Paloma and Early Literacy

A Quasi-Experimental Impact Analysis of K-2 Reading

An evaluation of beginning-of-year to end-of-year reading outcomes for kindergarten through second-grade students in a large urban school district, 2025-26. Paloma is a low-cost, home-based program that equips parents to deliver classroom-aligned reading practice. Effects are estimated against a propensity-matched within-district comparison group on DIBELS 8th Edition assessments, across a graded measure of platform exposure.

Paloma Learning · Research & Impact

SY25-26 Impact White Paper

Table of Contents

Table of Contents **2**
Executive Summary **4**
1. Introduction **5**
1.1 Why early reading matters 5
1.2 Why engagement is the crux 5
1.3 Research questions 5
2. How Paloma Works: A Theory of Change **6**
2.1 What families do each week 6
2.2 Why this should work 6
2.3 What this implies for measuring impact 6
3. Who Paloma Serves **6**
3.1 Setting and sample 6
3.2 A high-need population 7
3.3 Measures 7
3.4 Levels of analysis: nested exposure tiers 7
4. Constructing the Comparison **8**
4.1 Sample construction 8
4.2 Attrition by arm 8
4.3 Matching design 9
4.4 Baseline equivalence 9
4.5 Campus selection and confounding 9
5. Who Engages – and Why It Matters for the Results **10**
5.1 Engagement overview 10
5.2 The selection model 10
5.3 The load-bearing null 10
5.4 What predicts use: attendance and family routine 11
5.5 Dose-response: engagement and outcomes 11
5.6 A seat without use does nothing: the falsification test 12
6. Does Paloma Work? **12**
6.1 Estimation 12
6.2 Benchmark attainment: the primary result 12
6.3 Reading score: the magnitude of improvement 13
6.4 Growth-goal attainment 13
6.5 Effect sizes in context 13
6.6 Robustness 14
6.7 Specification robustness 14
7. For Whom It Works Best **15**
7.1 The below-benchmark population 15
7.2 Catch-up to grade level 15

7.3 The catch-up is real gain..... 16
7.4 No gap-widening..... 16
7.5 The equity arc, synthesized..... 16
8. Consistency with Prior Evidence..... 17
8.1 Two designs, two kinds of validity..... 17
8.2 The comparison must be made at the same dose..... 17
8.3 The effect scales with the exposure window..... 17
8.4 The specific study answers the general one's open questions..... 18
8.5 The grade-level nulls do not replicate..... 18
9. What This Means for the District..... 19
10. Cost-Effectiveness..... 20
11. Evidence Standards and Limitations..... 21
12. Conclusion..... 22
References..... 23

Executive Summary

Paloma is a low-cost, home-based early-literacy program. Each week it equips parents to deliver short, app-led reading lessons that mirror the exact scope and sequence of the child's classroom – reinforcing, at home, the specific skill the child is learning in school that week. At $50 per student per year, it is an order of magnitude less expensive than the threshold education researchers treat as "low cost." This paper estimates its effect on kindergarten through second-grade (K-2) reading outcomes in a large urban school district during the 2025-26 school year, following students from the beginning of the year (BOY) to the end of the year (EOY) on the DIBELS 8th Edition assessment against a propensity-matched within-district comparison group.

The central finding is that offering Paloma raised grade-level reading attainment, and that the effect grew with engagement. On the pre-specified primary outcome – meeting the DIBELS grade-level benchmark – the intention-to-treat effect was positive and statistically significant (adjusted difference = +3.3 pp, g = +0.07, p = .004). Among students who used the platform the effect was +7.0 pp (g = +0.15, p < .001), and among those who used it weekly – the program's intended dose – the effect was +10.3 pp (g = +0.23, p < .001).

Three features make the result credible. The comparison groups were equivalent at baseline on the matched intention-to-treat sample (standardized pretest difference g = −0.027). A falsification test on students who were offered the platform but completed no lessons returned a null effect on every outcome, indicating the result is not an artifact of which campuses adopted Paloma. And engagement was not sorted by students' starting reading level (OR = 1.04 per SD, p = .44), so the as-implemented effect is not a reflection of stronger readers selecting into use. The study is designed to meet the evidence requirements of ESSA 'Moderate Evidence' (Tier II); §11 maps the design against those standards.

The benefit was concentrated where it matters most. Among the 68 percent of students who began the year below grade-level benchmark, the share reaching grade level by spring exceeded matched comparison students at every level of exposure – 49% vs 44% (offered), 58% vs 47% (used), and 63% vs 46% (weekly). Students who began at or above benchmark, already near the ceiling, maintained their standing. Paloma therefore narrows the achievement gap rather than widening it. The pattern is not unique to this year: it reproduces a broader, multi-school study from the prior school year, two independent samples reaching the same conclusion by different routes.

Paloma users at these high-need campuses started 12 percentage points behind the district average on benchmark attainment. By the end of the year, they matched it. Set against the program's cost, the result is striking. At the intended weekly dose, Paloma's effect is comparable to the average effect of tutoring in the experimental literature (Nickow, Oreopoulos, & Quan, 2024) – a tutoring-grade result delivered through a parent-led model the evidence expects to be the weakest delivery channel, at $50 per student per year: roughly one two-hundredth of the annual per-student cost of formal schooling, and a fraction of what tutoring of similar effect costs.

1. Introduction

1.1 Why early reading matters

Reading on grade level in the early elementary years is among the most durable predictors of long-run academic success. The transition that occurs across kindergarten through second grade – from decoding words to reading fluently for meaning – sets a trajectory that is difficult to alter later. By the end of third grade, instruction increasingly assumes that students can read to learn; students who cannot are left to absorb new content through a skill they have not yet mastered, and the gap compounds.

The long-term stakes are well documented. Children who are not reading proficiently by the end of third grade are roughly four times less likely to graduate from high school on time than proficient readers, and the disparity is substantially larger for students from low-income families (Hernandez, 2011). Early reading is therefore not one academic outcome among many; it is a gateway whose effects propagate through a student's entire educational career. Interventions that move early reading – particularly for students who begin behind – are correspondingly high-leverage.

1.2 Why engagement is the crux

Education technology is an uneven lever on early literacy. Syntheses of experimental evidence find that the effects of digital learning programs vary widely and are frequently small, and that computer-assisted and "personalized learning" approaches deliver primarily to the extent that students actually engage with them (Escueta, Nickow, Oreopoulos, & Quan, 2020). A program that works well for engaged students can show a negligible average effect if a large share of students never use it.

Two consequences follow for any honest evaluation. First, the effect of providing access and the effect of engagement are distinct quantities, and conflating them either understates the program (by averaging in non-users) or overstates it (by reporting only the engaged). A credible analysis must report both. Second, because engagement is voluntary, the students who use a platform may differ systematically from those who do not, so an estimate based on users alone risks attributing to the program what is really selection. The design of this study is built around both concerns.

1.3 Research questions

The analysis is organized as a funnel that follows students from access through outcomes, structured around four questions: Reach – who does Paloma serve, and is the comparison fair? Use – who engages, what predicts engagement, and what does it produce? Impact – does Paloma improve reading outcomes, and by how much? Equity – does the benefit reach the students who began furthest behind? A cost-effectiveness analysis follows, positioning the measured effects against what the program costs.

2. How Paloma Works: A Theory of Change

2.1 What families do each week

Paloma is used at home, not in school. It is a parent-empowerment program: it equips parents to deliver structured reading practice to their child outside of class. Each week, the platform assigns a set of five short, app-led lessons, each keyed to a specific skill in that week's classroom sequence. The lessons mirror the exact scope and sequence of the child's classroom – not a parallel curriculum, but the same skills, in the same order, that the teacher is introducing that week. The child works through the lessons on the app while the parent acts as coach and supervisor, guiding and encouraging rather than teaching from scratch.

The intended pattern of use is weekly: a family using Paloma as designed engages with that week's aligned lessons each week, keeping home practice synchronized with classroom instruction. This weekly cadence is the dose the program is built around.

2.2 Why this should work

The mechanism is aligned reinforcement. A child receives additional, structured practice on precisely the skills the teacher is introducing, in the same sequence, at home – adding time on task on grade-level content, distributing that practice across the week, and bringing a parent into the learning. This design targets the two conditions the evidence says matter most for educational technology: engagement (Escueta et al., 2020) and alignment (Kraft, 2020).

2.3 What this implies for measuring impact

Because the program is home-based and parent-mediated, whether a child benefits depends on whether the family engages. That is why the analysis treats exposure as a graded condition rather than a binary one, and why it devotes a full section to who engages and why. The home-based design also clarifies what the comparison must rule out: not a difference in classroom instruction, but whether families who choose to engage differ from those who do not in ways that would inflate a usage-based estimate.

3. Who Paloma Serves

3.1 Setting and sample

The study covers the kindergarten, first-, and second-grade students at 14 campuses of one large urban school district that adopted Paloma during the 2025-26 school year. The platform was made available to families at these campuses by the district rather than selected by the vendor, and the comparison pool was drawn from 161 other campuses within the same district. Conducting the comparison entirely within one district holds constant the curriculum, calendar, assessment administration, and policy environment.

3.2 A high-need population

The campuses where Paloma was deployed serve a substantially higher-need population than the district average. Ninety-six percent of K-2 students at treatment campuses are economically disadvantaged, compared to 74 percent district-wide. Eighty-seven percent are classified as at-risk versus 73 percent. Sixty-one percent are emergent bilingual, compared to 46 percent across the district.

The starting-point gap is substantial. Students at treatment campuses began the year with a mean DIBELS composite score 13 points below the district average, and 53 percent started well below the grade-level benchmark compared to 40 percent district-wide. Only 13 percent began above benchmark, versus 25 percent in the district. These are not campuses where a positive result comes easily – they are campuses where the need is greatest and the baseline is lowest.

3.3 Measures

Reading was assessed with the DIBELS 8th Edition composite score, administered at the beginning and end of the school year. Three outcomes are carried through the analysis: benchmark attainment (primary), continuous composite score (secondary), and growth-goal attainment (supporting).

3.4 Levels of analysis: nested exposure tiers

Because the program can only affect children whose families use it, exposure is treated as a graded condition. Three nested levels are reported throughout the paper, all drawn from a single matched comparison with paired controls.

Exposure	Definition	Interpretation
Offered	All students whose families were given access to Paloma, regardless of usage (intention-to-treat).	**Causal floor**
Used	Completed at least one lesson.	As-implemented
Used weekly	Engaged on at least a weekly cadence – the intended dose.	Correlational

Table 1. Exposure levels.

4. Constructing the Comparison

4.1 Sample construction

The analytic sample was constructed through a series of exclusions, each named and counted to ensure full transparency.

Step	N
BOY-tested students (baseline)	34,952
− unresolvable duplicate BOY records	−28
= baseline-eligible	**34,924**
− lost to follow-up (no EOY record)	−3,235
− EOY present but unresolvable duplicate	−1
= BOY+EOY complete	**31,688**
− missing demographics	−112
− treatment campus, no usage record	−20
− missing growth classification	−22
= Analytic sample	**31,534**

Table 2. Sample construction waterfall.

Overall outcome attrition from baseline-eligible to BOY+EOY complete was 9.27 percent.

4.2 Attrition by arm

Sample loss between the beginning- and end-of-year assessments was low and favored retention in the treatment group. On the as-assigned frame – students with a baseline assessment and an arm assignment (33,724 students: 2,386 treatment, 31,338 comparison) – 2.18 percent of treatment students and 6.75 percent of comparison students were missing the end-of-year primary outcome, for overall attrition of 6.43 percent and differential attrition of 4.57 percentage points. Under the What Works Clearinghouse attrition standard, this combination constitutes low attrition under both the optimistic boundary (10.8 percentage points at this overall rate) and the more stringent cautious boundary (6.3 percentage points) – the study does not need the more favorable assumption to clear the bar. The as-assigned frame is narrower than the Table 2 waterfall, which retains students without demographic records and is the source of the 9.27 percent figure reported there; the two describe the same sample flow at different stages. The treatment group's lower attrition partly reflects how the offer is defined: a student enters the treatment frame at onboarding, which screens out families who left the district early in the year, while the comparison frame carries ordinary district mobility across the full year. Because this is a quasi-experimental design, the binding WWC requirement is baseline equivalence of the analytic sample (§4.4); the attrition accounting documents the sample flow and confirms that differential loss does not threaten the match.

4.3 Matching design

Because campuses were not randomized to Paloma, the comparison group was constructed by propensity-score matching without replacement. Each treated student was paired with comparison students from other district campuses, with exact matching on grade and campus program type and a caliper restriction on the propensity score. The propensity model used student-level pre-treatment characteristics: beginning-of-year DIBELS composite, attendance rate, sex, race and ethnicity, economic disadvantage, special-education status, emergent-bilingual status, and gifted-and-talented status.

Campus-level covariates – the campus mean BOY score and campus emergent-bilingual rate – were deliberately excluded from the propensity model and reserved for the outcome regression, a doubly-robust specification. This design lets the propensity model focus on matching students who look alike on individual characteristics, while the outcome model absorbs structural campus-level differences. Three matching ratios were estimated (1:1, 1:2, 1:3) as a robustness check; 1:2 is the primary specification.

4.4 Baseline equivalence

On the 1:2 matched offered sample, the treatment and comparison groups are equivalent on the beginning-of-year reading pretest to a standardized difference of $g = -0.027$, well inside the conventional 0.05 threshold for baseline equivalence. The worst student-level covariate imbalance was 0.055 (gifted-and-talented status), which falls in the range where WWC requires statistical adjustment – provided here by the covariate-adjusted estimator. All other covariates were balanced within 0.05 SD.

4.5 Campus selection and confounding

Platform campuses were selected by the district prior to the 2025–26 school year, before beginning-of-year testing, on the basis of prior underperformance relative to the district. Selection was therefore related to the outcome domain – but in the direction adverse to the program: as §3.2 documents, treatment campuses began the year substantially behind (a raw student-level baseline difference of −0.32 SD), so unobserved disadvantages concentrated at struggling campuses weigh against, not toward, the estimated effects. Known campus-level programs were observed in the data and handled by design: matching was exact within grade × campus-program-type cells, so treated students were compared only to comparison students at campuses with identical program status. As an external evaluator, we cannot enumerate every district-initiated support that may have accompanied the underperformance designation – but the identification strategy does not require that we do. Any campus-level co-intervention would apply equally to students at platform campuses whose families completed no lessons, and these students show null effects on every outcome at every matching ratio (§5.6). The observed gains therefore cannot be attributed to campus selection, co-occurring campus-level supports, or regression to the mean, none of which distinguish users from non-users within the same campuses. One asymmetry follows from this logic: the falsification test most directly protects the used and weekly tiers, which contrast exposure within campuses; the intention-to-treat estimate,

where offer and campus coincide by construction, leans on the exact-cell matching and the adverse selection direction – one reason this paper treats it as the conservative causal floor.

5. Who Engages – and Why It Matters for the Results

5.1 Engagement overview

Among families onboarded onto the platform, 52 percent completed at least one lesson. That figure frames the central operational fact of a home-based, usage-dependent program: roughly half of the consequential variation in outcomes is determined before the lessons themselves do any work, by whether a family engages at all.

5.2 The selection model

Whether engagement is sorted by student characteristics bears directly on how the impact estimates should be read. A logistic regression modeled platform use among onboarded students as a function of pre-treatment characteristics, with cross-validated discrimination assessed via 5-fold stratified cross-validation. The model's out-of-sample AUC was 0.60 – barely above chance. Baseline characteristics are poor predictors of who engages.

5.3 The load-bearing null

The result that matters most for the rest of the paper is a null: **starting reading level does not predict use** (OR = 1.04 per standard deviation, 95% CI [0.95, 1.14], p = .44). The confidence interval is tight, bounding any ability-based selection into engagement to at most a small effect in either direction. This is the empirical fact that licenses reading the as-implemented effect as a program effect rather than as stronger readers selecting into use.

Predictor	OR	95% CI	p
Starting DIBELS (+1 SD)	1.04	[0.95, 1.14]	.44
Attendance rate (+1 SD)	**1.27**	**[1.16, 1.40]**	**<.001**
Hispanic	1.18	[1.03, 1.35]	.02
Special education	0.91	[0.84, 0.99]	.03
Emergent bilingual	1.12	[1.00, 1.26]	.05

Table 3. Predictors of platform use among onboarded students (mutually adjusted logistic regression).

5.4 What predicts use: attendance and family routine

The strongest correlate of use is attendance (OR = 1.27 per standard deviation, $p < .001$). Because Paloma is completed at home, not at school, school attendance does not mechanically determine whether a family uses it. The association more plausibly reflects a household that maintains consistent routines – a family that gets a child to school reliably is also a family likely to maintain a home-learning cadence. Attendance is therefore best read as a marker of family routine and engagement capacity rather than as exposure to anything happening at school.

This connects directly to the theory of change. Paloma is a parent-coached model, and the families who engage are the ones with consistent routines, not the ones with higher-performing kids. The implication for the impact claims is substantial: the used-tier effects are correlational, but the selection is on routine, not on ability. That substantially narrows the plausible confounding story.

5.5 Dose-response: engagement and outcomes

Benchmark attainment rises monotonically with usage intensity. Students whose families completed no lessons attained benchmark at or below the comparison rate; every level of engagement above zero exceeded it, with the largest single increment occurring between no use and any use.

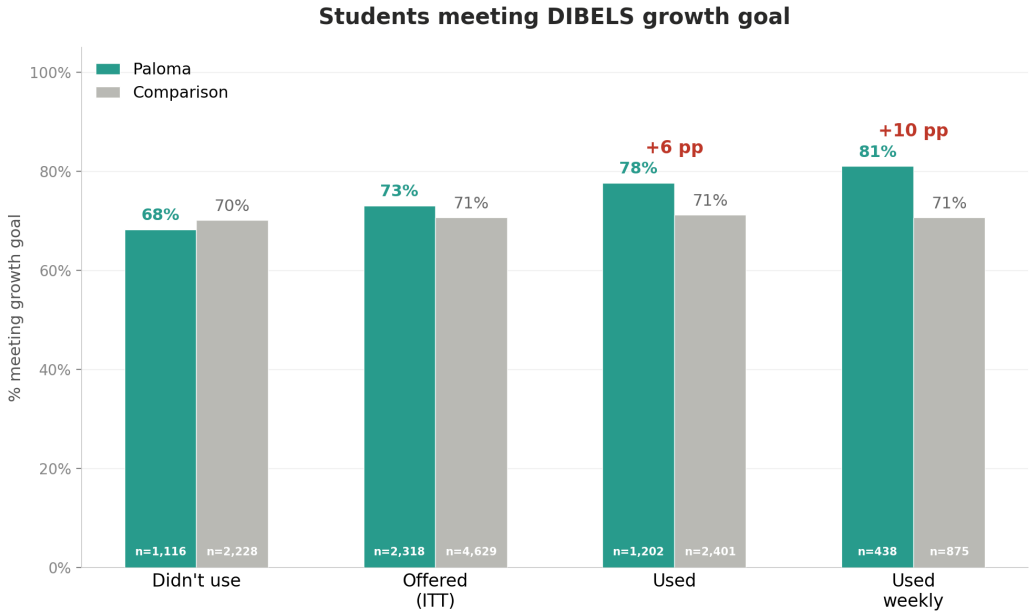


Figure 1. Dose-response: growth-goal attainment by usage level.

5.6 A seat without use does nothing: the falsification test

Compared against their paired controls, students whose families completed zero lessons show a null effect on every outcome – benchmark −0.3 pp (p = .85), score −1.5 points (p = .30), growth −1.2 pp (p = .48). Access without use confers no advantage. This is strong evidence that the effects observed among users are produced by use rather than by campus selection.

6. Does Paloma Work?

6.1 Estimation

Effects were estimated on the matched samples with covariate adjustment for the matching variables plus campus-level covariates – a doubly-robust specification – using heteroskedasticity-robust (HC1) standard errors. The paired-controls design ensures that when the treated side is subset to a usage tier, only that tier's matched comparison partners are retained.

6.2 Benchmark attainment: the primary result

On the primary outcome, the intention-to-treat effect is positive and statistically significant, and the effect rises monotonically with engagement.

Exposure	Paloma	Control	Adj. diff	p	g
Offered	63.9%	60.9%	+3.3 pp	.004	+0.07
Used	71.3%	63.4%	**+7.0 pp**	**<.001**	+0.15
Used weekly	77.4%	64.1%	**+10.3 pp**	**<.001**	+0.23

Table 4. Primary outcome – grade-level benchmark attainment by exposure (covariate-adjusted, 1:2 match).

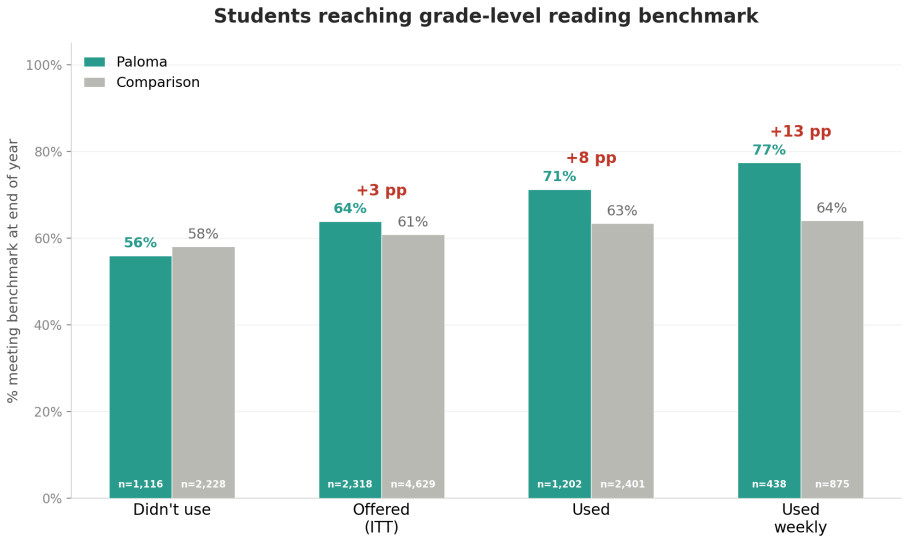


Figure 2. Students reaching grade-level reading benchmark, Paloma versus matched comparison, by exposure level.

6.3 Reading score: the magnitude of improvement

The continuous composite score confirms that benchmark gains reflect genuine movement in reading ability rather than students sitting just below the threshold being nudged over it.

Exposure	Paloma	Control	Adj. diff	p	g
Offered	450.4	449.1	+2.1 pts	.041	+0.04
Used	457.2	450.9	**+5.4 pts**	**<.001**	+0.11
Used weekly	466.6	452.4	**+10.5 pts**	**<.001**	+0.22

Table 5. Secondary outcome – continuous composite score by exposure (covariate-adjusted).

6.4 Growth-goal attainment

The growth-goal outcome confirms the pattern on a third measure.

Exposure	Paloma	Control	Adj. diff	p	g
Offered	73.1%	70.7%	+2.0 pp	.089	+0.04
Used	77.7%	71.2%	**+5.0 pp**	**.001**	+0.11
Used weekly	81.1%	70.7%	**+7.4 pp**	**.004**	+0.18

Table 6. Supporting outcome – growth-goal attainment by exposure (covariate-adjusted).

6.5 Effect sizes in context

Across nearly 750 randomized trials of education interventions, Kraft (2020) proposes empirical benchmarks of 0.05 SD (medium) and 0.20 SD (large), with higher thresholds for early literacy: 0.10 SD (medium) and 0.30 SD (large). Against that bar, the intention-to-treat effects sit at or near the medium threshold, the as-implemented effects are solidly medium, and the effects at the intended weekly dose approach or clear the large threshold.

6.6 Robustness

The estimates are remarkably stable across matching specifications. At 1:1, 1:2, and 1:3, the direction, magnitude, and significance of every finding hold. Specification robustness with respect to the attendance covariate is examined in §6.7.

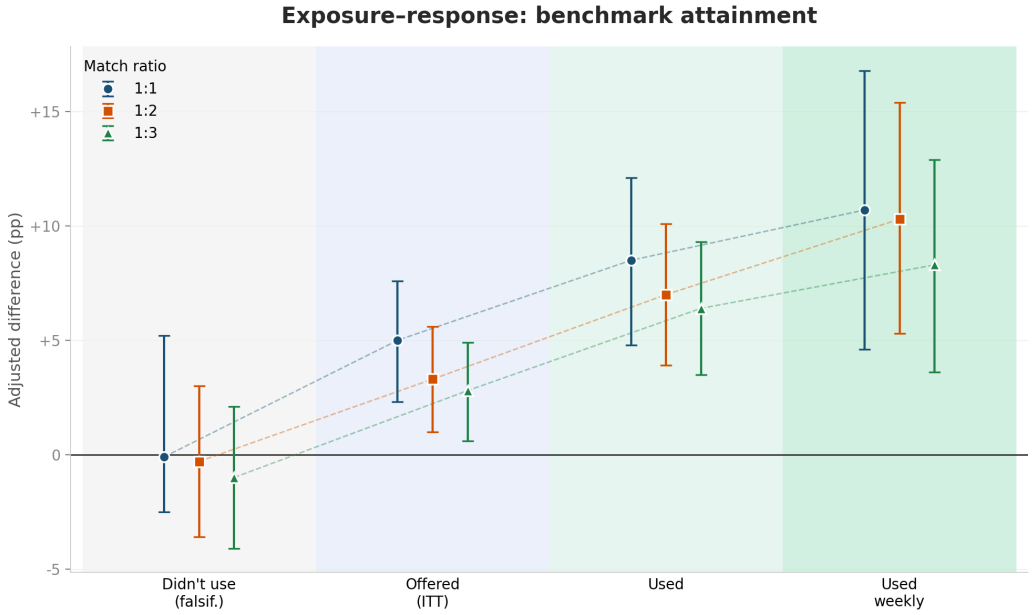


Figure 3. Exposure–response across matching ratios – benchmark attainment (pp). Results are stable across all three specifications.

6.7 Specification robustness

The analytic models include students' attendance rate as a covariate, per the district's specification. Because attendance was measured over the full intervention year rather than at baseline, WWC guidance treats such a variable as potentially endogenous – one the intervention could in principle influence. We tested this directly rather than assuming it: in the matched intention-to-treat sample, treatment status showed a small negative association with attendance (d = −0.05, p = .04). We judge this to reflect residual selection rather than a causal channel – a home-based reading program has no plausible mechanism for reducing school attendance – but we did not rest on that judgment. The full analysis was re-estimated with attendance excluded from both the propensity model and the outcome regressions, and the sample re-matched from scratch; pretest equivalence on the re-matched sample is g = −0.017, again within the range requiring no adjustment.

The re-estimated results tell the same story. On the primary outcome, the intention-to-treat effect is +2.9 percentage points (p = .014) against +3.3 reported, the used-tier effect +6.3 against +7.0, and the weekly-tier effect +8.1 against +10.3 – every estimate modestly smaller, every conclusion identical, with the continuous score and growth outcomes following the same pattern and the falsification tier null in both specifications. The one divergence runs in the excluded specification's favor: the intention-to-treat effect on growth-goal attainment reaches significance only without attendance (p = .009 versus p = .089 reported), and this paper relies

on it in neither. The body of the paper reports the attendance-inclusive estimates per the district's specification; the study's conclusions are unchanged under either, and readers applying WWC standards, under which contemporaneous covariates receive additional scrutiny, may treat the estimates in this section as the operative ones. Complete estimates for the attendance-free specification across all outcomes and matching ratios are available from the authors on request.

7. For Whom It Works Best

7.1 The below-benchmark population

Students scoring below the grade-appropriate DIBELS BOY benchmark cut constitute 68 percent of the treatment sample – a majority, not a tail. “Catching up” is the natural outcome for this population: beginning below the grade-level benchmark and reaching it by the end of the year.

7.2 Catch-up to grade level

Among students who began below benchmark, the share reaching grade level by spring exceeded the matched comparison at every level of exposure.

Exposure	Paloma	Control	Adj. diff	p	g
Offered	49.2%	43.9%	+5.3 pp	<.001	+0.11
Used	57.8%	46.9%	**+9.9 pp**	**<.001**	+0.20
Used weekly	63.1%	46.2%	**+13.6 pp**	**<.001**	+0.28

Table 7. Catch-up to grade level for students below benchmark at BOY (covariate-adjusted, 1:2 match).

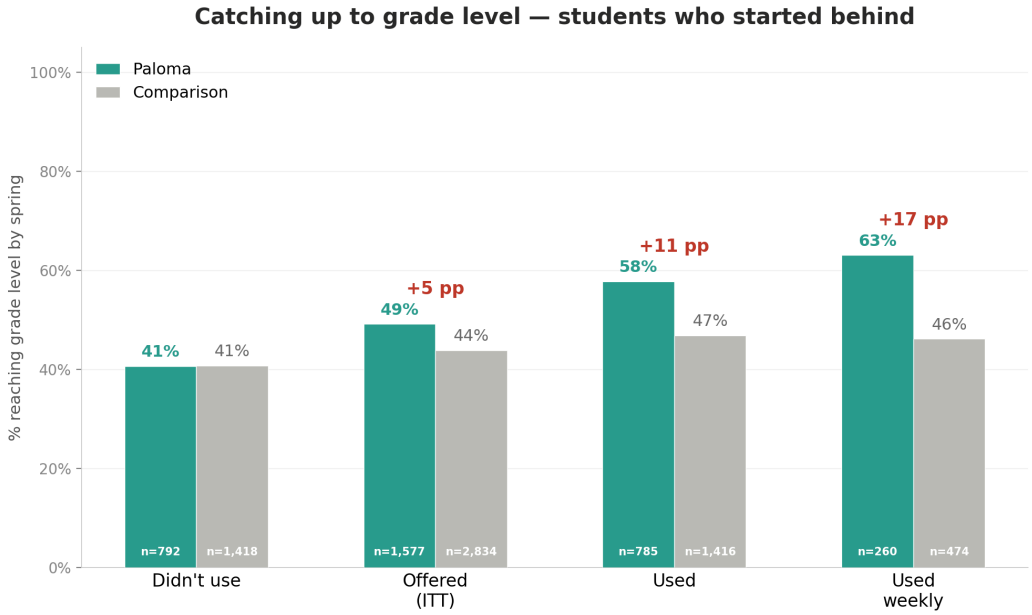


Figure 4. Catching up to grade level – students who started behind, by exposure level.

7.3 The catch-up is real gain

Within the below-benchmark stratum, the continuous score effect follows the same gradient – used +7.1 points (g = +0.17, p < .001), weekly +13.4 points (g = +0.31, p < .001) – confirming that students who reached grade level did so by improving as readers, not by sitting at the boundary and crossing it on noise.

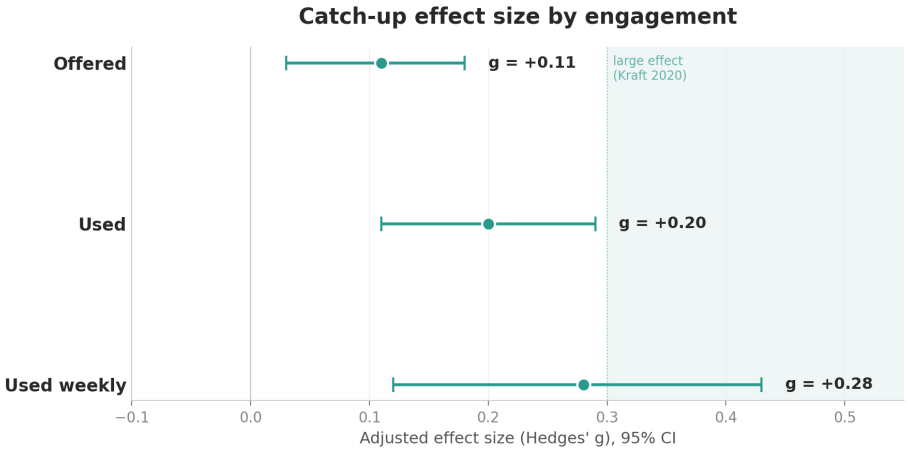


Figure 5. Catch-up (benchmark-attainment) effect sizes with 95% confidence intervals. The shaded region marks the range Kraft (2020) characterizes as a large effect for early-literacy interventions.

7.4 No gap-widening

The 32 percent of treatment students who began at or above benchmark show no measurable negative effect on any outcome at the offered tier: benchmark is maintained at 95 percent in both arms. The measurable benefit is concentrated among students who started behind, while students who started ahead held their position. Paloma narrows the achievement gap rather than widening it.

7.5 The equity arc, synthesized

Students who began behind were reached at the same rate as everyone else – starting reading level did not predict engagement. They benefited when their families engaged, at a magnitude exceeding the full sample. The effect was concentrated among students below benchmark. Each beat is supported by its own estimate, and the beats agree with one another.

8. Consistency with Prior Evidence

The findings reported here are not the first evidence that consistent Paloma use moves early reading. A year earlier, in SY24-25, Paloma evaluated the program across a broader and messier footprint – 4,895 kindergarten through second-grade students at 34 schools, assessed on several different instruments (MAP, DIBELS, i-Ready, and others) standardized within school-grade, with comparison students drawn from the same schools through Mahalanobis matching. That study estimated an overall effect of consistent use of d = 0.125 (90% CI [0.026, 0.224], p < .001) over a mid-year-to-end-of-year window. The two studies were built differently and answer to different strengths; read together, they corroborate one another more than either does alone.

8.1 Two designs, two kinds of validity

The studies sit at opposite ends of a familiar trade-off. The SY24-25 analysis was a breadth study: many schools, many assessments, a heterogeneous population. Its value is external – an effect that appears across that many settings and instruments is unlikely to be an artifact of any single one. The present study is a depth study: one district, one assessment (DIBELS 8th Edition), one calendar, with a propensity-matched within-district comparison group, a baseline-equivalent matched sample, a falsification test, and a selection model. Its value is internal – the effect survives a credible identification strategy. Neither design substitutes for the other. The breadth study cannot rule out confounding as tightly; the depth study cannot speak to settings beyond its district. The case for Paloma is stronger because both exist.

8.2 The comparison must be made at the same dose

A cross-year comparison is only meaningful between like exposures, and the two studies define exposure differently. SY24-25 classified a student as treated only if usage cleared a weekly-cadence-and-persistence bar – at least one lesson per week across at least four active weeks. In the vocabulary of this paper, that is not the intention-to-treat offered tier, nor the any use tier; it corresponds to the weekly, regular-use tier – the program's intended dose. The right cross-year comparison, then, is regular use against regular use.

The two studies' subgroup definitions are not commensurable in the same way, and should not be lined up against each other. SY24-25's "below peers" is a norm-referenced bottom tail ($z < -1.0$, roughly the lowest sixth of that sample); this paper's "below benchmark" is a criterion threshold that captures a majority (68 percent). A child can fall below the DIBELS benchmark while sitting near the middle of their local peer distribution. Because the two groups overlap poorly, the cross-year comparison is drawn at the overall regular-use level, not between subgroups.

8.3 The effect scales with the exposure window

At that matched dose, last year's regular-use effect (d = 0.125) is roughly half this year's ($g \approx 0.22$ on the composite score, $g \approx 0.23$ on benchmark attainment). The most direct reading of that difference is the measurement window. SY24-25 observed students over a mid-year-to-end-of-year span of roughly five months; this study follows them across the full year, roughly nine. The ratio of effects – about 1.8 – tracks the ratio of windows closely. Put in the units SY24-25 used, last year's effect accrued at roughly 0.025 SD per month of exposure; carried across a full nine-month year, that rate projects an effect near 0.22 – within range of

what this study observes at the weekly dose. An effect that grows in proportion to exposure time, across two independent samples, is the signature of a real treatment effect rather than a chance result.

Two qualifications keep that reading honest. First, the proportionality is directional, not exact: early-literacy growth is not linear across the year, the benchmark thresholds shift at mid-year, and the beginning-to-mid-year half of the year – which SY24-25 never observed – need not behave like the half it did. Second, the window is not the only thing that differs. This study's regular users are a higher-need population than last year's full sample, and effects run larger for students who begin further behind, so some portion of the difference reflects composition rather than time. The defensible statement is that this year's larger effect is consistent with a longer measurement window on a higher-need sample – not that the program became more effective. It is the same program, measured over more of the year.

8.4 The specific study answers the general one's open questions

The strongest link between the two papers is that this one resolves the limitations the earlier one named about itself. SY24-25 reported that its results, though significant, would not survive much unmeasured confounding – a Rosenbaum sensitivity bound of $\Gamma = 1.2$ – and singled out parental engagement and home literacy environment as the likely culprits. This study tests that worry directly: a student's starting reading level does not predict whether the family engages (OR = 1.04 per SD, $p = .44$), and students who were offered the platform but completed no lessons show a null effect on every outcome. The confounding channel SY24-25 could only flag, this study measures and finds wanting. SY24-25 also named assessment heterogeneity – many instruments, much noise – as an irreducible limitation; this study runs on a single DIBELS composite. The depth study does not merely agree with the breadth study; it closes the questions the breadth study left open.

Taken together, the two analyses describe the same program from two directions. One established that the effect appears across many schools and instruments; the other, that it withstands a credible causal design within a single district. They agree at the dose where they overlap, and the difference in their headline magnitudes is accounted for by how much of the year each one watched, and whom. That convergence – independent samples, different methods, a consistent conclusion – is what turns two studies into an evidence base.

8.5 The grade-level nulls do not replicate

SY24-25 reported one result this study is positioned to revisit directly: a strong effect for kindergartners ($d \approx 0.33$) but null effects for first and second graders ($d \approx 0.03$ and 0.06), which that paper attributed to the noise of combining several different assessment instruments. Two features of the present study make it a clean test of whether those nulls were real. It runs on a single instrument (DIBELS 8th Edition), removing the assessment heterogeneity SY24-25 named as the likely culprit; and its regular-use tier is the dose that matches SY24-25's treatment definition, so the comparison is made at the same exposure.

At that matched dose, the grade-level nulls do not reappear. On the continuous composite score – the unit comparable to SY24-25's effect sizes – every grade shows a positive, statistically significant effect among regular users: kindergarten $g = +0.24$ ($p = .01$), first grade $g = +0.24$ ($p = .005$), and second grade $g = +0.17$ ($p = .02$). The any-use tier, which is better powered, tells the same story on both score and benchmark in all three grades. The effect is not confined to

kindergarten; it reaches first and second graders too. If anything, the single largest grade-level effect on benchmark attainment falls in first grade (+14.1 pp, g = +0.33) – the precise grade SY24-25 estimated at zero. The grade pattern reads more sharply on benchmark than on score because the benchmark bar rises by grade: below-benchmark second graders gain on the continuous score but clear the steeper second-grade threshold less often, so benchmark understates their progress where score captures it. Read together, the two outcomes agree that the gains are broad across grades, with benchmark simply a higher hurdle in the later grade.

What this does and does not establish deserves the same discipline applied throughout this section. It establishes that SY24-25's grade-level nulls do not replicate: on a single high-quality instrument, over a full year, at the matched dose, the effect is positive in every grade. It does not, on its own, prove those nulls were caused by assessment noise. The cleaner instrument is one candidate explanation; the longer measurement window (a full year here versus half a year there) and the more responsive, higher-need population are two others, and this design cannot separate them – all three would predict the disappearance of a small grade-specific null. The honest claim is therefore the replication claim rather than the mechanism: where SY24-25 saw nothing in first and second grade, this study sees a consistent effect, which removes the grade-level nulls as a mark against the program even if it cannot fully diagnose why they appeared. As elsewhere in this comparison, the magnitudes are not lined up directly – SY24-25's figures are half-year mid-to-end-of-year effects and these are full-year effects – so what travels across the two studies is the pattern of sign and significance by grade, not the size of any single estimate.

9. What This Means for the District

The 14 campuses where Paloma was deployed started the year substantially behind the district average on benchmark attainment – 34 percent of Paloma users at or above grade level, compared to 46 percent district-wide, a 12 percentage-point gap. By the end of the year, students who used Paloma had effectively erased that gap.

	BOY	EOY	Gain
District average	46%	70%	+24 pp
Paloma users	**34%**	**71%**	**+37 pp**

Table 8. Benchmark attainment, BOY to EOY: Paloma users versus district average.

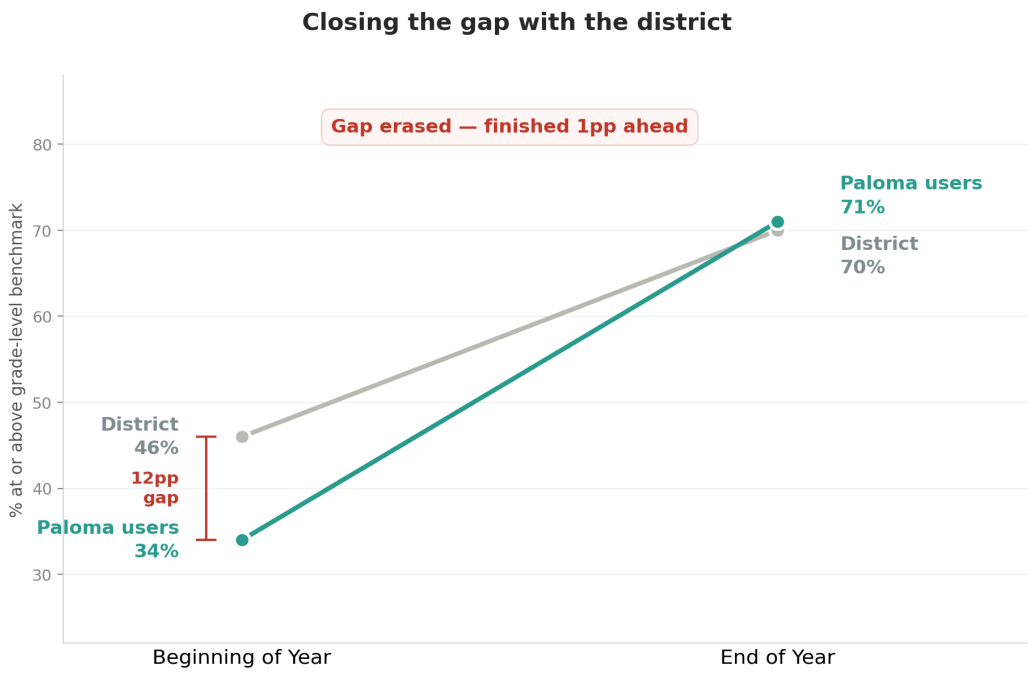


Figure 6. Closing the gap with the district. Paloma users started behind the district average and matched it by end of year.

This is a descriptive finding, not a causal one – some of the closure reflects regression to the mean, and the matched analysis in Section 6 is what isolates Paloma's contribution. But the descriptive fact is consistent with the causal story, and it is the version that speaks most directly to what a district wants to know: can this program help our highest-need campuses catch up?

10. Cost-Effectiveness

Effect sizes do not speak to policy on their own. As Kraft (2020) argues, effect-size benchmarks must be read alongside cost; he pairs his benchmarks with empirical per-pupil cost tiers of under $500 (low), $500 to $4,000 (moderate), and $4,000 or more (high). At $50 per student per year, Paloma sits an order of magnitude below the $500 ceiling – not merely inside the low-cost tier, but a tenth of the way into it.

What makes that price striking is the company Paloma's effects keep. The most rigorously evaluated tool for moving early reading is tutoring, and the largest meta-analysis of the experimental evidence – Nickow, Oreopoulos, and Quan (2024), pooling roughly 200 randomized studies – estimates an average tutoring effect of 0.288 SD. Paloma's effect at the intended weekly dose – g ≈ 0.22 on the composite score, g ≈ 0.23 on benchmark attainment, and g ≈ 0.28 to 0.31 among students who began below grade level – falls squarely in that range. In effect-size terms, weekly Paloma use is comparable to the average tutoring program.

On the dimension that matters most for a parent-led design, that comparison is conservative. The same meta-analysis finds tutoring effects are largest when delivered by teachers and paraprofessionals, with parent- and volunteer-delivered tutoring occupying the lower end of the

impact distribution. Paloma is a parent-coached model – the delivery channel the evidence expects to be weakest – and it nonetheless produces effects in the range of tutoring as a whole. It reaches a tutoring-grade result through the least expensive, and by the evidence historically least effective, delivery mechanism.

The cost contrast closes the argument. Tutoring that achieves effects of this size is labor-intensive by construction – frequent sessions with trained tutors – placing it in the moderate-to-high cost tiers of Kraft's schema, where per-student costs run into the hundreds or thousands. Paloma reaches a comparable effect at $50. Placing cost and effect together locates Paloma in the most favorable cell of Kraft's cost-by-effectiveness schema: medium-to-large early-literacy effects at the intended dose, delivered at the lowest cost.

Two honest qualifications. First, the 0.288 figure is the meta-analysis's grand pooled mean, averaged across grades, across reading and math, and across every type of tutor. A benchmark tailored to Paloma's exact case would move in two opposing directions – reading tutoring in the early grades runs above the pooled mean, raising the bar, while parent-delivered tutoring runs below it, lowering the bar – and these adjustments roughly cancel, which is what makes the pooled figure a fair reference rather than a flattering one. Second, the $50 price is Paloma's monetary cost and does not monetize parent time; even so, it is low enough that the cost-effectiveness conclusion holds under any reasonable valuation of the time families contribute.

11. Evidence Standards and Limitations

This study was designed and analyzed to meet the evidence requirements of ESSA "Moderate Evidence" (Tier II): a well-implemented quasi-experimental design with a statistically significant positive effect on a pre-specified primary outcome. The elements that standard scrutinizes are documented above – baseline equivalence on the matched analytic sample within the range requiring no statistical adjustment (§4.4), with covariate adjustment applied in the estimator regardless; attrition low under both WWC boundary assumptions (§4.2); a confounding assessment addressing campus selection and co-occurring supports (§4.5); and a pre-specified primary outcome, with secondary outcomes labeled as such throughout. Specification sensitivity, including the district-specified attendance covariate, is disclosed in §6.7. Formal evidence-tier designation rests with external reviewers, not with study authors; this paper's claim is that the design satisfies the published requirements, and the sections above are organized so a reviewer can verify each one directly.

Its limitations are the ones inherent to its design. The evidence comes from a single district in a single school year; the prior year's multi-school study (§8) extends the footprint but was built on a different design, and replication in additional districts remains the most valuable next evidence. The weekly-use tier is correlational, as labeled – families choose their engagement level, and while the selection model bounds ability-based sorting tightly (§5.3), it cannot rule out selection on unmeasured family factors. And the intention-to-treat estimate, while the study's most conservative causal quantity, coincides with campus-level receipt by construction, as §4.5

discusses. None of these limitations is unusual for a study of this type; all of them are stated so that the evidence can be weighed as it is.

12. Conclusion

The evidence assembled here points in one direction. Across a large urban district, students whose families were offered Paloma reached grade-level reading benchmark at higher rates than matched comparison students – and the more a family engaged, the larger the gain. That gradient is the signature of a program that works: a conservative intention-to-treat estimate that does not lean on usage, rising to a medium-to-large effect at the dose the program was built around.

What gives the finding its weight is everything it survived. Baseline equivalence was tight. A student who was offered the platform but never opened it gained nothing, ruling out the possibility that the result was an artifact of which campuses adopted Paloma. Engagement was not sorted by starting ability, so the effect among users is not stronger readers selecting into use. And the continuous reading score moved in step with benchmark attainment, confirming that students crossed the line by becoming better readers rather than drifting over it on noise. Each check that could have explained the result away instead left it standing.

The benefit, moreover, landed where it was needed most. Among the majority of students who began below grade level, the share reaching benchmark by spring outpaced matched comparison students at every level of exposure, while students already at grade level held their ground – Paloma narrowed the achievement gap rather than widening it. At the district's highest-need campuses, users began the year behind the district average and matched it by spring. And this was not a single year's good fortune: the pattern echoes the prior year's broader, multi-school study at the dose where the two overlap, two independent samples reaching the same conclusion by different routes.

Set against the cost, the result is hard to overstate. Paloma delivers reading effects comparable to the average tutoring program – through a parent-led model the evidence expects to be the weakest delivery channel, and at $50 per student per year, an order of magnitude below the field's low-cost threshold and a fraction of what tutoring of similar effect costs. A medium-to-large early-literacy effect, concentrated on the students furthest behind, consistent across two years, at a price that places it in the most favorable cell of the cost-effectiveness frontier. For a district asking whether its highest-need readers can catch up, and at what cost, the answer this study returns is unambiguous.

References

Escueta, M., Nickow, A. J., Oreopoulos, P., & Quan, V. (2020). Upgrading education with technology: Insights from experimental research. Journal of Economic Literature, 58(4), 897-996.

Hernandez, D. J. (2011). Double jeopardy: How third-grade reading skills and poverty influence high school graduation. The Annie E. Casey Foundation.

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

Nickow, A., Oreopoulos, P., & Quan, V. (2024). The Promise of Tutoring for PreK-12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence. American Educational Research Journal, 61(1), 74-107