as expected, statistical uncertainty is higher for the effects in Tables D41 and D42 because of the high degree of noncompliance. The two statistically significant effects from Tables D39 and D40 are now just outside traditional levels of confidence in Tables D41 and D42.

Table D41. Effect of Tutoring on Compliers in Cohort 1 (Winter)

Outcome	Effect	Standard Error	*P*-Value	95% Confidence Interval	Bandwidth	*N* Below	*N* Above
Stretch Growth	0.21	0.26	.42	(−.30, .71)	11.02	129	93
Typical Growth	0.48	0.30	.11	(−.11, 1.1)	11.55	144	104

Table D42. Effect of Tutoring on Compliers in Cohort 2 (Spring)

Outcome	Effect	Standard Error	*P*-Value	95% Confidence Interval	Bandwidth	*N* Below	*N* Above
Stretch Growth	0.23	0.31	.46	(−.39, .84)	12.15	158	148
Typical Growth	0.58	0.35	.10	(−.11, 1.3)	11.03	148	141

Figures D3 and D4 contain the effects in each cohort, both separately and combined, respectively. As can be seen by the blue boxes and confidence intervals, there is suggestive, but statistically uncertain evidence for an effect on typical growth in both the winter and spring outcome tests. The green diamonds represent the overall effect when combining each individual cohort effect. The overall ITT is a 28 percentage point increase in the probability of meeting typical growth. The overall effect of the tutoring is a 52 percentage point increase in the probability of meeting typical growth—a large and statistically significant gain for the students in tutoring who were near the eligibility threshold.

Figure D3. Combined Effect of Being Eligible for Tutoring on Typical Growth

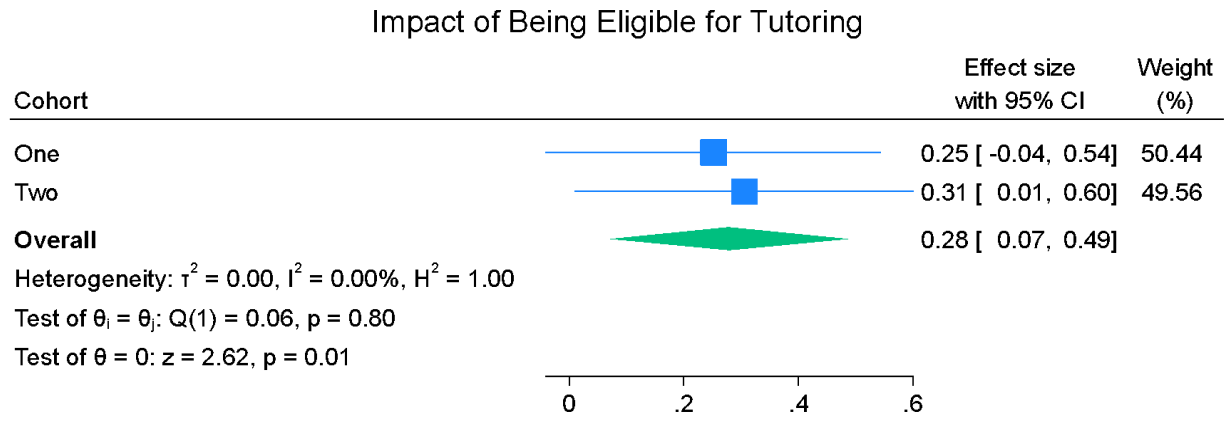
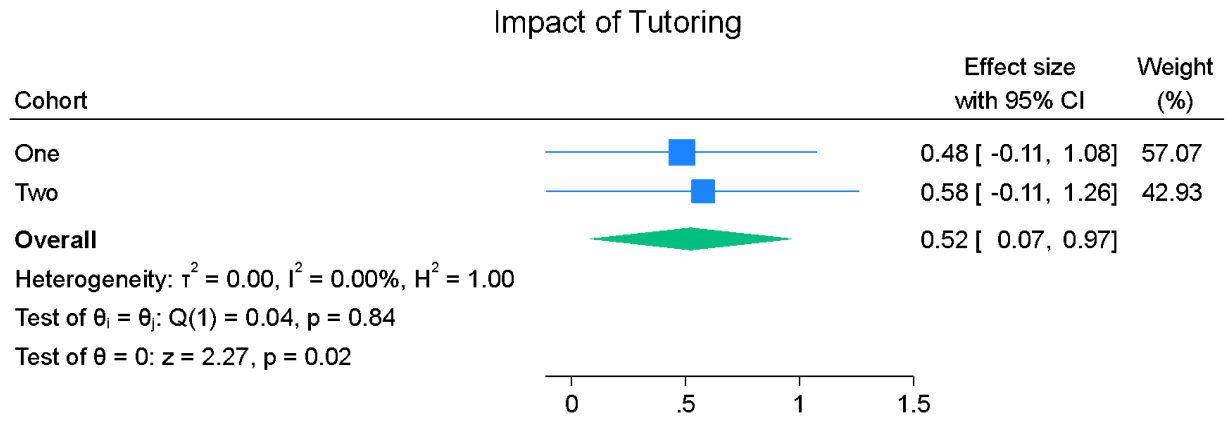


Figure D4. Combined Effect of Tutoring on Typical Growth



Finally, to ensure these results are not driven by the binary nature of the outcomes, we evaluate the impact of eligibility and tutoring on the underlying variable from which the binary growth variable was derived: the conditional growth percentile (CGP). CGP is a measure of how a student's growth compares to the growth of similar students that take the same set of tests. Figures D5 and D6 provide the impacts of CGP by cohort for being eligible and for actually receiving the treatment, respectively.

Figure D5. Combined Effect of Being Eligible for Tutoring on Conditional Growth Percentile

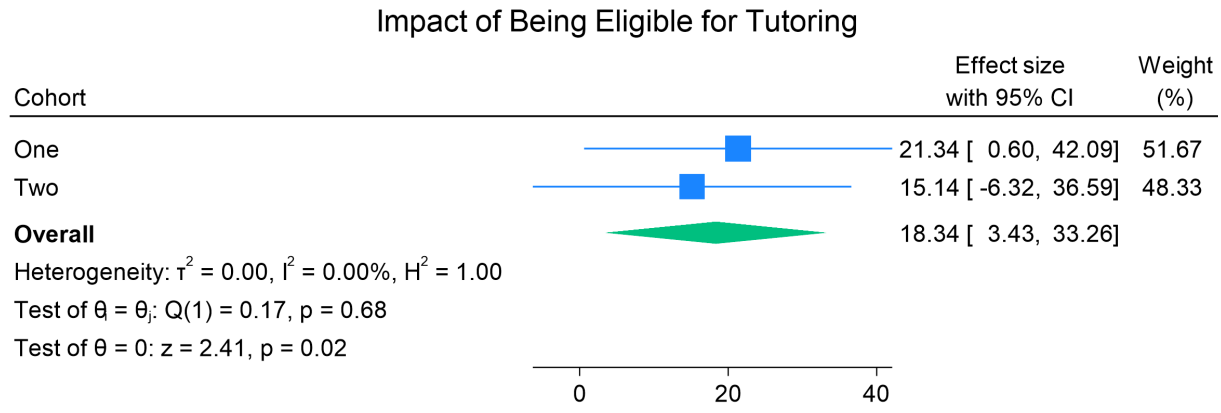
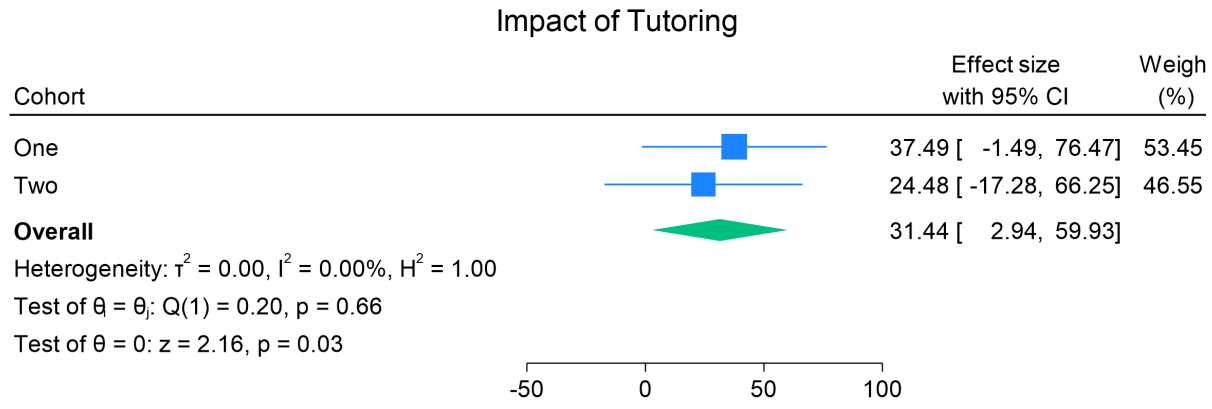


Figure D6. Combined Effect of Tutoring on Conditional Growth Percentile



In these robustness checks, we see the same pattern of effects. The gains from tutoring are larger than the gains associated with merely being eligible (as should be expected when half of the eligible students do not receive the tutoring). The magnitude of the gain—a 31-percentile point increase—is difficult to compare to a standard deviation. However, it is equivalent to moving a student from the 35th percentile (near the boundary for the lower third of the distribution) to the 66th percentile (near the boundary for the upper third of the distribution).

Implications

Intervention 1 had a large, meaningful effect on students' probability of meeting typical growth across two cohorts. Given that the tutoring duration is substantially less than a full academic year, these gains are impressive. The lack of effect on students' probability of

meeting their stretch growth targets may be related to the shorter duration of the tutoring. It should be noted that the effects on stretch growth are all positive and may have been distinguishable from zero with a higher compliance rate. This fact leads naturally to the first implication of these results—the Center for Outcomes Based Contracting (COBC) may want to consider how to strengthen fidelity to the eligibility criteria that districts set for students.

Two findings from interviews with district and school leaders may provide insight into some of the enabling conditions in the district that supported these results. First, leaders described deliberate efforts to build school-level ownership and support for the intervention. District leaders noted that implementation success depended on having dedicated staff at both the school level and the district level whose primary responsibility was to manage day-to-day implementation and maintain strong communication between the district and schools. These structures may have helped sustain compliance with eligibility criteria and address implementation issues as they arose.

Second, district leaders described strengthening progress monitoring in the second year of implementation by more directly involving principals. Principals were asked to attend meetings prepared to discuss an infographic summarizing quarterly attendance, which leaders felt "shifted our conversations to be more focused on implementation and helped principals understand the importance and purpose of OBC and high-impact tutoring." In addition, the provider sent a weekly "consecutive no-show" list so district leaders could follow up with site leaders about students' circumstances and attendance barriers. Together, these practices may have supported stronger implementation fidelity.

District E: Intervention 2

Design, eligibility, and payment structure are shared across both interventions; see the District Overview above for more details.

Descriptive Statistics

The students whose state assessment score performance qualified them for the intervention had (tautologically) lower prior ELA assessment scores than students in the same schools who were not eligible based on their prior spring's assessment performance. Given the explicit preference for students in specific student groups, the students were likely to be English Learners and foster students. Table D43 provides a comparison between the students whose prior assessment scores made them eligible for the intervention and their peers in the same schools for Cohort 1.

Table D43. Performance and Demographics by Eligibility for Cohort 1

	Ineligible Students *n* = 256	Eligible Students *n* = 752 students	*P*-Value
End-of-Year State ELA Assessment Percentile for Eligibility	53	15	<.001
Students in Foster Care	5 (2%)	5 (1%)	.290
English Learner	253 (99%)	748 (100%)	.013
Treated	0 (0%)	170 (23%)	<.001

Note. Socioeconomically disadvantaged status, student race and ethnicity, and student disability status were unavailable for these interventions.

Table D44 provides the same comparison for students in Cohort 2.

Table D44. Performance and Demographics by Eligibility for Cohort 2

	Ineligible Students *n* = 281	Eligible Students *n* = 661 students	*P*-Value
End-of-Year State ELA Assessment Percentile for Eligibility	55	17	<.001
Students in Foster Care	7 (3%)	3 (1%)	.005
English Learner	277 (99%)	658 (100%)	.110
Treated	0 (0%)	195 (30%)	<.001

Note. Socioeconomically disadvantaged status, student race and ethnicity, and student disability status were unavailable for these interventions.

Outcomes Analysis

We use four binary outcomes for this assessment of the tutoring intervention. The first two (for Cohort 1) are derived from growth between fall and winter ELA assessments—one indicating that typical growth has been met and another indicating that stretch growth has been met. The second two (for Cohort 2) are derived from growth between winter and spring ELA assessments and are similarly coded as indicating that typical or stretch growth has been met. These outcomes are linked directly to the outcomes based contract for tutoring in the district.

We had planned to take advantage of the eligibility criterion of lower performance on prior spring assessments, which would typically allow the use of a regression discontinuity design. RDD is a research method with strong causal validity that can be used when access or eligibility for a program is clearly defined by a cutoff—in this case, the prior spring's state assessment threshold (40th percentile) that determines intervention eligibility status.

As shown in Figures D7 and D8 below, however, an overwhelming majority of students with lower performance did not receive tutoring. Accordingly, it is challenging to determine whether differences in the outcomes of eligible and ineligible students are due to tutoring. Doing so is impossible in Cohort 1 and nearly so in Cohort 2. We present the "impact" of crossing the eligibility threshold for the sake of transparency. When focusing on eligibility, the results yield the impact of the ITT on student achievement, often referred to as the "policy effect" because it captures the effect on student outcomes of the district's eligibility policy, without regard to school site decisions about compliance with that policy.

Figure D7. Noncompliance With Eligibility Criterion, Cohort 1

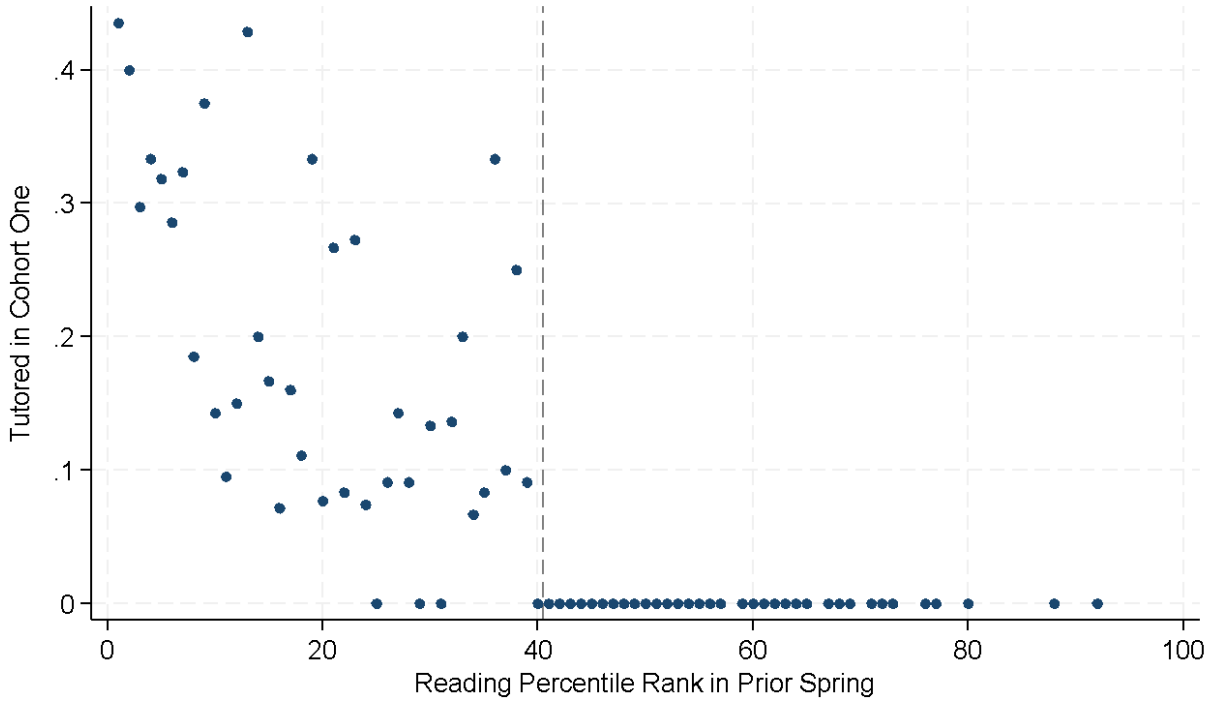
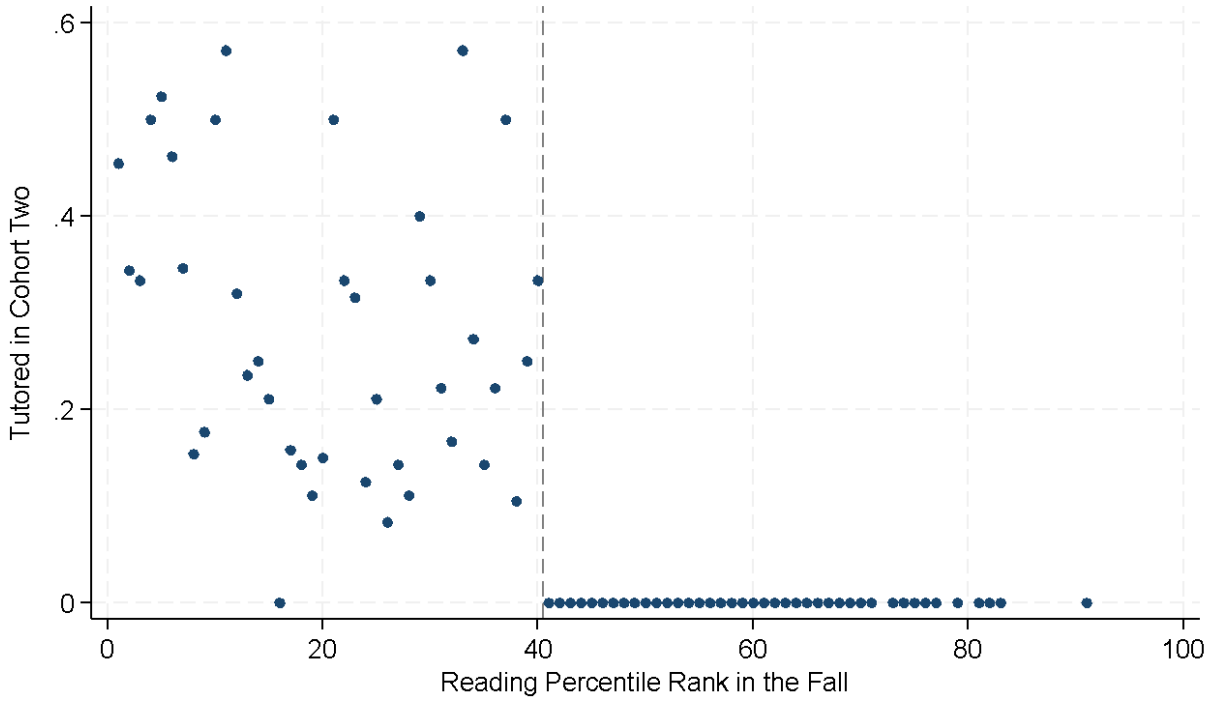


Figure D8. Moderate Noncompliance With Eligibility Criterion, Cohort 2



Tables D45 and D46 provide the results from the ITT analysis for students in the schools that participated in the intervention with Provider 2 in both cohorts. As expected, we were unable to find statistically significant differences in the outcome across the eligibility threshold. Assignment to the intervention appears not to have improved stretch growth or typical growth in the winter or spring, with effect sizes in the range of 13 to 16 percentage points, none of which are statistically distinguishable from zero. This should not be surprising, as only a small fraction of the eligible group of students actually received tutoring in the schools that were using Provider 2.

Table D45. ITT Effect of Tutoring in Cohort 1 (Winter)

Outcome	Effect	Standard Error	*P*-Value	95% Confidence Interval	Bandwidth	*N* Below	*N* Above
Stretch Growth	0.14	0.13	.25	(−.10, .39)	14.30	154	89
Typical Growth	0.13	0.17	.45	(−.21, .47)	13.13	147	83

Table D46. ITT Effect of Tutoring in Cohort 2

Outcome	Effect	Standard Error	*P*-Value	95% Confidence Interval	Bandwidth	*N* Below	*N* Above
Stretch Growth	0.13	0.17	.45	(−.19, .46)	11.20	139	110
Typical Growth	0.16	0.16	.32	(−.16, .48)	9.85	130	102

Implications

In District E, a low percentage of students who were eligible for tutoring with Intervention 2 actually received the treatment. Such noncompliance makes it difficult to conduct a causal analysis of the intervention's impact on student achievement. It also creates misalignment between tutoring implementation and one of the central tenets of the OBC model wherein a defined group of students is identified for support based on empirical achievement data.

Nevertheless, it should be noted that the effects on stretch and typical growth are all positive and may have been distinguishable from zero with a higher compliance rate. This fact leads naturally to the first implication of these results—the COBC may want to consider strengthening support for districts to ensure greater fidelity to the eligibility criteria that they set for students.

Two themes from interviews with the provider and district and school leaders may help explain these findings. First, leaders linked attendance and engagement challenges to scheduling and delivery format. One leader reported that holding sessions after school reduced attendance because some parents viewed tutoring as "extracurricular" and therefore as a lower priority; as a result, students were sometimes picked up early or did not attend. The provider similarly connected after-school scheduling to enrollment challenges. A district leader also suggested that the virtual format may have contributed to weaker attendance and engagement, noting a "big issue in terms of getting buy-in, even from school admin" due to a perception or misconception about virtual tutoring. Together, these factors may have hindered consistent participation.

Second, the provider reported that students were swapped out mid-program, which may have affected outcomes. The provider noted that the OBC offered a fairly long window (a full month) during which to finalize tutoring participants. They reported that this led to switching out students (e.g., due to behavioral issues) later in the program than they normally would have anticipated. The provider noted that this may have compromised outcomes, as the new students had less time in the program to achieve intended outcomes. In the provider's view, the extended window for finalizing tutoring groups also increased the number of roster changes, creating greater implementation volatility and making it harder to build strong, stable matches between students and tutors.

District F

District Overview

District F is a suburban district serving a highly diverse student population. During the 2023/24 and 2024/25 school years, District F entered into an OBC agreement with a virtual tutoring provider to address math and ELA gaps among elementary school–aged students. Approximately 70% of treated students met at least one of their targeted outcomes in the tutored subject; only about 4% met all of the targeted outcomes in the tutored subject.

Intervention Overview

During the 2023/24 and 2024/25 school years, District F had an outcomes based contract with a virtual tutoring provider over 2 school years. In both school years, the provider supplied tutoring in both reading and math. Across both years and for both subject areas, the eligibility criteria remained the same. Students in grades 3–6 who scored less than the 40th percentile on the diagnostic test administered at the start of the given school year were eligible to participate in the intervention. The intervention was designed to have services delivered twice a week for 45 minutes over a 24-week period in 2023/24 and a 20-week period in 2024/25 during after-school programming. In some cases, tutoring was provided during the school day depending on teacher approval.

For each student assigned to tutoring, the district paid a base payment of $648 and additional outcomes-based payments were structured based on how much of the yearly growth goal students achieved at the end of each diagnostic window (Diagnostics 2, 3, and 4 for 2023/24 and Diagnostics 2 and 3 for 2024/25). The additional outcomes-based payments were as follows:

- **$100** if a student achieved their typical goal on each diagnostic test
- **$85** additional if a student achieved their stretch goal on each diagnostic test
- **$127** if a student showed an 80% positive response on a student efficacy as a learner screener

The typical growth and stretch growth payments stack: For a student who met their stretch growth goal in a given diagnostic window, the provider earned both the $100 typical growth payment and the additional $85 stretch payment for that window, summed across all diagnostic windows in the year. The $127 efficacy-screener payment could be earned once per year. The per-student maximum payment, including the base payment, therefore was $1,330 (2023/24) and $1,145 (2024/25).

Eligibility Criteria

As noted above, eligibility for this tutoring intervention was based on a student's beginning-of-year diagnostic test performance. Rising 3rd–6th graders must have scored below the 40th percentile on a given subject's diagnostic test to be eligible for the tutoring intervention. All schools in the district (10 schools) participated in the tutoring intervention under an OBC.

Table D47 provides a comparison of eligibility and treatment status across all four intervention cells. As expected given the eligibility criteria, the number of students who qualified for tutoring substantially exceeded the number who received it in all four cells. Compliance rates ranged from 21% to 33% across years and subjects, reflecting the

district's characterization of the intervention as largely opt-in due to its out-of-school delivery format. A small but notable share of students—between 11% and 17% of treated students in each cell—received tutoring despite scoring above the eligibility threshold, indicating that school- and student-level discretion played a role in assignment, alongside the formal eligibility criteria.

Table D47. Tutoring Assignment and Compliance by Subject and Year

	2023/24 Math Tutoring	2023/24 ELA Tutoring	2024/25 Math Tutoring	2024/25 ELA Tutoring
Eligible for Tutoring	875	821	665	743
Received Tutoring	186	197	218	185
Received Tutoring Without Meeting Eligibility Criteria	25	26	38	22

Descriptive Statistics

As expected given the eligibility criteria, the students whose diagnostic test performance qualified them for the intervention had lower beginning-of-year math and reading diagnostic scores than students in the same schools who were not eligible based on their diagnostic test performance. Across the 2 years and for both tutored subjects, the eligible students were also more likely to be English Learners, have a disability, be Black, or be Hispanic.

Table D48 provides a comparison of the students whose reading diagnostic test scores made them eligible for the intervention and their peers in the same schools for school year 2023/24.

Table D48. Performance and Demographics by Eligibility for the 2023/24 Reading Intervention

	Ineligible Students *n* = 2,296	Eligible Students *n* = 844 students	*P*-Value
Reading Diagnostic 1: Mean Percentile	79	18	<0.001
English Learner	360 (16%)	448 (53%)	<0.001
Students With Disabilities	97 (4%)	216 (26%)	<0.001
Black/African American	22 (1%)	24 (3%)	<0.001
Hispanic	242 (11%)	354 (42%)	<0.001
Treated	23 (1%)	174 (21%)	<0.001

Table D49 provides a comparison of the students whose math diagnostic test scores made them eligible for the intervention and their peers in the same schools for school year 2023/24.

Table D49. Performance and Demographics by Eligibility for the 2023/24 Math Intervention

	Ineligible Students *n* = 2,235	Eligible Students *n* = 906 students	*P*-Value
Math Diagnostic 1: Mean Percentile	80	19	<0.001
English Learner	386 (17%)	425 (47%)	<0.001
Students With Disabilities	94 (4%)	219 (24%)	<0.001

	Ineligible Students *n* = 2,235	Eligible Students *n* = 906 students	*P*-Value
Black/African American	17 (1%)	29 (3%)	<0.001
Hispanic	195 (9%)	402 (44%)	<0.001
Treated	22 (1%)	164 (18%)	<0.001

Table D50 provides a comparison between the students whose reading diagnostic test scores made them eligible for the intervention and their peers in the same schools for school year 2024/25.

Table D50. Performance and Demographics by Eligibility for the 2024/25 Reading Intervention

	Ineligible Students *n* = 2,365	Eligible Students *n* = 774 students	*P*-Value
Reading Diagnostic 1: Mean Percentile	78	19	<0.001
English Learner	363 (15%)	422 (55%)	<0.001
Students With Disabilities	81 (3%)	225 (29%)	<0.001
Black/African American	25 (1%)	19 (2.5%)	0.004
Hispanic	264 (11%)	337 (44%)	<0.001
Treated	17 (1%)	168 (22%)	<0.001

Table D51 provides a comparison of the students whose math diagnostic test scores made them eligible for the intervention and their peers in the same schools for school year 2024/25.

Table D51. Performance and Demographics by Eligibility for the 2024/25 Math Intervention

	Ineligible Students *n* = 2,468	Eligible Students *n* = 673 students	*P*-Value
Math Diagnostic 1: Mean Percentile	82	20	<0.001
English Learner	448 (18%)	338 (50%)	<0.001
Students With Disabilities	92 (4%)	214 (32%)	<0.001
Black/African American	20 (1%)	24 (4%)	<0.001
Hispanic	279 (11%)	322 (48%)	<0.001
Treated	37 (2%)	181 (27%)	<0.001

Note. Socioeconomically disadvantaged status was unavailable for these interventions.

Outcomes Analysis

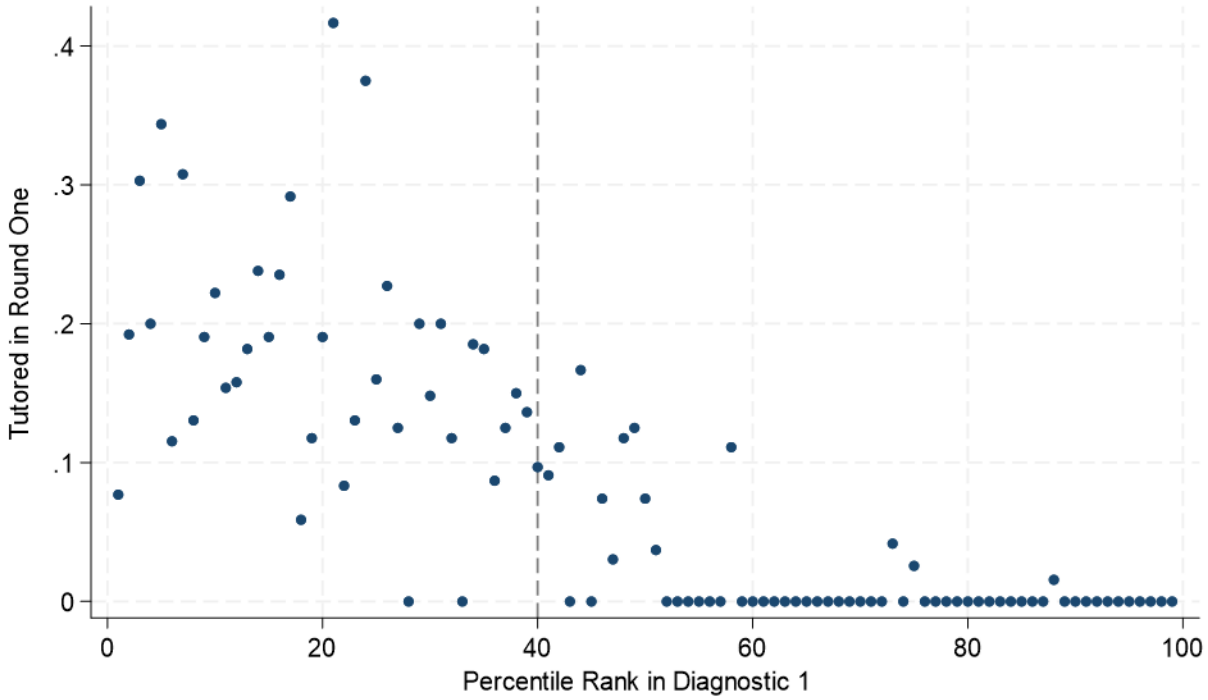
We use 12 binary outcomes for this assessment of the tutoring intervention: two indicators for each of three diagnostic assessments in both math and ELA. The diagnostic assessments are given in the winter, the spring, and the end of the school year, separately for ELA and math. For each of these six tests, we create two binary indicators: one for meeting "typical growth" on the diagnostic test and another for meeting "stretch growth" on the same test. These outcomes are linked directly to the outcomes based contract for tutoring in the district. We do not analyze the student-efficacy-as-a-learner screener that triggered the $127 payment; the outcomes reported here are limited to the typical- and stretch-growth indicators tied to the $100 and $85 payments.

To assess the impact of the OBC intervention on student outcomes, we compared students who were eligible and received tutoring to students who were eligible and did not receive tutoring. Such a comparison cannot be considered causal because we cannot account for all of the unobservable ways in which the treated students are different from the untreated students. We had planned to take advantage of the eligibility criterion of lower performance on fall diagnostic tests, which would typically allow the use of a regression discontinuity design. RDD is a research method with strong causal validity that can be used when access

or eligibility for a program is clearly defined by a cutoff—in this case, the fall diagnostic score threshold (40th percentile) that determines intervention eligibility status by subject (math and ELA).

As shown in Figure D9 below, the test score eligibility criteria did little to determine treatment status near the 40th percentile cutoff value. Some students with higher performance received tutoring, and the majority of students with lower performance did not receive tutoring.

Figure D9. Noncompliance With Tutoring Eligibility for the Math 2023/24 Intervention



The magnitude of noncompliance found here complicates causal comparisons for a number of reasons. First, it removed the possibility of a rigorous RDD because assignment to treatment is continuous around the threshold; there is not a sharp increase in the probability of receiving tutoring when scores dip below the 40th percentile. Second, the fact that a mere fifth of students who were eligible actually received tutoring indicates there was considerable unobserved selection into the treatment. Undoubtedly, the fact that tutoring was offered after school played an important role in determining who was treated, but we lack data on which students attended the after-school sessions. Instead, we compare outcomes for lower scoring students, a fifth of whom enrolled in tutoring due to a constellation of forces that we cannot measure.

With all of those caveats in mind, we run a comparison between students who received tutoring and those who did not, all of whom were eligible based on their diagnostic score, which we control for in the following simple model:

$$Outcome_i = \alpha + \beta_1 Treated_i + \beta_2 Diagnostic1 + e_i$$

In this model, Outcome is the outcome for the *i*th student, Treated is a binary indicator for participating in tutoring, Diagnostic1 is the score on the *i*th student's first diagnostic exam in the relevant subject (the one that assigns them to tutoring), and *e* is a random error term. In the tables below, we report 1.

Tables D52 through D55 provide the results from this descriptive analysis. For 8 of the 12 outcome measures in the 2023/24 school year, the differences between the tutoring group and the untreated group are not statistically distinguishable from zero. In three of the remaining four, the tutoring group was between 7 and 9 percentage points more likely to meet the particular outcomes—differences that can be distinguished from zero at or near conventional levels. In the final comparison, the tutoring group was 6 percentage points less likely to meet their stretch goal in Diagnostic 2. Of the eight outcome measures in the 2024/25 school year, four are distinguishable from zero at conventional levels, indicating that the tutoring group was between 6 and 10 percentage points less likely to meet the relevant outcomes. We stress again that these should not be interpreted as causal.

Table D52. "Effect" of Reading Tutoring in District F in 2023/24

Outcome	Diagnostic Wave	Difference	Standard Error	*P*-Value	95% Confidence Interval	R2
Stretch	4	−0.014	0.042	0.734	(−.10, .07)	0.000
Stretch	3	0.019	0.039	0.627	(−.06, .10)	0.000
Stretch	2	−0.029	0.030	0.348	(−.09, .03)	0.001
Typical	4	0.029	0.039	0.456	(−.05, .10)	0.001
Typical	3	0.075	0.041	0.065	(−.01, .15)	0.004
Typical	2	0.024	0.042	0.570	(−.06, .11)	0.001

Note. The *n*-size of eligible students for the 2023/24 reading tutoring cohort was 844.

Table D53. "Effect" of Math Tutoring in District F in 2023/24

Outcome	Diagnostic Wave	Difference	Standard Error	*P*-Value	95% Confidence Interval	R2
Stretch	4	0.088	0.043	0.041	(.00, .17)	0.006
Stretch	3	−0.025	0.033	0.455	(−.09, .04)	0.005
Stretch	2	−0.059	0.022	0.008	(−.10, −.02)	0.009
Typical	4	0.089	0.039	0.021	(.01, .17)	0.006
Typical	3	0.000	0.043	0.998	(−.08, .08)	0.006
Typical	2	0.009	0.040	0.815	(−.07, .09)	0.020

Note. The *n*-size of eligible students for the 2023/24 math tutoring cohort was 906.

Table D54. "Effect" of Reading Tutoring in District F in 2024/25

Outcome	Diagnostic Wave	Difference	Standard Error	*P*-Value	95% Confidence Interval	R2
Stretch	3	−0.055	0.037	0.139	(−.13, .02)	0.007
Stretch	2	−0.060	0.030	0.042	(−.12, .00)	0.006
Typical	3	0.037	0.041	0.376	(−.04, .12)	0.003
Typical	2	−0.094	0.042	0.026	(−.18, −.01)	0.013

Note. The *n*-size of eligible students for the 2024/25 reading tutoring cohort was 774.

Table D55. "Effect" of Math Tutoring in District F in 2024/25

Outcome	Diagnostic Wave	Difference	Standard Error	*P*-Value	95% Confidence Interval	R2
Stretch	3	−0.076	0.029	0.010	(−.13, −.02)	0.048
Stretch	2	−0.059	0.023	0.010	(−.10, −.01)	0.040
Typical	3	−0.016	0.043	0.703	(−.10, .07)	0.031
Typical	2	−0.057	0.040	0.158	(−.14, .02)	0.042

Note. The *n*-size of eligible students for the 2024/25 math tutoring cohort was 673.

Implications

A potential implication from District F is that additional attention may be merited in how districts set eligibility and implementation rules when they define the target student population for an intervention under an outcomes based contract. In District F, assignment to treatment was only loosely related to the eligibility criteria set forth in the contract, as only one fifth of eligible students received the intervention, and some who were not eligible received it. Such noncompliance with eligibility criteria does not allow us to conduct a causal analysis of the intervention's impact on student achievement. It also creates misalignment between tutoring implementation and the central tenets of the OBC model wherein a defined population of students is identified and prioritized for support based on empirical achievement data.

Qualitative data gathered from interviews with district and school leaders suggest that flexibility in the timing of the intervention may have contributed to noncompliance. School leaders reported that the intervention was offered after school but that teachers could also choose to schedule the intervention during small-group instruction time during the school day. One school leader noted that some teachers were reluctant to use small-group time for the intervention because they were concerned it would disrupt regular instruction; in these cases, students only participated after school.

Teacher discretion over whether to offer the intervention during the school day may have important implications for implementation fidelity. One leader perceived meaningful differences in fidelity across these settings. Attendance and implementation fidelity appeared stronger when the intervention was delivered during the school day, whereas

after-school sessions were associated with lower student engagement and more variable attendance.

Although the sample sizes were too small to separate effects by grade level, qualitative data also suggest that students in different grade levels may have responded differently to the virtual tutoring intervention. One school leader suggested that younger students benefited the most from the tutors' individual attention and differentiated support, as suggested by these students' visible enthusiasm and growth in achievement that the district observed over time. In contrast, the same leader reported that sustaining engagement among older students was more difficult. In the leader's view, pulling older students out of regular class time for the intervention was disruptive; students appeared more distracted and fatigued, and teachers expressed concern that students missed important classroom content.

A school leader also reported that the online format posed challenges for students with individualized education programs (IEPs). Teachers told the leader they spent substantial time helping these students log on and remain engaged for the required duration, which raised concerns that these barriers may have reduced the intervention's effectiveness for those students.

Interview data also suggest that the district experienced some initial implementation challenges that improved over time. For example, school leaders reported that teachers were initially skeptical about the intervention's effectiveness due to its online format and uncertainty about tutors' qualifications. Provider-led trainings, combined with increased experience using the intervention, helped address these concerns over time. As a result, teachers more consistently reviewed the platform's progress reports and showed stronger buy-in. School leaders said this shift was reflected in teachers' greater willingness to integrate the intervention into small-group instruction during the school day and, in some cases, to use progress-report data to inform instruction. Such efforts may have improved implementation fidelity and student achievement over time.

School leaders also shared that student attendance and overall engagement were not very strong at the start of implementation. To help address this issue, the district and provider suggested that implementation leaders provide incentives for students. Such incentives included "school currency" for students to use on treats or a pizza party for the entire class, which were awarded for meeting attendance and achievement benchmarks. School leaders perceived that attendance improved because of this change and that students' interest and confidence in their performance grew as a result, too. The provider shared the sentiment that these efforts played an important role in strengthening student attendance.

Finally, in the district leader's view, implementing an intervention under OBC opened up communication pathways between the district, provider, school staff, and families at multiple stages throughout the process, and these communication pathways strengthened implementation over time. For example, the leader shared that planning meetings to

develop the contract brought together principals, teachers, after-school staff, and the provider. The leader indicated that lessons learned from this process informed strategies to better involve school staff in the second year of implementation and to make further efforts to ensure their involvement in contract development and implementation planning.

The district leader also reported that they designated teachers on special assignment to facilitate communication between teachers, tutors, and families about the intervention and to provide structured progress updates. Bringing these interest holders together increased buy-in from families and teachers and supported ongoing dialogue about implementation between classroom staff and tutors.

Additional Graphs

Figure D10. Noncompliance With Tutoring Eligibility for the Reading 2023/24 Intervention

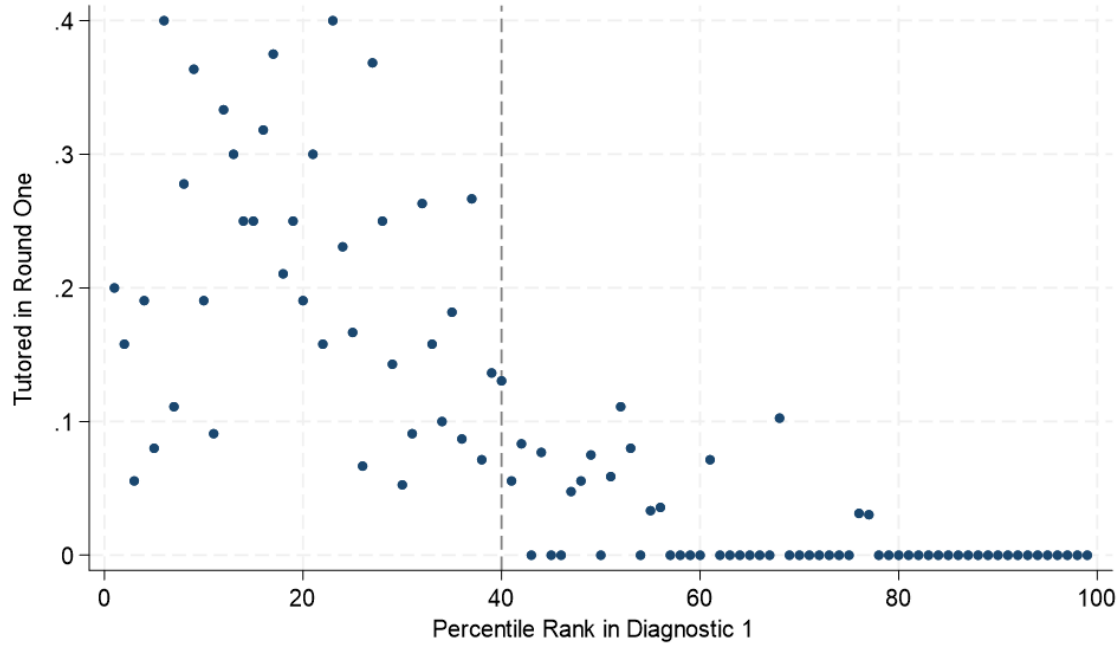


Figure D11. Noncompliance With Tutoring Eligibility for the Math 2024/25 Intervention

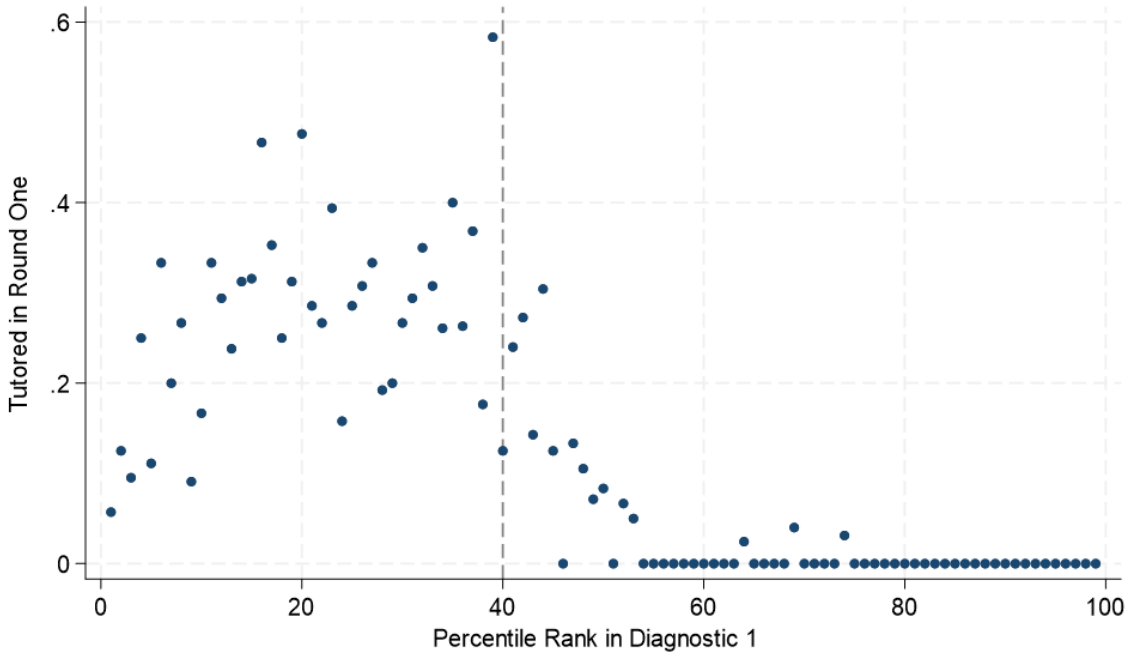
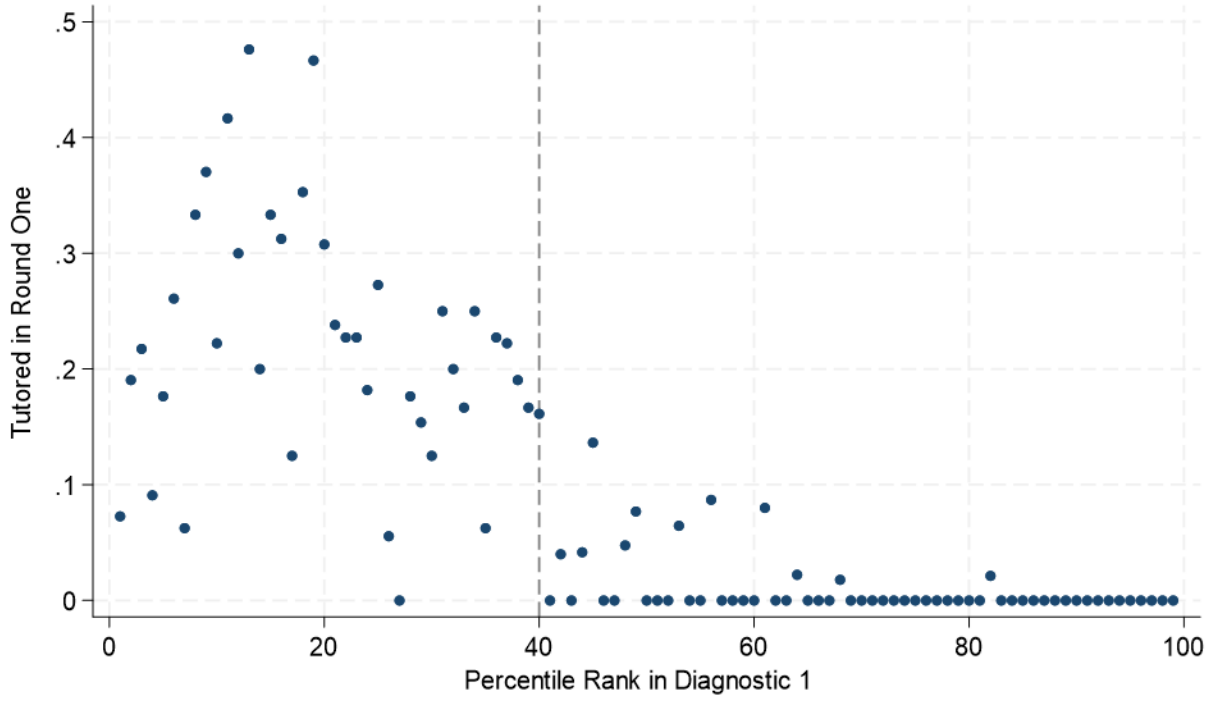
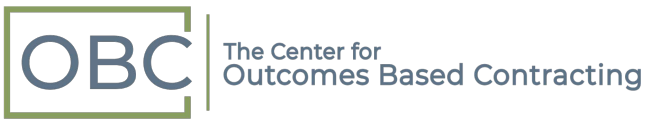


Figure D12. Noncompliance With Tutoring Eligibility for the Reading 2024/25 Intervention




WestEd


OBC
The Center for
Outcomes Based Contracting