WORKING PAPER Nº IDB-WP-1863

When Does EdTech Work?

Evidence from a Scalable Home-Based Intervention with School Complementarity

Juanita Bloomfield
Ana Balsa
Alejandro Cid

Inter-American Development Bank
Gender and Diversity Division
Department of Research and Chief Economist

September 2026

When Does EdTech Work?

Evidence from a Scalable Home-Based Intervention with School Complementarity

Juanita Bloomfield*
Ana Balsa*
Alejandro Cid*

* Universidad de Montevideo

Inter-American Development Bank
Gender and Diversity Division
Department of Research and Chief Economist

September 2026

Abstract*

Despite growing investment in education technology, evidence on how to integrate digital tools effectively into public education systems remains limited. Substituting classroom instruction with computer-aided learning has often failed to improve outcomes, even though schools remain the most direct route to scale educational innovations. This paper studies an alternative approach: delivering personalized educational software to households, where smartphone access is widespread, while experimentally varying school involvement. We evaluate Leo Leo, an adaptive early literacy app for young children, using a randomized controlled trial in 105 public schools in Nuevo León, Mexico. Schools were assigned to either (i) a family-only intervention providing access to the app, (ii) a combined family-and-school intervention that additionally provided teachers and principals with information on children's progress, or (iii) a control group that received neither the app nor school support. The home-based technology generates robust gains in early literacy, with suggestive positive spillovers to early numeracy, when complemented by school support; the family-only intervention produces no significant learning gains. Descriptive mechanism analyses are most consistent with greater child engagement with the software as the leading channel. We find no indication that changes in parental inputs play a role, based on a household survey with a lower response than the student assessment. Overall, the results show that while home-based delivery of educational technology can enhance scalability, institutional involvement remains critical for translating scalable technology into learning gains.

JEL codes: I21, C93, O15

Keywords: randomized controlled trials, EdTech, early childhood development

*Department of Economics, Universidad de Montevideo, Av. Dra. María Luisa Saldún de Rodríguez 2097, Montevideo, Uruguay; phone: +59827074461. Juanita Bloomfield: j.bloomfield@um.edu.uy; Ana Balsa: abalsa@um.edu.uy; Alejandro Cid: acid@um.edu.uy. We acknowledge financial support from the GDLAB from the Inter-American Development Bank. We greatly appreciate comments from Julian Cristia, Belen Michel, and Agustin Pardo. We are grateful to Ignacia Donoso, Guadalupe Almeida, and Nicolás Anido for excellent assistance and supervision in the implementation and evaluation of the program. This study was registered on the AEA RCT Registry (AEARCTR-0014912).

1 Introduction

Around the world, even when children attend school, many fail to acquire foundational literacy and numeracy skills, particularly in low- and middle-income countries (World Bank Group, 2022; Glewwe and Muralidharan, 2016; Alvarez Marinelli et al., 2025). A central challenge is heterogeneity: Children within classrooms often differ by several grade levels in their skills, making it difficult for teachers to tailor instruction to individual learning needs using standard curriculum-driven approaches (Banerjee et al., 2007; Pritchett and Beatty, 2015; Muralidharan et al., 2019). This mismatch between instruction and learning levels contributes to persistent learning gaps.

Educational technology (EdTech) has attracted growing interest as a potentially scalable way to address this challenge. Adaptive software and computer-assisted learning tools can personalize instruction to children's developmental levels and provide practice and feedback at low marginal cost (Bulman and Fairlie, 2016; Escueta et al., 2020; Rodriguez-Segura, 2022; Singh et al., 2025). Delivering these technologies through the home could substantially expand access while reducing many of the implementation challenges associated with classroom-based programs. However, moving instruction outside the classroom also removes the institutional environment that may be critical for translating access to technology into learning. Whether adaptive EdTech remains effective under home-based delivery, and whether schools continue to play an important complementary role under this delivery model, therefore remains an open question. Although home-based delivery offers a potentially more scalable alternative, there is remarkably little experimental evidence addressing this question.

Most studies evaluate EdTech interventions implemented through schools. While results are mixed, a growing body of evidence suggests EdTech improves learning when it complements teachers and existing institutions but often fails when it substitutes for them (Malamud and Pop-Eleches, 2011; Beuermann et al., 2015; Cristia et al., 2017; Bettinger et al., 2023; Beg et al., 2022). At the same time, integrating technology into regular classroom instruction is operationally complex, requiring hardware investments, teacher training, adjustments to instructional time, and sustained supervision. These implementation frictions help explain why many promising EdTech interventions struggle to scale effectively through public school systems (Muralidharan and Singh, 2025). Home-based delivery offers a potentially more scalable alternative, but little is known about whether EdTech outside the classroom can generate meaningful learning gains or whether institutional support is necessary.

This paper studies whether personalized EdTech can improve learning when delivered through the home rather than directly through classroom instruction and whether institutional complements remain important under this delivery model. We assess the impact of Leo Leo, an adaptive early literacy program designed for children aged five. The program delivers personalized literacy activities through a mobile app designed to be used at home in short, frequent sessions.

We evaluate the program through a randomized controlled trial conducted during the 2024–25 academic year in 105 public schools in Nuevo León, Mexico. Schools were randomly assigned to one of three groups: a control group that was not exposed to the app; a family-only treatment that provided families with access to the app; and a combined family-and-school treatment that additionally provided teachers and principals with regular progress reports on children's use of the app. This three-arm design allows us to separately identify the effects of home-based access to adaptive technology and the additional role of institutional involvement in sustaining engagement and supporting learning.

We examine a broad set of outcomes capturing both impacts and mechanisms. We assess children's literacy and numeracy skills, measure engagement and progression within the app, and study intermediate outcomes related to parental involvement and school support. We also examine potential unintended effects on socioemotional development and screen use.

We find positive and statistically significant effects of the combined family-and-school intervention on children's progression within the literacy app and on several learning outcomes, including literacy and early numeracy skills. By contrast, the family-only intervention generates no robust learning gains relative to the control group. Descriptive mechanism analyses point to greater child engagement with the app as the leading channel, but because engagement is itself an outcome of the intervention rather than an independently assigned mediator, we interpret this as suggestive of a channel rather than as a causally identified decomposition. We find no indication that changes in parental investments play a role, though this comes from a household survey with a lower response rate than our main sample, and should be interpreted with caution. Instead, children in schools receiving teacher and principal support use the app more intensively, and schools receiving progress reports appear to reinforce usage and incorporate the technology into instructional practices. We find no adverse socioemotional effects of the combined intervention, while the family-only treatment is associated with suggestive increases in hyperactivity and problematic screen use.

The paper makes two main contributions, complemented by a third contribution. While a large literature evaluates school-based EdTech, only a handful of studies examine home-delivered EdTech experimentally (Beuermann et al., 2015; Bresciani et al., 2026; Arnold et al., 2021; Niklas et al., 2025). None of them, to our knowledge, vary school involvement, so none can separately identify the role of institutional support once the technology reaches the home. Our study tests a home-based intervention at the scale of a public school system in a middle-income country and is the first, to our knowledge, to randomize school-level information provision in order to isolate the incremental role of schools once household access to the technology is held constant. We examine these questions in the context of early literacy, extending a literature on adaptive EdTech that has focused primarily on mathematics, for which content is more easily modularized and aligned with adaptive instruction (Escueta et al., 2020); literacy acquisition relies on a broader set of foundational skills and may pose distinct challenges for scalable digital delivery.

The most direct antecedent for our study is by Bresciani et al. (2026), who conduct a six-month randomized experiment (About TIME) with 459 preschool-age children in Chicago, comparing a group that receives a tablet preloaded with math apps to a group that receives analog math materials, without any accompanying school or teacher involvement. On average, they find no effect of the math app on children's math skills, though children of parents without a college degree gain 0.17 standard deviations, closing roughly one third of the initial skill gap relative to children of college-educated parents. Unlike About TIME, which supplies each family with a dedicated tablet, Leo Leo runs on families' own smartphones, so our design tests a delivery model that is more scalable because it does not require hardware provision.

Our family-only arm is similarly limited to household access to Leo Leo, but unlike in Bresciani et al. (2026) it produces no significant gains even among children of less educated caregivers; in our setting, gains emerge only when teachers and principals also receive information on children's engagement with the app.

The second contribution of this paper is to a growing literature studying low-cost digital interventions aimed at supporting children's development through mobile phones and digital messaging (Doss et al., 2019; Mayer et al., 2019; Bloomfield et al., 2023; Balsa et al., 2023, 2024). Most of these interventions seek to improve child outcomes indirectly by changing parental behavior through information provision, coaching, or behavioral nudges. By contrast, Leo Leo delivers personalized instructional content directly to the child while preserving a highly scalable home-based delivery model. This design allows us to distinguish improvements arising from institutional support from those operating through changes in parental investments.

Finally, we contribute to a growing literature examining the socioemotional consequences of EdTech. The growing use of EdTech has raised concerns about its implications for children's socioemotional well-being and mental health (Abrahamsson, 2026; Braghieri et al., 2022; Allcott et al., 2020). As concerns about children's screen exposure have become increasingly central to education policy, experimental evidence on such effects remains limited, particularly for interventions targeting young children. By jointly evaluating children's learning outcomes and socioemotional well-being, we provide evidence on both the educational benefits and the potential unintended consequences of adaptive EdTech.

The remainder of the paper is structured as follows: Section 2 describes the Leo Leo program; Section 3 presents the experimental design and data; Section 4 reports the results; and Section 5 concludes.

2 Context and the Leo Leo Program

Mexico is an upper-middle-income country with an estimated population of 126 million. According to the 2022 PISA assessment, Mexico ranked 35th out of 37 OECD countries evaluated, with reading scores 61 points below the average. Nearly half of its students did not reach the minimum proficiency level in reading.

The intervention took place in Nuevo León, a relatively prosperous and industrialized state in northeastern Mexico, and was implemented in collaboration with its secretary of education, which facilitated outreach to participating schools and families. Maternal education in the state is relatively high (mothers have a literacy rate of 97.8% and an average of 10.6 accumulated years of schooling), though wide economic disparities remain—in terms of educational outcomes and access to resources—across population groups.

Nuevo León exhibits levels of digital connectivity above the national average, which justifies the assumption of widespread smartphone access underlying our scalability argument. According to Mexico's National Survey on Availability and Use of Information Technologies in Households, Nuevo León ranks among the states with the highest share of households with internet access and internet-connected smart devices (INEGI, 2025, 2026), well above both the national average and, in particular, to rural areas nationwide, where internet use is markedly lower than in urban areas (INEGI, 2024). Even so, the program launch required bringing satellite internet devices to activation events (Section 3.3), suggesting that stable connectivity was not universal even in this comparatively favorable context. Generalizing our results to rural settings or to countries with lower smartphone penetration should therefore be done with caution.

Leo Leo is an educational mobile app on the Wumbox educational platform. The app is designed for children to learn to read and write and is adapted to children's different ability levels. The main theory guiding the intervention is the Response to Intervention (RTI) model (Jiménez, 2019), a widely used educational framework that emphasizes early identification of learning needs, continuous monitoring, and data-based decision-making to modify instruction according to each student's needs and provide support. According to Jiménez (2019), the ingredients necessary for developing reading proficiency include vocabulary, alphabetic knowledge, phonemic awareness, fluency, and comprehension. The RTI framework also highlights strategies that have proven effective in reading instruction. The design of Leo Leo is closely aligned with these principles and integrates these components into its environment.

The app is easy to use and features seven interactive games and activities (see Appendix A of the Supplementary Material), each targeting a specific literacy skill through explicit teaching, scaffolding, and gamified elements designed to keep children engaged. The difficulty level adapts to a child's progress, moving from fully guided activities toward tasks that demand greater autonomy, with scaffolds and explanatory videos available whenever a child struggles. Between activities, short mini games reinforce complementary skills such as fine motor coordination, and children can track their progress toward their own goals with immediate feedback.

The app is compatible with low- and mid-range mobile devices across all operating systems and does not require a constant internet connection, making it particularly suitable for low-connectivity contexts. Its design enables short daily sessions that end after 20 minutes of use. The program also integrates real-time data collection to monitor children's progress. Based on estimated monthly sessions and total minutes, and assuming three sessions per week at 20 minutes each, a child would reach 100% progress in the app after 54 twenty-minute sessions over five months of play, equivalent to about 3.6 hours per month.

3 Impact evaluation: Methodology

3.1 Experimental design: Randomization

The evaluation uses a randomized controlled trial with 105 public schools in Nuevo León, Mexico. Schools were randomly assigned to one of three groups, with 35 schools each. Randomization was stratified by school size (schools with student populations above or below 60, the average enrollment for third-year preschool).

Treatment 1: Family- and school-level intervention

This group received a combined intervention at both the family and school levels. Families were given access to Leo Leo and encouraged to use it with their children for about 20 minutes, at least three times per week. In addition, they received automated WhatsApp messages tailored to their child's in-app activity. When the child was active and making progress, families received two messages per week reinforcing continued use (see Appendix B of the Supplementary Material for message examples). If activity declined, automated reminders were triggered to encourage re-engagement.

When the child was inactive, the messaging system followed a simple escalation protocol. If the child remained inactive for one week, a re-engagement message was sent to the family and the teacher was notified. If inactivity persisted for two weeks, the family received an additional message to help address potential barriers. Only in cases of prolonged inactivity were teachers prompted to follow up directly. Families also had access to a dedicated WhatsApp contact for technical support if needed.

In the school-based component of the treatment, teachers and principals were given access to the app to track individual student progress and receive reports every other week through WhatsApp highlighting each student's learning status. These reports classified students into three categories to guide teachers: (i) OK (the student is progressing as expected); (ii) Priority (the student is slightly behind or has not used the app for two weeks); and (iii) Critical (the student is significantly behind or has not used the app for three or more weeks). Reports also included information on the school's average progress and activation rates relative to expected benchmarks, along with messages tailored to the school's overall engagement and progress (see Appendix B of the Supplementary Material for examples of reports and messages). Teachers and principals were encouraged to reinforce the importance of using Leo Leo at home, particularly for students classified as Priority or Critical.

School leaders and teachers were also invited to participate in a series of virtual workshops on the use of the app and related literacy topics. The workshops were held monthly and combined practical guidance on the app with broader discussions of early literacy skills.[*]

Treatment 2: Family-level intervention only

Families in this group received the same family-level intervention as in Treatment 1. They were given access to the app and encouraged to use it with their children for 20 minutes, at least three times a week. Families received personalized messages via WhatsApp triggered by the child's in-app activity and had access to a dedicated WhatsApp contact from the

[*] The modules overview is the following: (i) Introduction to the Program (2 hours), (ii) Phonological and Alphabetic Awareness (1 hour), (iii) The Role of Motor Skills in Writing (1.5 hours), (iv) Reading Speed and Fluency (1 hour), (v) Vocabulary Development (1 hour), (vi) Reading Comprehension (1 hour), (vii) Specific Learning Difficulties (1.5 hours).

Wumbox team for technical support. Unlike with Treatment 1, there were no school reports or monthly workshops.

Control group

The control group did not receive any part of the intervention at either the family or school level. Control schools were offered the opportunity to implement the intervention in the subsequent academic year with the subsequent cohort.[†]

This design allows us to identify the overall effect of the intervention (Treatment 1 vs. Control; Treatment 2 vs. Control) and the incremental impact of school involvement (Treatment 1 vs. Treatment 2).

3.2 Study sample

The sample consisted of 105 participant schools with 6,332 children. Each treatment group consisted of 35 schools with 2,087 students in the control, 2,156 students in Treatment 1, and 2,089 students in Treatment 2. All children enrolled in the third year of preschool classes at the participating schools were included in the study. In Mexico, preschool education consists of three grades, typically serving children aged three to five. Third-grade preschool occurs in the final year before primary school, which children usually enter at age six. As preschool is compulsory, enrollment at age five is nearly universal.

Treatment was randomized at the school level. For data collection, all third-grade preschool classes within each school were considered; however, because of resource constraints, in schools with more than three classes, up to three were randomly selected for assessment and all children in the selected classes were evaluated.

Of the 105 schools initially randomized, seven declined to participate before the intervention began (Appendix Figure B1 summarizes the construction of the analysis sample). (Table C1 of the Supplementary Material shows that school participation was unrelated to treatment assignment.) This reduced the sample of schools to 98 and the sample of children to 5,905

[†]We believe the likelihood of significant contamination of the control group by either Leo Leo or comparable interventions is minimal. We never disclosed the name of the app to control schools or families, and we requested that supervisors and the secretary of education do the same. Moreover, even if a family in the control group downloaded Leo Leo independently, the app only allows free access to a small number of games. To unlock the full contents or continue using the app, a payment is required. At low socioeconomic levels, willingness to pay for such digital content is low, reducing even more the likelihood of significant use among control-group families. Unfortunately, we do not have information on the use of alternative educational apps, so we cannot fully rule out that some families may have independently used such apps. However, there is no indication that there is a widely used, comparable app that provides a similar pedagogical approach or skill coverage.

(see Table 1). Consequently, our intent-to-treat (ITT) estimates refer to the population of schools that agreed to participate following random assignment. Because no intervention activities occurred prior to these refusals, treatment exposure was unaffected.

Table 1: Sample after baseline attrition (schools that implemented the program)

	Control	Treatment 1	Treatment 2	Total
Schools	31	33	34	98
Children	1,874	1,992	2,039	5,905

Notes: Of the 105 initially randomized schools, 7 declined to participate before any intervention activities began (see Appendix Figure B1). Because no baseline or endline data were collected in these schools, the number of children is imputed using the average enrollment per school within the corresponding treatment arm. Treatment exposure was unaffected by these refusals, as they occurred prior to the start of implementation.

In Section 3.5 we present balance in the sample that implemented the intervention.

3.3 Implementation

In September 2024, families assigned to the treatment groups were invited to participate in a program launch event. During event sessions, parents received information about Leo Leo and were encouraged to engage with it at least three times per week for 20 minutes per session.

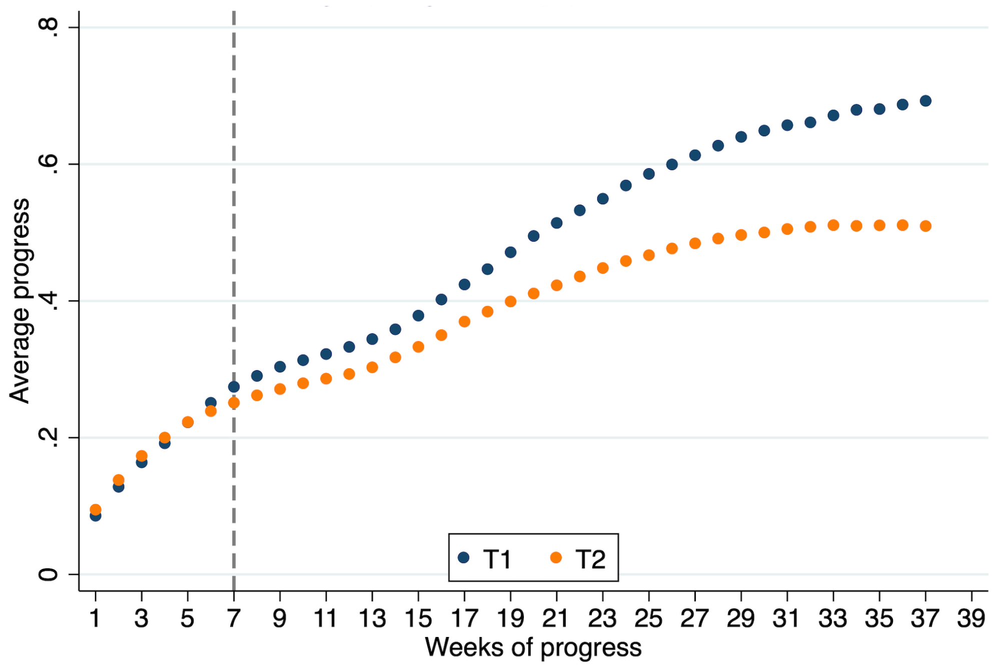
Given that many schools had limited or no Wi-Fi access, and since not all families had mobile data plans, we took satellite internet devices to the launch events to facilitate app downloads. This ensured families could install and activate the app, which requires internet connectivity for initial setup. Students received individual login credentials.

To accommodate families unable to attend the in-person launch, a virtual launch event was conducted via WhatsApp. This remote launch also served as a follow-up mechanism for those who participated in person, reinforcing login accuracy and preventing user account errors.

Following the program launch, implementation proceeded as planned. As of May 29, 2025, there were 1,657 active users in Treatment 1 and 1,285 in Treatment 2, implying 83% and 63% activation rates, respectively. Average app progress reached 69% in Treatment 1 and 51% in Treatment 2. Beginning in week 7, teachers received updates every two weeks on their students' app usage and progress. Figure 1 presents children's progress by week of the intervention; the dashed vertical line marks the point at which reports began to be distributed. After teachers started receiving the regular progress reports, student progress within the app increased in Treatment 1 relative to Treatment 2, and this gap persisted throughout the remainder of the intervention. Prior to the introduction of the reports, both groups exhibit similar trends. After week 7, Treatment 2 provides a plausible counterfactual for Treatment 1, suggesting that the acceleration in Treatment 1 reflects the impact of school support on sustained app usage.

Attendance at the workshops was heterogeneous and generally low among both teachers and principals in Treatment 1 schools, with average completion of the workshop series being 31.9% among teachers and 20.5% among principals; attendance declined after an initial peak in both groups, and roughly a third of teachers and half of principals did not attend any sessions. Despite this low and declining attendance, Treatment 1 generated substantial and robust learning gains relative to both the control group and Treatment 2, a pattern suggesting the WhatsApp progress reports, rather than the workshops, sustained engagement. Because attendance at the workshops was not randomized and instead reflected teachers' and principals' own choices, any association between attendance and outcomes would likely be confounded by unobserved characteristics such as motivation or effort, so we do not use it to draw causal conclusions about the workshops' contribution; the low and declining attendance nonetheless suggests their marginal contribution was likely limited relative to the automated reports. This distinction is also relevant for scalability since automated reporting is far less costly to replicate at scale than teacher workshops.

Figure 1: Progress in the app, by week of implementation



Notes: The figure plots average cumulative progress in the Leo Leo app by week of implementation for Treatment 1 (family + school) and Treatment 2 (family only). The dashed vertical line indicates the introduction of teacher progress reports (week 7).

3.4 Data

Our analysis combines primary data collected at baseline and endline on children, families, and teachers with administrative school records and system-generated data from the Leo Leo platform. Primary data collection was carried out by the research team between May and June 2025 for the endline evaluation, following approximately one academic year of program exposure.[‡] We briefly describe each data source below; detailed descriptions of instruments, item wording, and outcome construction are provided in Appendix A. Table G4 of the Supplementary Material shows descriptive statistics of outcomes assessed at follow-up in the control group.

3.4.1 Student learning outcomes

Student learning outcomes were measured at endline during regular school hours by trained enumerators. We focus on early literacy as the primary outcome domain and examine early numeracy as a non-targeted outcome to assess potential spillovers. Emerging research suggests that some cognitive skills contribute to learning across language, reading, and mathe-

[‡]The academic year in Mexico runs from September to June.

matics domains (Pham et al., 2025).

Literacy skills were measured using two complementary instruments. First, children completed the Leo Test, a tablet-based assessment available on the same educational platform as Leo Leo that evaluates progression across multiple stages of early literacy, including alphabetic knowledge, phonological awareness, reading, and writing. We focus on two main outcomes: the overall score and the number of subtests completed. Because the test stops when a child does not perform well on a specific subtest, completing more subtests indicates higher performance. We also examine performance on each individual subtest using indicator variables equal to one if the child successfully completed that subtest.

Second, children were assessed individually using IPAL (Alphabetic Knowledge and Phonological Awareness), which provides standardized measures of letter-name knowledge, letter-sound correspondence, and phonological awareness. Unlike the Leo Test, IPAL was administered independently by trained enumerators and therefore provides an external measure of foundational literacy skills. Considering both assessments allows us to distinguish improvements in children's performance within the digital learning environment from gains that generalize to an independently administered assessment.

To capture broader developmental effects beyond literacy, children completed the numeracy tasks of the International Development and Early Learning Assessment (IDELA), which cover number recognition and basic arithmetic operations.

The Leo Test was administered in small group settings using tablets, while IPAL and IDELA assessments were conducted individually in a quiet space within the school. Each assessment required approximately 15 minutes per child. For comparability across instruments, all continuous outcome measures are standardized to have a mean of zero and standard deviation of one in the control group. All participating families and schools provided informed consent prior to data collection.

3.4.2 Leo Leo data

The Leo Leo platform generates detailed system data on children's engagement and progression throughout the intervention period. Each child accesses the app using a unique identifier, which allows us to link platform activity to the experimental sample. Usage logs record the progression through literacy activities and completion of different stages of the learning sequence.

We use these data to construct indicators for whether children reach specific milestones within the app and an indicator for whether the child ever activated the app. These system-generated outcomes play a central role in our analysis of mechanisms, allowing us to examine whether differences in learning outcomes across treatment arms are linked to differences in app usage.

3.4.3 Secondary data on schools

We supplement the primary data with school-level administrative records from the Ministry of Education ("911" data). This source provides information on enrollment, staffing, and infrastructure. In addition, we merge municipality-level census data capturing socioeconomic characteristics such as maternal education in the school zone.

These secondary data are used to describe the study context and assess baseline balance.

3.4.4 Teacher questionnaire

Teachers completed questionnaires at both baseline and endline. In both evaluation periods, the survey collected information on teaching experience, classroom characteristics, and views on student learning. At endline, teachers additionally reported on their exposure to and use of Leo Leo progress reports in schools assigned to the combined family-and-school treatment.

In the endline evaluation, teachers also completed an adapted Strengths and Difficulties Questionnaire (SDQ) for each child in their classroom, which we use to measure children's socioemotional development. We focus on domains related to hyperactivity/inattention, peer relationships, and somatic complaints. These teacher-reported measures allow us to examine potential unintended consequences of increased technology use.

3.4.5 Family questionnaire

We asked families to complete a questionnaire at baseline to collect information on household characteristics. Surveys were administered remotely via an electronic questionnaire or telephone call. Because response rates were low in some schools, baseline family-reported data are supplemented with teacher-reported information on basic demographics.

At endline, families reported children's use of Leo Leo in Treatment 1 and Treatment 2, as well as information on home learning practices and children's media use, and completed a validated measure of problematic screen use. These data allow us to study whether the intervention affected home inputs and to assess potential trade-offs between EdTech use and children's broader media consumption.

3.5 Balance and descriptive statistics

Table 2 shows the descriptive statistics of schools, teachers, and children at baseline by treatment arm, as well as the outcome of the balance tests. This information comes from administrative records compiled by the Ministry of Education and from the family and teacher questionnaires. The table reports means and standard deviations for each group, along with differences between groups and their corresponding standard errors. The table includes the sample of schools that implemented the intervention.[§] Results are organized across five panels: school-level characteristics (Panel A), teacher characteristics (Panel B), child characteristics (Panel C), household background variables (Panel D), and municipality background variables (Panel E). Sample sizes vary across variables because responses to baseline questionnaires were incomplete; the balance tests use available observations.

Overall, baseline characteristics are well balanced across treatment arms (Table 2), and the few significant differences are consistent with what would be expected by chance given the large number of comparisons. The main exceptions are communicational difficulties among enrolled children, lower in Treatment 1 than in the control group (Panel A, significant at 10%); teachers' years of experience, 4.15 years higher in Treatment 1 than in Treatment 2 (Panel B, significant at 5%); the share of female children (4 percentage points higher in Treatment 1 than in control), and completion of the second year of preschool (12 percentage points higher in Treatment 1 than in Treatment 2) (Panel C, significant at 5% and 10%, respectively); and caregiver's years of education, 0.58 years higher in Treatment 1 than in Treatment 2 (Panel C, significant at 5%). Household and municipality background variables (Panels D and E) show no significant differences across arms.

To account for any residual imbalance, all subsequent impact estimates include randomization-strata fixed effects and the baseline covariates described in Section 3.7 below. Taken together, the balance results support the internal validity of the experimental design.

[§] In Table F1 of the Supplementary Material, we show descriptive statistics for the full sample of randomized schools. Baseline variables remain balanced in that sample.

Table 2: Descriptive statistics and balance

Variable	Control			Treatment 1			Treatment 2			Diff. Treatment 1 – Control			Diff. Treatment 2 – Control			Diff. Treatment 1 – Treatment 2		
	Mean	s.d.	N	Mean	s.d.	N	Mean	s.d.	N	Diff.	s.e.	N	Diff.	s.e.	N	Diff.	s.e.	N
Panel A: School																		
Number of children enrolled (total)	146.290	64.720	31	142.580	61.880	33	132.410	51.430	34	-0.980	11.400	64	-12.600	10.550	65	11.370	9.950	67
Number of children enrolled (5 years old)	45.840	23.170	31	43.940	21.220	33	42.680	19.130	34	-1.050	4.400	64	-2.740	4.130	65	1.670	3.700	67
Total enrolled with behavior issues	1.970	2.850	31	1.640	2.130	33	1.180	1.700	34	-0.330	0.640	64	-0.790	0.590	65	0.470	0.470	67
Total enrolled with communicational difficulties	3.680	4.680	31	1.940	3.030	33	2.090	3.330	34	-1.720*	1	64	-1.600	1.020	65	-0.140	0.780	67
Total enrolled with learning disabilities	1.390	2.920	31	2	3.330	33	2.150	4.520	34	0.650	0.790	64	0.780	0.950	65	-0.100	0.890	67
Number of third-grade groups	2.610	1.020	31	2.670	1.080	33	2.500	0.930	34	0.100	0.200	64	-0.090	0.190	65	0.190	0.190	67
Number of third-grade teachers	2.650	1.050	31	2.700	1.100	33	2.500	0.930	34	0.100	0.200	64	-0.130	0.190	65	0.220	0.190	67
Number of third-grade schoolrooms	2.580	0.960	31	2.670	1.080	33	2.440	0.890	34	0.130	0.200	64	-0.120	0.180	65	0.240	0.190	67
Double shift	0.160	0.370	31	0.150	0.360	33	0.060	0.240	34	-0.010	0.090	64	-0.100	0.080	65	0.090	0.080	67
Location = Monterrey	0.100	0.300	31	0.180	0.390	33	0.120	0.330	34	0.080	0.090	64	0.020	0.080	65	0.060	0.090	67
Panel B: Teacher																		
Female	0.980	0.120	65	0.990	0.110	76	1	0	72	0	0.020	141	0.020	0.010	137	-0.010	0.010	148
Weekly hours of work	26.640	10.160	65	24.560	8.360	70	24.390	7.750	67	-2.160	1.980	135	-2.150	1.720	132	0.170	1.520	137
Weekly hours of work at the studied center	22.620	6.140	62	21.170	4.960	72	21.790	3.970	68	-1.410	0.940	134	-0.840	0.990	130	-0.550	0.710	140
Years of experience as preschool teacher	14.460	9.030	60	16.420	7.840	74	12.360	7.640	71	2.090	1.940	134	-2.030	1.900	131	4.150**	1.610	145
Years of experience at the studied center	8.300	8.250	61	7.780	6.220	74	6.660	5.420	70	-0.470	1.450	135	-1.640	1.510	131	1.170	1.090	144
Years of education	21.200	2.020	61	20.910	2.450	76	20.720	2.060	71	-0.220	0.420	137	-0.470	0.370	132	0.210	0.430	147
Panel C: Child																		
Female	0.480	0.500	1789	0.510	0.500	1972	0.500	0.500	2030	0.040**	0.020	3761	0.030	0.020	3819	0.010	0.020	4002
Age	5.860	0.300	1755	5.870	0.290	1966	5.860	0.290	1981	0.010	0.010	3721	0	0.010	3736	0.010	0.010	3947
Completed second grade of preschool	0.800	0.400	1793	0.870	0.330	1972	0.750	0.430	2015	0.080	0.060	3765	-0.050	0.080	3808	0.120*	0.060	3987
Caregiver's years of education	13.420	2.650	1536	13.730	2.670	1762	13.160	2.460	1624	0.310	0.350	3298	-0.280	0.340	3160	0.580**	0.280	3386
Caregiver's age	32.690	6.490	1505	32.960	6.580	1643	32.500	6.530	1466	0.270	0.510	3148	-0.210	0.450	2971	0.470	0.410	3109
Panel D: Household																		
Number of books index	1.370	0.780	331	1.320	0.750	400	1.340	0.760	340	-0.050	0.070	731	0.000	0.080	671	-0.040	0.080	740
Frequency of parental involvement: reading	1.300	0.820	331	1.360	0.840	400	1.350	0.800	340	0.060	0.060	731	0.050	0.060	671	0.010	0.060	740
Frequency of parental involvement: games	2.760	0.920	331	2.730	0.890	400	2.760	0.880	340	-0.020	0.070	731	0.020	0.070	671	-0.040	0.080	740
Panel E: Municipality																		
Caregiver completed primary education at the municipality level	0.990	0.010	30	0.990	0.010	33	0.990	0.010	34	0.000	0.000	63	0.000	0.000	64	0.000	0.000	67
Literacy at the municipality level	0.990	0.000	30	0.990	0.000	33	0.990	0.000	34	0.000	0.000	63	0.000	0.000	64	0.000	0.000	67
Average years of schooling of caregiver at the municipality level	10.350	0.640	30	10.230	0.400	33	10.410	0.530	34	-0.130	0.140	63	0.050	0.150	64	-0.180	0.110	67

Notes: Difference estimates and standard errors come from regressions of each baseline characteristic on treatment indicators (one comparison at a time), including randomization-strata fixed effects. Standard errors are cluster-robust at the school level. The number-of-books index is based on parental reports of how many children's books are available at home, coded as none (0), 1 to 3 (1), 3 to 10 (2), and more than 10 (3). The frequency of parental involvement in reading and play is measured with a five-category scale: never (0), once a week (1), several times a week (2), almost every day (3), and every day (4). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

3.6 Attrition

The estimation sample for children's outcomes is restricted to children who completed both the Leo Test and IPAL, ensuring that estimates across the two primary literacy measures are directly comparable.[¶] Outcome completion rates range from approximately 58% to 64% across treatment groups. As shown in Appendix Figure B1 and Table 3, completion rates do not differ significantly between either treatment group and the control group or between the two treatment groups, suggesting that differential attrition is unlikely to drive the estimated treatment effects.

Table 3: Early literacy outcomes completion rates.

	Treatment 1 vs. Control	Treatment 2 vs. Control	Treatment 1 vs. Treatment 2
Treatment	0.015	0.029	-0.014
	(0.050)	(0.046)	(0.037)
Constant	0.584***	0.613***	0.641***
	(0.043)	(0.041)	(0.028)
N	3863	3907	4024

Notes: Dependent variable: indicator equal to one if the child completed both the Leo Test and IPAL assessments (1 = yes, 0 = no). Cluster-corrected standard errors are shown in parentheses. N= Number of observations. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Response rates for other outcomes for IDELA, SDQ, the family questionnaire, and the teacher questionnaire are not statistically significant across any of the comparisons (see Table G1 of the Supplementary Material). Overall, these findings indicate that response rates are balanced across experimental groups and are unlikely to pose a major threat to internal validity. Nevertheless, we examine post-attrition balance in Table G2 of the Supplementary Material. When comparing Treatment 1 and Treatment 2, both the probability of having attended preschool before enrollment and caregivers' years of education are higher in Treatment 1 than in Treatment 2, significant at the 5% and 10% levels, respectively. Teachers' years of experience are also higher in Treatment 1. No significant baseline differences are observed between either treatment group and the control group. These imbalances therefore arise only in the comparison between the two treatment groups after restricting the sample to children with endline assessments. The common set of baseline covariates included in every

[¶]The common-sample restriction applies to all analyses of children's outcomes, including outcomes drawn from the family questionnaire, such as problematic screen use (Table 7) and parental involvement (Table 10, Panel C). An exception is the teacher survey, for which responses to individual questions were optional. Teacher-survey outcomes are therefore estimated using the largest available sample for each item, and sample sizes vary accordingly.

regression (Section 3.7), together with strata and enumerator fixed effects, addresses these imbalances.

As an additional robustness check, we reestimate the main literacy specifications using stabilized inverse-probability weights that account for differential non-response to the joint Leo Test and IPAL assessment across treatment arms; the two-step weight-construction procedure (a LASSO-selected logit model for the denominator and a treatment-only logit for the numerator, trimmed at the 1st and 99th percentiles) is described in the notes to Table G5 of the Supplementary Material. Results from this exercise are discussed in Section 4.1.

Table G3 of the Supplementary Material reports baseline balance for the subsample of children who completed the endline family questionnaire. While response propensity does not differ significantly by treatment arm (Table G1 of the Supplementary Material), two baseline household variables—the frequency of parental reading and play with the child—are imbalanced within this subsample: Children in Treatment 2 whose families responded report less frequent reading at baseline relative to Control, and children in Treatment 1 whose families responded report less frequent play at baseline relative to Control. Because these variables are closely related to the parental-involvement outcomes, we control for them directly in that specification, in addition to the covariates used elsewhere in the paper.

3.7 Estimation approach

We estimate treatment effects on learning outcomes at the end of the school year, after approximately eight months of exposure to the program.

We use ordinary least squares to estimate ITT effects through pairwise comparisons across experimental arms. For each comparison, we estimate the model using only the two relevant groups. Specifically, we run three separate regressions: (i) Treatment 1 vs. Control, (ii) Treatment 2 vs. Control, and (iii) Treatment 1 vs. Treatment 2. The estimated equation is:

$$Y_{is} = \beta_0 + \beta_1\, PairTreat_s + \gamma' X_{is} + \varepsilon_{is} \tag{1}$$

Here, Y_{is} is the outcome of interest for child i in school s and $PairTreat_s$ is an indicator equal to 1 if school s belongs to the first group in the pairwise comparison and 0 if it belongs to the second group.

We adjust this specification for the following set of covariates contained in X_{is}: (i) randomization-stratum fixed effects (based on school size), (ii) enumerator fixed effects, and (iii) a common set of baseline covariates included in all specifications to improve precision and maintain a common specification across treatment contrasts.

In the estimates, we use cluster-robust standard errors. To account for multiple-hypothesis testing, we group conceptually related outcomes into pre-specified families and adjust p-values within each family using the List et al. (2019) procedure, controlling the family-wise error rate.

As a robustness check, we also estimate a pooled specification using the full sample, including indicators for Treatment 1 and Treatment 2 and using the control group as the omitted category. We then perform F-tests to assess whether the coefficients for the two treatment arms are statistically different from each other. Results are reported in Table G6 of the Supplementary Material and are consistent with the main pairwise comparison estimates.

4 Results

4.1 Main impacts on early literacy

Table 4 presents ITT estimates of the intervention's effects on early literacy. Odd columns report the ITT coefficient without covariates. Even columns report the ITT coefficient from specifications controlling for baseline unbalanced controls. Panel A reports outcomes from the Leo Test, and Panel B reports outcomes from the independently administered IPAL assessment. Across both instruments, we find a consistent pattern: Meaningful learning gains arise only when the home-based app is complemented by school support (Treatment 1), whereas providing the app to families alone (Treatment 2) does not generate robust improvements. This pattern remains after adjusting for multiple-hypothesis testing for the primary outcomes.

In the Leo Test (Panel A), children in Treatment 1 outperform the control group by 0.30 standard deviations on the overall score and by 0.21 standard deviations on the number of subtests completed. These effects are statistically significant at the 1% level and remain significant after adjusting for multiple-hypothesis testing. By contrast, Treatment 2 yields much smaller and statistically nonsignificant estimates: 0.08 standard deviations on the overall score and 0.04 standard deviations on the number of subtests.

The contrast between the two treatment arms is also economically meaningful. When comparing Treatment 1 to Treatment 2, the Leo Test score is higher in Treatment 1 by 0.23 standard deviations in the baseline-unadjusted specification and remains significant at conventional levels when controlling for baseline covariate imbalances between the two treatment arms. However, this difference is not robust to multiple-hypothesis-testing adjustments in specifications with controls. A similar pattern holds for the number of subtests completed: Treatment 1 exceeds Treatment 2 by 0.18 standard deviations, although statistical significance is sensitive to multiple-testing adjustments.

Panel B of Table 4 shows that these gains extend to the IPAL assessment, which measures foundational literacy skills outside the technological environment. Relative to the control group, Treatment 1 increases phonological awareness by 0.28 standard deviations, improves letter-name knowledge by 0.17 standard deviations, and improves letter-sound knowledge by 0.16 standard deviations. These effects are statistically significant at the 1%, 5%, and 10% levels. After adjusting for multiple-hypothesis testing, only the effect on phonological awareness remains statistically significant. In contrast, estimates for Treatment 2 are again smaller and statistically indistinguishable from zero across all IPAL outcomes. Direct comparisons between Treatment 1 and Treatment 2 indicate higher phonological awareness and letter-sound knowledge in Treatment 1 (though these differences are not robust to multiple-testing adjustments), consistent with the hypothesis that institutional support raises the effectiveness of the same home-delivered technology.

Table G6 of the Supplementary Material reports pooled estimates using the full sample, including indicators for Treatment 1 and Treatment 2 and using the control group as the omitted category. Overall, the results are consistent with the main pairwise comparison estimates. In most outcomes, the estimated effects remain similar in magnitude and statistical significance to those obtained in the pairwise specifications. The only outcomes that lose statistical significance relative to the main specification are letter-name and letter-sound knowledge, although the coefficients remain positive and similar in magnitude.

Table G5 of the Supplementary Material reports estimates reweighted using the stabilized inverse-probability weights described in Section 3.6. The results are very similar to our main findings: Point estimates are close in magnitude to the unweighted specifications in Table 4, and no coefficient changes statistical significance at conventional levels.

Table 4: Effects of the intervention on literacy (coefficients in standard deviations)

	Treatment 1 vs. Control		Treatment 2 vs. Control		Treatment 1 vs. Treatment 2	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Leo Test						
Overall score	0.301***	0.264***	0.078	0.125	0.226***	0.130*
	(0.091)	(0.081)	(0.104)	(0.081)	(0.074)	(0.074)
	[2139]	[2139]	[2087]	[2087]	[2182]	[2182]
	{0.020}	{0.020}			{0.012}	{0.210}
Number of subtests taken	0.209***	0.188***	0.037	0.066	0.176***	0.115*
	(0.070)	(0.068)	(0.075)	(0.062)	(0.055)	(0.059)
	[2139]	[2139]	[2087]	[2087]	[2182]	[2182]
	{0.033}	{0.052}			{0.015}	{0.191}
Panel B: IPAL Test						
Phonological awareness	0.278***	0.218***	0.110	0.099	0.148*	0.070
	(0.085)	(0.080)	(0.091)	(0.084)	(0.081)	(0.078)
	[1955]	[1955]	[1973]	[1973]	[2126]	[2126]
	{0.012}	{0.038}			{0.160}	
Letter-name knowledge	0.174**	0.136*	0.047	0.029	0.077	0.062
	(0.081)	(0.078)	(0.075)	(0.070)	(0.066)	(0.068)
	[2136]	[2136]	[2181]	[2181]	[2275]	[2275]
	{0.132}	{0.245}				
Letter-sound knowledge	0.155*	0.103	0.061	0.038	0.153**	0.108
	(0.086)	(0.084)	(0.082)	(0.078)	(0.068)	(0.066)
	[2131]	[2131]	[2183]	[2183]	[2274]	[2274]
	{0.146}				{0.138}	

Notes: Cluster-corrected standard errors are shown in parentheses. For significant coefficients, p-values adjusted for multiple-hypothesis testing are shown in braces. We define a primary-literacy family as consisting of (i) the Leo Test overall score and (ii) the IPAL phonological-awareness score. Additional literacy measures (Leo items completed, IPAL-subdomains letter-name knowledge, and letter-sound knowledge) are treated as secondary outcomes and adjusted within a separate family. Numbers of observations are reported in square brackets. Odd columns report the ITT coefficient without covariates. Even columns report the ITT coefficient from specifications controlling for baseline unbalanced controls: child's education, caregiver's education, and teacher experience. We impute the mean value of the variable in observations with missing values in covariates. All regressions include strata fixed effects and enumerator fixed effects. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

To better understand which skills drive improvements in the Leo Test score, Table 5 decomposes the assessment into its subtests. The results indicate that Treatment 1 primarily increases the probability that children reach more advanced components of the literacy sequence. Effects are concentrated in subtests that capture phonological awareness beyond basic letter recognition (initial and intermediate sound awareness) and early reading skills (word reading and reading comprehension). For instance, Treatment 1 increases completion of the intermediate-sound-awareness subtest by 0.10 standard deviations and the word-reading subtest by 0.11 standard deviations, both significant at the 1% level. We also find positive effects on reading comprehension (0.07 standard deviations). In contrast, there is no detectable effect on the earliest subtest related to basic letter-name and letter-sound recognition, suggesting that the intervention's main impact is not simply to raise minimal exposure, but to help children progress into more complex literacy tasks.

Table 5: Effects of the intervention on subtests of Leo Test (coefficients in standard deviations)

	Treatment 1 vs. Control		Treatment 2 vs. Control		Treatment 1 vs. Treatment 2	
	(1)	(2)	(3)	(4)	(5)	(6)
Letter-name and letter-sound recognition	0.003	0.003	-0.001	0.000	0.005	0.002
	(0.005)	(0.005)	(0.006)	(0.006)	(0.005)	(0.005)
	[2139]	[2139]	[2087]	[2087]	[2182]	[2182]
Initial sound awareness	0.035**	0.029*	0.010	0.017	0.029*	0.009
	(0.015)	(0.015)	(0.017)	(0.016)	(0.015)	(0.014)
	[2139]	[2139]	[2087]	[2087]	[2182]	[2182]
	{0.114}	{0.213}			{0.133}	{0.800}
Intermediate sound awareness	0.097***	0.090***	0.037	0.048*	0.067***	0.040
	(0.030)	(0.029)	(0.032)	(0.028)	(0.024)	(0.026)
	[2139]	[2139]	[2087]	[2087]	[2182]	[2182]
	{0.032}	{0.042}		{0.037}	{0.063}	
Word reading	0.106***	0.098***	0.042	0.054*	0.071***	0.043
	(0.032)	(0.031)	(0.035)	(0.030)	(0.026)	(0.028)
	[2139]	[2139]	[2087]	[2087]	[2182]	[2182]
	{0.026}	{0.045}		{0.351}	{0.072}	
Reading comprehension	0.074**	0.067**	-0.000	0.009	0.067***	0.053**
	(0.028)	(0.027)	(0.028)	(0.023)	(0.022)	(0.024)
	[2139]	[2139]	[2087]	[2087]	[2182]	[2182]
	{0.095}	{0.143}			{0.045}	{0.217}
Word writing	0.029	0.022	-0.012	-0.003	0.039**	0.028
	(0.025)	(0.025)	(0.026)	(0.020)	(0.017)	(0.017)
	[2139]	[2139]	[2087]	[2087]	[2182]	[2182]
					{0.128}	

Notes: Cluster-corrected standard errors are shown in parentheses. We define a single family as consisting of the six Leo Test subtests reported in this table, adjusted jointly using the List et al. (2019) procedure. For significant coefficients, p-values adjusted for multiple-hypothesis testing are shown in braces. Numbers of observations are reported in square brackets. Odd columns report the ITT coefficient without covariates. Even columns report the ITT coefficient from specifications controlling for baseline unbalanced controls: child's education, caregiver's education, and teacher experience. We impute the mean value of the variable in observations with missing values in covariates. All regressions include strata fixed effects and enumerator fixed effects. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

4.2 Impacts on non-targeted child outcomes

4.2.1 Numeracy

Table 6 reports the effects of the intervention on early numeracy skills measured through the IDELA assessment. Although numeracy is not explicitly targeted by the Leo Leo app, examining these outcomes allows us to assess whether improvements in literacy engagement spill over into broader cognitive skills.

Children in Treatment 1 perform significantly better than the control group in arithmetic operations, with an effect size of 0.32 standard deviations, statistically significant at the 1% level. However, this effect is not statistically significant after adjusting for multiple-hypothesis testing. This suggests that gains generated through the literacy-focused intervention may extend to related domains. By contrast, Treatment 2 shows smaller and statistically nonsignificant effects relative to the control group (0.11 standard deviations), reinforcing the broader result that access to the app alone is insufficient to generate measurable improvements in child outcomes.

Direct comparisons between the two treatment arms further confirm the importance of school involvement. Children in Treatment 1 outperform those in Treatment 2 by approximately 0.17 standard deviations in arithmetic operations, with estimates remaining stable across specifications with and without baseline controls. However, these differences are not robust to multiple-hypothesis-testing adjustments.

In contrast, we do not observe statistically significant effects on number identification. Coefficients are small and imprecisely estimated across all comparisons, indicating that the intervention does not affect basic number recognition. Taken together, these findings suggest that any spillover effects on numeracy should be interpreted with caution and are concentrated in more cognitively demanding tasks involving simple arithmetic rather than in foundational number-identification skills.

Table 6: Effects of the intervention on numeracy (coefficients in standard deviations)

	Treatment 1 vs. Control		Treatment 2 vs. Control		Treatment 1 vs. Treatment 2	
	(1)	(2)	(3)	(4)	(5)	(6)
Correct numbers identified	0.111	0.069	-0.004	-0.012	0.056	0.009
	(0.072)	(0.063)	(0.081)	(0.078)	(0.061)	(0.061)
	[2121]	[2121]	[2177]	[2177]	[2262]	[2262]
Correct arithmetic operations	0.321***	0.299***	0.107	0.100	0.169**	0.140**
	(0.120)	(0.119)	(0.077)	(0.078)	(0.068)	(0.069)
	[2098]	[2098]	[2149]	[2149]	[2223]	[2223]
	{0.273}	{0.372}			{0.139}	{0.217}

Notes: Cluster-corrected standard errors are shown in parentheses. We define a single family as consisting of the two IDELA outcomes reported in this table, adjusted jointly using the List et al. (2019) procedure. For significant coefficients, p-values adjusted for multiple-hypothesis testing are shown in braces. Numbers of observations are reported in square brackets. Odd columns report the ITT coefficient without covariates. Even columns report the ITT coefficient from specifications controlling for baseline unbalanced controls: child's education, caregiver's education, and teacher experience. We impute the mean value of the variable in observations with missing values in covariates. All regressions include strata fixed effects and enumerator fixed effects. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

4.2.2 Problematic screen use

Table 7 presents impacts on problematic screen use based on parental reports. Parent survey measures should be interpreted as suggestive because response rates are incomplete. The sample consists of children in the common Leo Test/IPAL sample that also completed the family questionnaire (25.6%, 28.4%, and 25.8% in Control, Treatment 1, and Treatment 2 respectively). Children in Treatment 1 show no statistically significant increases across any of the screen-use indicators. While point estimates are generally positive, they are smaller in magnitude and not statistically significant.

In Treatment 2, children show increases in several dimensions of problematic screen use. Parents report that screen entertainment becomes a stronger source of motivation (0.22 standard deviations), children think more frequently about screen use (0.19 standard deviations), and screen use interferes more with family activities (0.18 standard deviations), with effects statistically significant at conventional levels. We also observe a marginally significant increase in parents' reporting that children's desire to use screens is rising over time. However, none of these four effects remain statistically significant after adjusting for multiple-hypothesis testing.

The differences between Treatment 1 and Treatment 2 are not statistically significant, and coefficient magnitudes are broadly similar across the two treatment arms. Accordingly, the results are more consistent with a modest difference between the control group and both treatment arms, with somewhat larger and more precisely estimated effects in Treatment 2.

While the program does not appear to generate large increases in problematic media use, there is some suggestive evidence of increases in certain dimensions, particularly in Treatment 2, though these patterns should be interpreted with caution given the lack of robustness to multiple-hypothesis testing. Table 7 reports estimates that already control for the baseline imbalance in parental reading and play frequency identified among family questionnaire respondents (Table G3 of the Supplementary Material). Relative to a specification without these controls, coefficients change little in magnitude, though two of the four significant estimates in Treatment 2 fall from the 1% to the 5% level, and from the 5% to the 10% level, respectively.

Overall, these results highlight an important distinction between access and implementation. Providing children with digital learning tools at home is not neutral in terms of behavior; however, the evidence does not support strong differences across implementation models, and increased exposure may translate into less regulated screen use rather than more productive learning engagement.

Table 7: Effects of the intervention on problematic screen use (coefficients in standard deviations)

	Treatment 1 vs. Control	Treatment 2 vs. Control	Treatment 1 vs. Treatment 2
	(1)	(2)	(3)
It is hard for my child to stop using screens	0.043	0.031	-0.045
	(0.083)	(0.080)	(0.086)
	[579]	[562]	[617]
Screen entertainment is my child's unique motivation	0.110	0.209**	-0.099
	(0.084)	(0.082)	(0.090)
	[579]	[562]	[617]
		{0.153}	
My child is always thinking about screen entertainment	0.138	0.172**	-0.024
	(0.086)	(0.082)	(0.087)
	[579]	[562]	[617]
		{0.293}	
My child's screen use interferes with family activities	0.155	0.178*	-0.016
	(0.095)	(0.091)	(0.085)
	[579]	[562]	[617]
		{0.346}	
My child's screen use causes problems for the family	0.130	0.149	-0.032
	(0.091)	(0.095)	(0.089)
	[579]	[562]	[617]
My child gets frustrated when not allowed to use screens	0.043	0.004	-0.008
	(0.088)	(0.084)	(0.079)
	[579]	[562]	[617]
My child's desire to use screens is increasing	0.131	0.162*	-0.036
	(0.080)	(0.093)	(0.074)
	[579]	[562]	[617]
		{0.453}	
My child hides to use screens	-0.070	0.047	-0.083
	(0.081)	(0.077)	(0.071)
	[579]	[562]	[617]
My child turns to screens for comfort on bad days	0.046	0.049	-0.001
	(0.077)	(0.079)	(0.076)
	[579]	[562]	[617]

Notes: Outcomes are drawn from the family questionnaire, completed by approximately 26% to 27% of children in the common Leo Test/IPAL analytic sample (see Table G1 of the Supplementary Material and Section 3.6). All outcomes are standardized. We define a single family as consisting of the nine items reported in this table, adjusted jointly using the List et al. (2019) procedure. Cluster-corrected standard errors are shown in parentheses. Regressions control for the baseline frequency of parental reading to and playing with the child (imputing the mean for missing values), in addition to strata fixed effects; see Table G3 of the Supplementary Material for the baseline imbalance in these variables among family-questionnaire respondents, which motivates this control. Numbers of observations are reported in square brackets. ***p<0.01, **p<0.05, *p<0.10.

4.2.3 Socioemotional development

Table 8 reports the effects of the intervention on socioemotional outcomes measured through the SDQ instrument. Across all specifications, we find no evidence that Treatment 1 affects children's socioemotional development. Point estimates for somatic complaints, hyperactivity, and peer relations are small and statistically indistinguishable from zero, suggesting that increased engagement with the literacy program did not generate adverse behavioral or emotional consequences when school support was present.

Results differ somewhat for Treatment 2. While we do not detect significant impacts on somatic complaints or peer relations, we observe a positive and marginally significant effect on hyperactivity (approximately 0.18 standard deviations, significant at the 10% level). However, this effect is not statistically significant after adjusting for multiple-hypothesis testing. Although the estimate is modest and should be interpreted cautiously, it suggests that children who accessed the app without school involvement may have experienced changes in behavior. Direct comparisons between Treatment 1 and Treatment 2 show negative but statistically nonsignificant coefficients, indicating that socioemotional differences across treatments are suggestive rather than definitive.

Overall, the SDQ results indicate that the intervention does not generate broad socioemotional effects. The absence of negative impacts in Treatment 1 is particularly relevant given concerns that increased exposure to digital learning tools at early ages could displace social or classroom interactions. At the same time, the suggestive increase in hyperactivity in the family-only arm is consistent with the idea that institutional structure may help regulate how children engage with digital tools, though this pattern is not robust to multiple-hypothesis-testing adjustments.

Table 8: Results SDQ Test (coefficients in standard deviations)

	Treatment 1 vs. Control		Treatment 2 vs. Control		Treatment 1 vs. Treatment 2	
	(1)	(2)	(3)	(4)	(5)	(6)
Somatic complaints	0.054	0.060	0.009	0.005	0.038	0.048
	(0.095)	(0.096)	(0.068)	(0.067)	(0.087)	(0.087)
	[2066]	[2066]	[2084]	[2084]	[2170]	[2170]
Hyperactivity	0.109	0.106	0.179*	0.216**	-0.065	-0.091
	(0.092)	(0.085)	(0.101)	(0.105)	(0.083)	(0.095)
	[2105]	[2105]	[2106]	[2106]	[2187]	[2187]
			{0.278}	{0.192}		
Peer relations	-0.169	-0.151	-0.104	-0.125	-0.054	-0.018
	(0.174)	(0.156)	(0.132)	(0.138)	(0.154)	(0.136)
	[2076]	[2076]	[2090]	[2090]	[2174]	[2174]

Notes: All outcomes are standardized. Cluster-corrected standard errors are shown in parentheses. We define a single family as consisting of the three SDQ domains reported in this table, adjusted jointly using the List et al. (2019) procedure. For significant coefficients, p-values adjusted for multiple-hypothesis testing are shown in braces. Numbers of observations are reported in square brackets. Odd columns report the ITT coefficient without covariates. Even columns report the ITT coefficient from specifications controlling for baseline unbalanced controls: child's education, caregiver's education, and teacher experience. We impute the mean value of the variable in observations with missing values in covariates. All regressions include strata fixed effects and enumerator fixed effects. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

4.3 Heterogeneity of treatment effects

Our theory of change suggests that the app provides more benefit to more disadvantaged students, as it is personalized to each child's learning level and needs. Therefore, in this section we assess heterogeneous effects along socioeconomic disadvantage, proxied by caregiver education. Results are presented in Table 9. Each pair of columns shows the coefficients and standard errors from an OLS regression of the outcome in each row on the ITT main effect, on the interaction between ITT status and the variable capturing the mechanism, on the main effect of the mechanism, and on randomization strata. The table displays only the ITT main effect and the interaction of ITT with the mechanism.

Across outcomes, we find some evidence that the intervention generates larger gains for children from more disadvantaged households, defined as those with lower levels of maternal education. Although not all interaction coefficients are statistically significant, the pattern of estimates suggests that children from less educated families may benefit more from the program. This pattern is consistent with the program's adaptive design, which tailors content to each child's skill level and may be particularly valuable for children facing greater learning constraints at home.

Panel A (Leo Test) shows that for the comparison of Treatment 1 versus Control, the interaction with caregiver education is positive and statistically significant for the number of subtests completed, though not for the overall score; comparing Treatment 1 directly with Treatment 2 shows no clear evidence of disproportionate benefit for disadvantaged students on this instrument. Panel B (IPAL) points in the same direction, particularly for letter-name knowledge in the Treatment 1-versus-Control comparison, although the Treatment 1-versus-Treatment 2 interaction for phonological awareness is negative. Panel C (IDELA numeracy) shows a similar pattern, with a significant interaction for number identification in the Treatment 1-versus-Control comparison.

Table 9: Heterogeneous effects by caregiver education (coefficients in standard deviations)

	Treatment 1 vs. Control		Treatment 2 vs. Control		Treatment 1 vs. Treatment 2	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Leo Test						
Leo score	0.200*	0.155	0.107	0.053	0.096	0.096
	(0.116)	(0.125)	(0.121)	(0.126)	(0.087)	(0.098)
	[1991]	[1991]	[1849]	[1849]	[1974]	[1974]
Number of subtests completed	0.117	0.181*	0.060	0.075	0.063	0.102
	(0.089)	(0.102)	(0.094)	(0.103)	(0.076)	(0.093)
	[1991]	[1991]	[1849]	[1849]	[1974]	[1974]
Panel B: IPAL						
Phonological awareness	0.217**	0.036	0.031	0.169	0.149	-0.186**
	(0.104)	(0.104)	(0.108)	(0.108)	(0.104)	(0.083)
	[1814]	[1814]	[1758]	[1758]	[1940]	[1940]
Letter-name knowledge	0.053	0.199*	0.040	0.058	0.010	0.087
	(0.107)	(0.106)	(0.096)	(0.095)	(0.087)	(0.085)
	[1988]	[1988]	[1945]	[1945]	[2069]	[2069]
Letter-sound knowledge	0.029	0.181	0.011	0.083	0.093	0.035
	(0.117)	(0.117)	(0.103)	(0.103)	(0.090)	(0.089)
	[1983]	[1983]	[1945]	[1945]	[2066]	[2066]
Panel C: IDELA						
Correct numbers identified	0.011	0.188*	0.000	0.058	-0.027	0.062
	(0.085)	(0.103)	(0.107)	(0.095)	(0.079)	(0.095)
	[1974]	[1974]	[1940]	[1940]	[2054]	[2054]
Correct arithmetic operations	0.295*	0.028	0.022	0.108	0.206**	-0.093
	(0.151)	(0.125)	(0.092)	(0.095)	(0.090)	(0.104)
	[1951]	[1951]	[1916]	[1916]	[2019]	[2019]

Notes: Cluster-corrected standard errors are shown in parentheses. Numbers of observations are reported in square brackets. Odd columns report the ITT coefficient (main effect). Even columns report the interaction between the ITT and a dummy equal to one if the caregiver has 12 years of education or less. All regressions include strata fixed effects and enumerator fixed effects. Regressions control for child's education, caregiver's education, and teacher experience. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Overall, the analysis suggests that the intervention may generate larger gains for children from less educated households. While the interaction effects are not always precisely estimated, the pattern across outcomes is consistent with the possibility that adaptive digital instruction can be especially beneficial for children facing greater socioeconomic disadvantage. This directional pattern echoes the subgroup finding in Bresciani et al. (2026), even though, as discussed in the introduction, our own family-only arm does not reproduce their result of significant gains among less educated households.

4.4 Mechanisms: How did the program work?

To understand why the combined family-and-school intervention generated larger learning gains than the family-only arm, we next examine potential mechanisms. Specifically, we study whether the differential impacts across treatment arms are associated with differences in children's engagement with the app, changes in parental involvement at home, or shifts in teachers' practices and expectations. We view this as a descriptive decomposition rather than a causally identified mediation analysis: App engagement, parental involvement, and teacher practices are themselves outcomes of the treatment, not independently randomized moderators, so differences across arms in these intermediate measures are suggestive of a channel rather than proof of one. Evidence comes from administrative app-usage logs, family surveys, and teacher questionnaires.

Table 10 presents results on engagement with the Leo Leo app. Panel A shows differences in usage patterns between treatment arms.[‖] Relative to Treatment 2, children in Treatment 1 make significantly greater progress within the app and are less likely to exhibit zero engagement by the end of the intervention period. In addition, children in Treatment 1 use the app more frequently and spend more time per session. These results suggest that the added school component substantially increases adherence and intensity of use, which is consistent with the idea that institutional support helps sustain engagement with the technology over time. Given that both treatment arms receive access to the same app, these differences highlight the role of schools in transforming access into sustained use.

Panel B examines total screen use based on parental reports. While overall screen time increases for children in Treatment 2 relative to the control group, no statistically significant increase is observed in Treatment 1. This pattern is consistent with the interpretation that school involvement may help channel screen exposure toward more structured and educational use, rather than increasing total screen consumption.

[‖] Sample sizes for app-engagement outcomes are smaller than in the full analysis sample because usage data are not available for all children. Missing data of any kind typically reflect loss of interest from the family, loss of interest from the child, or problems with the device, including loss, damage, or reduced availability. To assess whether missing engagement data differ systematically across treatment arms, we regress an indicator for observing engagement data on treatment assignment, restricting the sample to Treatment 1 and Treatment 2. The coefficient is small and not statistically significant (Table G1 of the Supplementary Material), indicating that differences in reported engagement between Treatment 1 and Treatment 2 are unlikely to reflect differential data completeness rather than differences in actual app use. Beyond the response rate itself, Table G8 of the Supplementary Material further shows that among children with observed engagement data, baseline child and household characteristics are balanced between Treatment 1 and Treatment 2, with no statistically significant differences in any of the eight variables examined. This indicates that the composition of children with engagement data, and not only its incidence, is comparable across arms, reinforcing the conclusion that the engagement results in Table 10 are unlikely to be driven by differential attrition.

Panel C explores whether differences in outcomes are explained by greater parental involvement in literacy activities. Across measures such as reading with children, playing literacy-related games, or helping with writing, we find no systematic differences between Treatment 1 and Treatment 2. In a specification without baseline controls, families in Treatment 1 report a marginally lower frequency of reading books relative to Control; once we control for the baseline imbalance in reading and play frequency among family-questionnaire respondents (Table G3 of the Supplementary Material; see Section 3.6), this difference disappears (Table 10, Panel C). These findings are suggestive rather than conclusive, given the lower response rate of the family questionnaire in comparison to the main sample, but they do not point to additional parental effort as the channel behind the combined treatment's effects.

Table 10: Progress in Leo Leo app at the end of the intervention

	Treatment 1 vs. Control	Treatment 2 vs. Control	Treatment 1 vs. Treatment 2
Panel A: Use of the Leo Leo app			
Progress until May 29			0.483***
			(0.091)
			[1231]
			{0.000}
Students with zero engagement by May 29			-0.039**
			(0.015)
			[1761]
			{0.038}
Frequency of Leo Leo app use			0.257**
			(0.000)
			[483]
			{0.044}
Average duration of Leo Leo app sessions			0.186**
			(0.003)
			[483]
			{0.025}
Panel B: Overall screen use			
Total use of screens	0.123	0.189***	-0.058
	(0.080)	(0.069)	(0.075)
	[579]	[562]	[617]
		{0.005}	
Panel C: Frequency of parental involvement			
Frequency of reading books with child	-0.142	-0.034	-0.082
	(0.087)	(0.075)	(0.091)
	[579]	[562]	[617]
Frequency of playing writing games (letters/name)	-0.011	0.012	-0.036
	(0.078)	(0.081)	(0.065)
	[579]	[562]	[617]
Frequency of teaching child to write words	-0.004	0.036	-0.014
	(0.080)	(0.086)	(0.086)
	[579]	[562]	[617]

Notes: The first outcome corresponds to cumulative standardized progress in the Leo Leo app up to May 29. The second outcome is an indicator equal to one if cumulative progress by May 29 was below 5%. Panel A reports only the comparison of Treatment 1 versus Treatment 2 since app log data are not available for the control group, which never received access to the app. The first two outcomes in Panel A (progress and zero engagement) are drawn from Leo Leo administrative log data; the last two outcomes (frequency of use and average session duration) are drawn from the endline family questionnaire and therefore share the same lower response rate as Panels B and C. Panel B and Panel C outcomes are drawn from the family questionnaire, completed by approximately 26% to 27% of children in the common analytic sample (see Table G1 of the Supplementary Material and Section 3.6). Panel C regressions additionally control for the baseline frequency of parental reading to and playing with the child (imputing the mean for missing values), given the baseline imbalance in these variables among family-questionnaire respondents documented in Table G3 of the Supplementary Material. All regressions include strata fixed effects. The four outcomes in Panel A form a single family, the outcome in Panel B is adjusted as its own family, and the three outcomes in Panel C form a separate single family, each adjusted jointly using the List et al. (2019) procedure. Sample sizes vary across engagement outcomes because usage data were not reported for all children, either because the app was never activated (in turn because of limited device access, limited family interest, or a child's diagnosis) or because play was discontinued after activation (in turn because of lack of family or child interest or because of device problems). Regressions in Panel A include a control for the date of the baseline evaluation. ***p<0.01, **p<0.05, *p<0.10.

We next examine whether school involvement influenced teachers' practices, perceptions, and integration of Leo Leo. Table 11 reports the main results from the endline teacher survey. Additional outcomes are presented in Table G7 of the Supplementary Material. Although estimates are imprecise because of the small sample size, several patterns emerge. Teachers in Treatment 1 report a greater tendency to group children according to language comprehension levels and display more positive assessments of families' role in language learning.

They also report higher expectations for students' educational trajectories, including university attendance and completion, relative to teachers in Treatment 2. While these measures are self-reported and should be interpreted cautiously, they suggest that providing teachers with structured information on student progress may have encouraged closer monitoring and strengthened coordination between school and home environments.

Additional results indicate that school involvement also influenced the use and integration of Leo Leo. Relative to Treatment 2, teachers in Treatment 1 report significantly greater classroom use of Leo Leo and increased use of the app outside the classroom. They also report stronger support from school leadership for the use of supplementary educational tools. While differences in reminders to students and families for home app use are positive but not statistically significant, these findings suggest that institutional engagement facilitated the incorporation of the technology into both instructional practices and school routines.

Table 11: Teacher-questionnaire results (coefficients in standard deviations)

	Treatment 1 vs. Control	Treatment 2 vs. Control	Treatment 1 vs. Treatment 2
	(1)	(2)	(3)
Panel A: Teacher Practices and Expectations			
Groups children by language comprehension level (frequency)	0.494	0.173	0.379*
	(0.309)	(0.346)	(0.217)
	[34]	[36]	[46]
Expects students to start university (percentage)	0.361	-0.117	0.566*
	(0.335)	(0.340)	(0.291)
	[34]	[36]	[46]
Evaluates the role of families in children's language learning	0.192	-0.135	0.463**
	(0.279)	(0.331)	(0.222)
	[34]	[36]	[46]
Considers parental help at home frequent (perception)	0.326	0.010	0.392*
	(0.307)	(0.368)	(0.228)
	[34]	[36]	[46]
Panel B: Use and Integration of Leo Leo			
Frequency of reminders to students for home app use			0.295
			(0.214)
			[74]
Frequency of reminders to families for home app use			0.260
			(0.233)
			[74]
Classroom time using Leo Leo			0.517**
			(0.232)
			[74]
Use of Leo Leo outside the classroom			0.430*
			(0.220)
			[74]
Use of other literacy apps			0.085
			(0.216)
			[74]
School leadership support for the use of supplementary tools such as Leo Leo			0.574***
			(0.200)
			[73]

Notes: Panel B reports only the comparison of Treatment 1 versus Treatment 2 since these outcomes describe the integration of Leo Leo into classroom practice and therefore do not apply to teachers in schools that never received the app. Cluster-corrected standard errors are shown in parentheses. Numbers of observations are reported in square brackets. The teacher survey consisted of a single questionnaire in which responses to individual questions were optional. As a result, the number of observations varies across outcomes because some teachers completed only a subset of the questions. Overall survey participation was similar across treatment groups. All regressions include strata fixed effects. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

5 Conclusion

This paper evaluates the impact of Leo Leo, an adaptive early literacy app, using a randomized controlled trial in 105 public preschools in Nuevo León, Mexico. Schools were randomly assigned to a control group, a family-only intervention, or a combined family-and-school intervention that provided teachers with structured support and feedback.

Our central finding is that institutional involvement substantially enhances the effectiveness of personalized digital instruction. The combined family-and-school treatment generated sizable and robust gains in early literacy. We also find suggestive, though less statistically robust, positive spillovers to early numeracy, concentrated in arithmetic operations rather than number identification, while the family-only intervention produced no robust improvements relative to the control group. The strongest evidence for the incremental role of schools comes from the digital literacy assessment, with supporting evidence from independently administered measures of foundational literacy skills, particularly phonological awareness. Holding technology constant, adding school-based support substantially increased impact.

Importantly, Leo Leo was delivered through the home and designed for use outside the classroom, but it was not institutionally detached from schools. Both treatment arms received the same household access to the technology; the key difference was whether schools were given information and tools to support its use. Our results therefore suggest that home delivery may help overcome some of the implementation constraints associated with school-based EdTech but does not render schools unimportant. Rather, schools appear to play a critical role in structuring, legitimizing, and reinforcing the use of EdTech. The effects identified in this study are substantial when framed within the existing literature. For context, the median effect size is typically around 0.1 standard deviations (Evans and Yuan, 2022).

We also find little evidence that the learning gains generated by the combined family-and-school intervention were accompanied by adverse socioemotional effects. An additional point raised by these results concerns children's screen exposure, a topic that has become central to discussions of EdTech and child development. Our findings suggest that the relevant question is not whether young children should use screens, but what kind of screen experience they are given. Leo Leo was designed as a structured, time-limited, and pedagogically focused tool rather than an open-ended entertainment platform, nor as a substitute for parents, teachers, or books. Sessions close automatically after 20 minutes, and progression through the app is tied to explicit literacy content rather than free browsing. In this sense, the design itself builds in a form of parental control since limits on use do not depend on daily negotiation between caregivers and children but are embedded directly in the technology. Consistent with this interpretation, we find no increase in total screen time in Treatment 1 relative to the control group, even though this arm showed the largest gains in literacy and the highest engagement with the app. By contrast, screen time rose significantly in Treatment 2, the arm that also shows suggestive increases in hyperactivity and problematic use, consistent with weaker regulation when engagement relies solely on household initiative. This pattern is consistent with substitution rather than addition: When school support is present, children appear to replace less structured or lower-quality screen time with time spent on a curated learning tool, rather than simply accumulating more screen time overall. More broadly, our findings suggest that discussions about EdTech should focus not on the quantity of screen exposure, but on the quality, structure, and purpose of children's digital experiences. As digital learning tools become increasingly embedded in education systems, understanding how to design and integrate technology in ways that channel screen time toward productive uses will become increasingly important.

Descriptive evidence helps explain why the combined intervention was more effective. Relative to the family-only treatment, children in the combined family-and-school treatment used the app more frequently, spent more time using it, progressed further through the curriculum, and were less likely to remain inactive. By contrast, we find little evidence that the intervention changed parental literacy practices at home, based on a household survey with a lower response rate than our main analytic sample; we interpret this specific result as suggestive. We also find suggestive evidence that teachers in the combined treatment were more likely to monitor student progress, encourage use of the platform, and incorporate information from Leo Leo into their educational practices. Finally, impacts were larger among children from more disadvantaged households, indicating that institutional support may be particularly valuable where household constraints are greater.

These findings speak directly to a central tension in the EdTech literature. Home-based delivery relaxes scalability constraints: The software license costs approximately \$1.25 per child per year, for a total first-cohort cost of about \$17.50 per child in Treatment 1,[**] the arm that generated learning gains, a favorable figure given that most of these costs (software licensing, teacher training, coordination) are low or fixed and would fall further as the program scales to more schools and cohorts. Technology delivered outside the classroom, however, does not eliminate the need for institutions: Implementation here relied on routine coordination with schools and education authorities, as well as integration into existing public-sector structures, suggesting that the institutional supports that make the technology effective can themselves be embedded into existing public-sector delivery structures.

This cost places our evidence within the broader landscape of education cost-effectiveness evidence. The average effect size for computer-assisted learning interventions in McEwan's (2015) meta-analysis of developing-country randomized controlled trials was 0.15 standard

[**] One-time implementation costs totaled \$49,120: teacher and principal training (\$16,000), program coordination (\$31,000), and satellite internet devices for launch events (\$2,120). Teacher and principal training was delivered only in Treatment 1 schools; dividing this cost by the 1,992 children in Treatment 1 (Table 1) yields \$8.03 per child. Program coordination and satellite devices applied to both treatment arms; dividing this \$33,120 by the 4,031 children in Treatment 1 and Treatment 2 (Table 1) yields \$8.22 per child. Summing these components, one-time implementation costs amount to approximately \$16.25 per child in Treatment 1 and \$8.22 per child in Treatment 2.

deviations, roughly half the effect size we estimate for Treatment 1. Angrist et al. (2025) propose the learning-adjusted years of schooling (LAYS) framework to compare interventions on a common metric that combines effect size and cost; they find that some of the most cost-effective interventions in their review of over 200 studies deliver the equivalent of more than three years of high-quality schooling for as little as $100 per child. Because LAYS combines cost and learning gains into a single measure, and because this figure reflects the top tier of a wide distribution of interventions rather than a general benchmark, a direct comparison requires expressing our effect size in the same metric rather than comparing per-child costs alone, a conversion we leave for future work. With that caveat in mind, our estimated first-cohort cost of approximately $17.50 per child is modest by comparison, though we cannot yet say how Leo Leo would rank on a LAYS basis once both the size and duration of its effects are taken into account.

Our findings are consistent with a growing literature suggesting that the effectiveness of EdTech depends not only on the technology itself, but also on the institutional environments in which it is embedded (Ma et al., 2024). Our evidence suggests that light-touch school involvement can sustain engagement and transform access to technology into meaningful learning gains even when adaptive technology is delivered primarily through the home rather than the classroom. The present design does not allow us to determine whether these institutional effects operate through accountability, active teacher follow-up, better coordination between home and school, or the combination suggested by our evidence. Identifying the minimal institutional inputs required to sustain engagement remains an important question for future research. Nonetheless, our findings show that personalized digital tools delivered at home can generate meaningful gains at scale, but only when embedded within institutional structures that support consistent participation. Future research should isolate which institutional inputs are sufficient to sustain engagement and improve learning.

This study evaluates a program designed and implemented with scalability in mind. The intervention reaches thousands of children from a representative population and is delivered in ordinary public schools under standard conditions. We did not select elite institutions or provide intensive support beyond what is feasible within existing systems. By embedding the program within normal school routines, the evaluation reflects real-world implementation constraints and enhances the external validity of the findings. This design aligns with the "Option C" approach proposed by List (2026), which emphasizes incorporating scalability constraints into the intervention architecture from the outset.

Our findings contribute to a growing body of large-scale evidence on EdTech in public systems. As emphasized by Muralidharan and Singh (2025), evaluating interventions under real-world constraints is essential to understanding their effectiveness. We show that EdTech delivered at home can be both scalable and effective, but only when complemented by institutional support: Teacher engagement sustains usage and facilitates effective implementation, even when delivery occurs at home.

A Evaluation Scales

Leo Test

The Leo evaluation consists of nine subtests. These subtests are designed to evaluate different areas of early literacy. Test 1 measures verbal comprehension, while test 2 and 3 measure alphabetic knowledge. Tests 4 and 5 evaluate phonological knowledge, specifically initial and middle sound recognition. Test 6, 7 and 8 evaluate reading (including words and pseudowords reading). Finally, test 9 measures writing capability. Children must complete test i to access test $i + 1$. To evaluate the treatment effect, we constructed eleven variables based on the student's performance. Nine of these are dummy indicators, each corresponding to the completion of one of the tests. For example, the dummy for test 2 is equal to one if the child completed test 2 at endline (and, by construction, all previous tests); similarly, the dummy for test 3 equals one if the child completed test 3 at endline, and the same defined through test 9. If one child completed test 1 but not test 2 at endline, the dummy for test 2 is equal to zero. Because of the design of the Leo Test, all children who completed test 2 also completed test 3, and the same pattern for tests 7 and 8. Therefore, we consider the dummy of test 3, but not of test 2, and the dummy of test 8, but not of test 7. Because all the children in the sample evaluated answered test 1, we do not consider this variable as an outcome (the indicator equals one for all observations).

In addition, in order to summarize overall performance, we created two more variables. The first is the number of subtests completed, which is a variable in the range of 1 to 9. This variable captures the highest stage reached by a child in the Leo Leo sequence. The second variable is the Leo score, a weighted index of outcomes that assigns the following weights: 5% to test 1, 15% to test 2, 3 and 4, and 10% to tests from 5 to 9. The number of subtests completed, and the weighted index are useful to measure the general results.

IPAL Test

We also constructed additional outcomes based on IPAL (Alphabetic Knowledge and Phonological Awareness) results. This test is divided in three sections: phonological awareness, knowledge of letter names, and knowledge of letter sounds. Specifically, in the study we consider from IPAL the following outcomes: 1) correct answers in phonological awareness, 2) correct answers in letter name knowledge, and 3) correct answers in letter sound knowledge.

IDELA Test

We considered items related to number identification and mathematical operations. Specifically, we considered the number of correct numbers identified, and the number of correct arithmetic operations.

SDQ Test

We focus on three domains: hyperactivity/inattention, peer relationships, and somatic complaints. Each item is rated by the child's teacher, who reports on the child's behavior in the classroom, using a three-point scale (1 = "not true," 2 = "somewhat true," 3 = "certainly true").

The hyperactivity/inattention subscale captures children's ability to regulate their activity level and attention. Its five items reflect tendencies such as restlessness, distractibility, and impulsivity, as well as the capacity to sustain concentration and complete tasks. From these items, we created two measures: a total score, obtained by summing across the five items (range 5–15), and a mean score, defined as the average across items (range 1–3).

The peer relationships subscale captures children's social integration and the quality of their interactions with other children. It also consists of five items, which assess whether the child tends to be solitary, has at least one close friend, is generally liked by peers, is bullied, and gets along better with adults than with other children. As with hyperactivity, we created both a total score and a mean score, with higher values indicating more positive peer relationships.

Finally, we included a single item that captures somatic complaints, asking whether the child often suffers from headaches, stomach - aches, or nausea. This item is rated on the same three-point scale, with higher values reflecting more frequent somatic complaints.

Teacher questionnaire

Teachers are being surveyed by phone to collect information on their teaching practices and use of the intervention's feedback reports.

Family evaluation

Families are being surveyed by phone to collect data on reading habits using three items from the STIM-Q READ scale and on children's screen use through the Problematic Media Use Measure – Short Form (PMUM-SF), based on parental reports. In addition, we collected information on children's use of the Leo Leo app.

- STIM-Q: We use selected items from STIM-Q, an instrument designed to measure the home learning environment. Parents report the number of days per week, in which they engage in literacy-related activities with their child, such as reading books, practicing letters or words. Higher values in the total score (range 0-21) and mean score (range 0-7) indicate in this case more frequent stimulation.

- PMUM-SF: As part of this instrument, we include a set of nine items specifically focused on problematic screen use, covering difficulties in limiting screen time, interference with family life, frustration when screens are unavailable, reliance on screens for comfort, or secrecy about its use. Each item is rated on a five-point scale (1="never", 5="always"). As with the SDQ subscales, we constructed a total score (range 9-45) and a mean score (range 1-5), with higher values indicating more problematic patterns of screen use.

- Leo Leo use: Parents report the number of days per week the child used the app (range 0-7), and the average duration of each session (range 0-20 minutes). These measures are only being collected for those children in the treatment arms and are therefore used exclusively for comparisons across treatment groups.

B Sample Construction, Outcome Completion, and Baseline Balance

Figure B1 summarizes the construction of the analysis sample from school randomization through the main analysis sample.

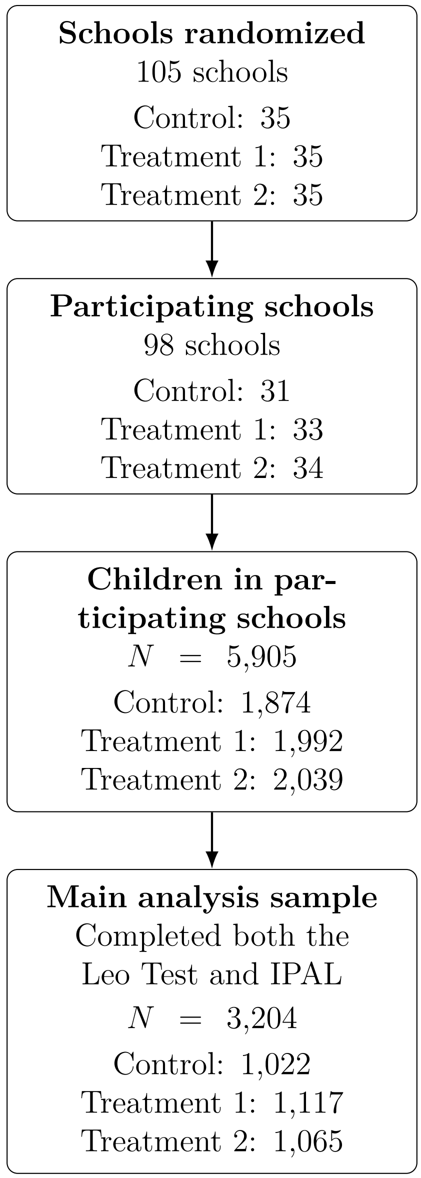


Figure B1: Sample construction

Notes: Seven schools declined participation after random assignment but before the intervention began. The intervention was implemented in the remaining 98 schools. The main analysis sample consists of children who completed both the Leo Test and IPAL.

References

Abrahamsson, S. (2026). Smartphone bans, student outcomes and mental health. *Journal of Human Resources*.

Allcott, H., Braghieri, L., Eichmeyer, S., and Gentzkow, M. (2020). The welfare effects of social media. *American economic review*, 110(3):629–676.

Alvarez Marinelli, H., Boggild-Jones, I., Crawford, M., Dubeck, M. M., Jhingran, D., Lack, C. J., Mohohlwane, N., Oviedo Buitrago, M. E., Piper, B., Saavedra, J., et al. (2025). Effective reading instruction in low-and middle-income countries: What the evidence shows.

Angrist, N., Evans, D. K., Filmer, D., Glennerster, R., Rogers, F. H., and Sabarwal, S. (2025). How to improve education outcomes most efficiently? a review of the evidence using a unified metric. *Journal of Development Economics*, 172.

Arnold, D. H., Chary, M., Gair, S. L., Helm, A. F., Herman, R., Kang, S., and Lokhandwala, S. (2021). A randomized controlled trial of an educational app to improve preschoolers emergent literacy skills. *Journal of Children and Media*, 15(4):457–475.

Balsa, A., Boo, F. L., Bloomfield, J., Cristia, A., Cid, A., de La Paz Ferro, M., Valdés, R., and del Luján González, M. (2023). Effect of crianza positiva e-messaging program on adult–child language interactions. *Behavioural Public Policy*, 7(3):607–643.

Balsa, A. I., Gómez Muzzio, E., González, M. L., Bloomfield, J., Cid, A., and Valdés, R. (2024). Crianza positiva: combining group workshops and e-messages to strengthen parenting competences. In *Child & Youth Care Forum*, volume 53, pages 719–740. Springer.

Banerjee, A. V., Cole, S., Duflo, E., and Linden, L. (2007). Remedying education: Evidence from two randomized experiments in india. *The Quarterly Journal of Economics*, 122(3):1235–1264.

Beg, S., Halim, W., Lucas, A. M., and Saif, U. (2022). Engaging teachers with technology increased achievement, bypassing teachers did not. *American Economic Journal: Economic Policy*, 14(2):61–90.

Bettinger, E., Fairlie, R., Kapuza, A., Kardanova, E., Loyalka, P., and Zakharov, A. (2023). Diminishing marginal returns to computer-assisted learning. *Journal of Policy Analysis and Management*, 42(2):552–570.

Beuermann, D. W., Cristia, J., Cueto, S., Malamud, O., and Cruz-Aguayo, Y. (2015). One laptop per child at home: Short-term impacts from a randomized experiment in peru. *American Economic Journal: Applied Economics*, 7(2):53–80.

Bloomfield, J., Balsa, A., and Cid, A. (2023). Using behavioral insights in early childhood interventions: The effects of crianza positiva e-messaging program on parental investment. *Review of Economics of the Household*, 21(1):95–130.

Braghieri, L., Levy, R., and Makarin, A. (2022). Social media and mental health. *American Economic Review*, 112(11):3660–3693.

Bresciani, D., Kalil, A., Liu, H., Mayer, S. E., and Shah, R. (2026). Closing early math gaps by parental education with technology at home. *Economics of Education Review*, 111:102774.

Bulman, G. and Fairlie, R. W. (2016). Technology and education: Computers, software, and the internet. In *Handbook of the Economics of Education*, volume 5, pages 239–280. Elsevier.

Cristia, J., Ibarrarán, P., Cueto, S., Santiago, A., and Severín, E. (2017). Technology and child development: Evidence from the one laptop per child program. *American Economic Journal: Applied Economics*, 9(3):295–320.

Doss, C. J., Fahle, E. M., Loeb, S., and York, B. N. (2019). More than just a nudge: Supporting kindergarten parents with differentiated and personalized text messages. *Journal of Human Resources*, 54(3):567–603.

Escueta, M., Quan, V., Nickow, A. J., and Oreopoulos, P. (2020). Education technology: An evidence-based review. *Review of Educational Research*, 90(3):479–529.

Evans, D. K. and Yuan, F. (2022). How big are effect sizes in international education studies? *Educational Evaluation and Policy Analysis*, 44(3):532–540.

Glewwe, P. and Muralidharan, K. (2016). Improving education outcomes in developing countries: Evidence, knowledge gaps, and policy implications. In *Handbook of the Economics of Education*, volume 5, pages 653–743. Elsevier.

INEGI (2024). Encuesta nacional sobre disponibilidad y uso de tecnologías de la información en los hogares (endutih) 2023. Comunicado de prensa 372/24, Instituto Nacional de Estadística y Geografía.

INEGI (2025). Encuesta nacional sobre disponibilidad y uso de tecnologías de la información en los hogares (endutih) 2024. Comunicado de prensa 57/25, Instituto Nacional de Estadística y Geografía.

INEGI (2026). Encuesta nacional sobre disponibilidad y uso de tecnologías de la información en los hogares (endutih) 2025. Reporte de resultados 19/26, Instituto Nacional de Estadística y Geografía.

Jiménez, J. E. (2019). *Modelo de respuesta a la intervención*. Comercial Grupo ANAYA, SA.

List, J. A. (2026). The economics of scaling early childhood programs: Lessons from the chicago school. *Journal of Political Economy*, 134(1):1–48.

List, J. A., Shaikh, A. M., and Xu, Y. (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics*, 22(4):773–793.

Ma, Y., Fairlie, R., Loyalka, P., and Rozelle, S. (2024). Isolating the tech from edtech: Experimental evidence on computer-assisted learning in china. *Economic Development and Cultural Change*, 72(4):1923–1962.

Malamud, O. and Pop-Eleches, C. (2011). Home computer use and the development of human capital. *The Quarterly Journal of Economics*, 126(2):987–1027.

Mayer, S. E., Kalil, A., Oreopoulos, P., and Gallegos, S. (2019). Using behavioral insights to increase parental engagement: The parents and children together intervention. *Journal of Human Resources*, 54(4):900–925.

McEwan, P. J. (2015). Improving learning in primary schools of developing countries: A meta-analysis of randomized experiments. *Review of Educational Research*, 85(3):353–394.

Muralidharan, K. and Singh, A. (2025). Adapting for scale: Experimental evidence on technology-aided instruction in india. Technical report, National Bureau of Economic Research.

Muralidharan, K., Singh, A., and Ganimian, A. J. (2019). Disrupting education? experimental evidence on technology-aided instruction in india. *American Economic Review*, 109(4):1426–1460.

Niklas, F., Birtwistle, E., Mues, A., and Wirth, A. (2025). Learning apps at home prepare children for school. *Child Development*, 96(2):577–590.

Pham, T., Joanisse, M. F., Ansari, D., Oram, J., Stager, C., and Archibald, L. M. (2025). Early cognitive predictors of language, literacy, and mathematics outcomes in the primary grades. *Early Childhood Research Quarterly*, 70:187–198.

Pritchett, L. and Beatty, A. (2015). Slow down, youre going too fast: Matching curricula to student skill levels. *International Journal of Educational Development*, 40:276–288.

Rodriguez-Segura, D. (2022). Edtech in developing countries: A review of the evidence.

Singh, A., Navarro-Sola, L., and Oreopoulos, P. (2025). Education technology. VoxDevLit, 20(1).

World Bank Group (2022). *World Development Report 2018: Learning to Realize Education's Promise*. World Bank Publications.

Supplementary Material

When Does EdTech Work? Evidence from a Scalable Home-Based Intervention with School Complementarity

Juanita Bloomfield Ana Balsa Alejandro Cid
June 2026

This Supplementary Material accompanies the main paper. It contains Appendices A, B, C, F, and G, referenced throughout the main text. Appendices D (Evaluation Scales) and E (Sample Construction, Outcome Completion, and Baseline Balance) appear in the main paper.

A Games in Leo Leo App

The app includes the following games:

1. The *Tracing Letters* activity consists of tracing letters expressed in different typographies, each one providing access to the development of different cognitive areas of the child. It has five pedagogical objectives: learning to trace letters, recognizing graphemes, stimulating visual-motor coordination, associating types of letters, and stimulating working memory, attention and inhibitory control.

2. The *Written Pipes* game is about completing/ordering letters/syllables to form different words. It is intended to achieve four objectives: the development of the ability to write words by ordering letters or syllables, to strengthen the alphabetic principle, to encourage reading by syllables or letter by letter, and to incorporate new vocabulary.

3. *Spelling Coconuts* is an activity in which the child must recognize a series of graphemes quickly. Its objectives are to recognize graphemes at an increasing speed, to exercise the association between phoneme and grapheme, to identify ways of writing letters, to develop visual discrimination and visual-motor coordination, and to stimulate sustained attention.

4. The *Fishing for Words* activity consists of the association of phonemes to letters or words in a multiple-choice context. It is intended to achieve the following objectives: develop greater phonological awareness, identify words that contain a specific phoneme, associate graphemes to phonemes, the development of laterality and stimulation of attention and working memory.

5. The *River Sounds* game is developed as an exercise to distinguish the sound of a grapheme within a word. It seeks to achieve four objectives: a greater development of phonological awareness, the identification of words that begin with a specific phoneme, the association of graphemes and phonemes, and the stimulation of sustained attention.

6. *Reading Tides* consists of reading speed training and word reading automation. It has five objectives: to train reading speed and automate reading, promote reading fluency, stimulate auditory discrimination, develop visual word recognition and processing speed, and stimulate attention and working memory.

7. *Navigating Phrases* game is about solving riddles in a multiple-choice context. Its objectives are to develop listening and reading comprehension, strengthen reading fluency, broaden general knowledge, draw inferences from clues and riddles, and stimulate concentration.

B Examples of Messages to Families and Teachers

Families receive two messages per week via WhatsApp, aimed at fostering consistent engagement with the app. Messages are tailored based on each child's progress and activity within the app. Table B1 presents three examples of messages sent to families.

Table B1: Example of messages sent to parents

Type of message	**Example**
Motivation	Hi [caregiver's name]! 🎉 We wanted to congratulate [child's name] on their progress with Leo Leo. She is very close to reaching her goal! Every step counts, and with just a bit more practice, we're confident she will achieve great things ⭐ If [child's name] reaches at least letter G, she will receive her recognition! 🏆 How about we set a goal for her to play three times this week to make it happen? I'll be here to support and celebrate each step forward.
Follow-up	Hi [caregiver's name], How are you? 😊 Just a quick message to say hello and let you know that [child's name] finished last week with XX% progress. Great job! 🏆 If he plays every day, he'll be able to earn his diploma. I'll be here to support you throughout the entire journey.
No progress	Just a quick message to ask why [child's name] hasn't continued playing 😔. Her classmates are making great progress, and I'd really love for her to earn her diploma too. Let me know how I can support you.

In Treatment group 1, teachers are given access to the Leo Leo app to track individual student progress and receive biweekly reports highlighting each student's status. Reports are sent on Monday mornings. The message that teachers receive is the following:

Table B2: Example of messages sent to teachers

Type of message	Example
School with higher progress than expected	Hello [Name of the school] team! I hope you are starting a very nice week. As usual, I am sending the biweekly progress report of your students. On average, you have [X]% progress, which is above the expected for the date. Congratulations! This speaks of the perseverance and commitment of their families and their support. However, we have a group of children who are below [X]% and we are concerned about them. Tell me if it is due to lack of access to the device, lack of interest, or lack of motivation. They are: • Name of the child 1 • Name of the child 2 • Name of the child 3 • Name of the child 4 You can see the detail of each child in the attachment. OK: Means he/she is in line with the expected percentage (40% or more). Priority: About to reach the expected percentage or hasn't played in two weeks. Critical: Below the expected percentage and/or hasn't played in three weeks or more.

Continued on next page

Type of message	Example
School with lower progress than expected	Hello, teachers, how are you? I hope all is well. As I do every 15 days, I come with my biweekly report. I just updated the metrics, and the school has an average of [X]% progress, so we are a little below the expected [X]% for the date. So let's continue to motivate your families and commit to having your children play at least three times a week. We have a group of children who are below [X]% and we are concerned about them. Tell me if it is due to lack of access to the device, lack of interest, or lack of motivation. They are: • Name of the child 1 • Name of the child 2 • Name of the child 3 • Name of the child 4 You can see the detail of each child in the attachment. OK: Means he/she is in line with the expected percentage (40% or more). Priority: About to reach the expected percentage or hasn't played in two weeks. Critical: Below the expected percentage and/or hasn't played in three weeks or more.

In addition to this message, we sent teachers an excel file with the following information of each child in the school: (i) date of report (ii) child's identification, (iii) class group, (iv) status of activation the app – active or inactive. In addition, we categorized each child to guide teachers. We included the following three categories: (i) OK – the student was progressing as expected; (ii) Priority – the student was slightly behind or has not used the app for two weeks; and (iii) Critical – the student was significantly behind or has not used the app for three or more weeks.

Table B3: Example of spreadsheet sent to teachers

First Name	Last Name	Email	Section	Activation Status	Follow-up	Progress Jan. 14
Person 1	Person A	`user1@example.com`	A	Active	Critical	32%
Person 2	Person B	`user2@example.com`	A	Active	Critical	15%
Person 3	Person C	`user3@example.com`	A	Active	OK	75%
Person 4	Person D	`user4@example.com`	A	Active	OK	48%
Person 5	Person E	`user5@example.com`	A	Active	OK	84%
Person 6	Person F	`user6@example.com`	A	Active	Critical	15%
Person 7	Person G	`user7@example.com`	A	Inactive	–	–
Person 8	Person H	`user8@example.com`	A	Active	OK	78%

C School compliance with treatment assignment

Table C1: Rate of schools that participated in the program

Comparison	Difference in participation probability	s.e.	Constant	s.e.	N
T1 vs. Control	0.057	(0.068)	0.860***	(0.072)	70
T2 vs. Control	0.086	(0.061)	0.846***	(0.071)	70
T1 vs. T2	-0.029	(0.049)	0.959***	(0.046)	70

Notes: Table shows results from a regression of a school participation dummy on treatment assignment and randomization strata fixed effects. Cluster-corrected standard errors are shown in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

F Balance of randomized schools

Table F1: Descriptive statistics of schools at baseline by treatment arm and balance

	Control			Treatment 1			Treatment 2			Diff. T1 – Control			Diff. T2 – Control			Diff. T1 – T2		
Variable	Mean	s.d.	N	Mean	s.d.	N	Mean	s.d.	N	Diff.	s.e.	N	Diff.	s.e.	N	Diff.	s.e.	N
Number of children enrolled (5 years old)	44.486	22.293	35	43.371	21.037	35	42.343	18.946	35	-1.114	4.027	70	-2.143	3.802	70	-1.029	3.538	70
Total enrolled with behavior issues	1.771	2.734	35	1.657	2.086	35	1.143	1.683	35	-0.114	0.586	70	-0.629	0.547	70	-0.514	0.448	70
Total enrolled with communicational difficulties	3.400	4.526	35	2.171	3.240	35	2.029	3.303	35	-1.229	0.947	70	-1.371	0.954	70	-0.143	0.784	70
Total enrolled with learning disabilities	1.314	2.784	35	2.171	3.519	35	2.086	4.468	35	0.857	0.747	70	0.771	0.886	70	-0.086	0.878	70
Number of third year groups	2.571	0.979	35	2.629	1.087	35	2.486	0.919	35	0.057	0.187	70	-0.086	0.175	70	-0.143	0.187	70
Number of third grade teachers	2.600	1.006	35	2.657	1.110	35	2.486	0.919	35	0.057	0.188	70	-0.114	0.177	70	-0.171	0.188	70
Number of third grade schoolrooms	2.543	0.919	35	2.629	1.087	35	2.429	0.884	35	0.086	0.181	70	-0.114	0.169	70	-0.200	0.188	70
Double shift	0.143	0.355	35	0.171	0.382	35	0.057	0.236	35	0.029	0.088	70	-0.086	0.072	70	-0.114	0.076	70
Location = Monterrey	0.086	0.284	35	0.200	0.406	35	0.114	0.323	35	0.114	0.084	70	0.029	0.073	70	-0.086	0.088	70
Literacy at the municipality level	0.991	0.004	35	0.99	0.004	35	0.99	0.004	35	-0.001	0.001	70	-0.001	0.001	70	0	0.001	70
Caregiver completed primary education at the municipality level	0.987	0.007	35	0.968	0.007	35	0.986	0.007	35	-0.001	0.002	70	-0.001	0.002	70	0	0.002	70
Average years of schooling of caregiver at the municipality level	10.364	0.599	35	10.23	0.385	35	10.392	0.134	35	-0.134	0.121	70	0.027	0.134	70	0.162	0.108	70

Notes: Difference estimates and standard errors come from regressions of each baseline characteristic on treatment indicators (one comparison at a time), including randomization strata fixed effects. Standard errors are cluster-robust at the school level. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

G Outcome Completion Rates, Balance Tables and Descriptive Statistics at Follow-up

Table G1: Outcome specific completion rates

Outcome	Comparison	Difference in response probability	s.e.	Constant	s.e.	N
IDELA†	T1 vs. Control	0.000	(–)	1.000	(–)	2139
	T2 vs. Control	0.000	(–)	1.000	(–)	2187
	T1 vs. T2	0.000	(–)	1.000	(–)	2282
SDQ	T1 vs. Control	-0.012	(0.018)	0.995***	(0.015)	2139
	T2 vs. Control	-0.040	(0.028)	1.010***	(0.019)	2187
	T1 vs. T2	0.028	(0.030)	0.980***	(0.015)	2282
Family questionnaire	T1 vs. Control	0.027	(0.021)	0.269***	(0.018)	2139
	T2 vs. Control	0.001	(0.022)	0.268***	(0.019)	2187
	T1 vs. T2	0.026	(0.022)	0.264***	(0.017)	2282
Teacher questionnaire	T1 vs. Control	0.118	(0.082)	0.540***	(0.074)	169
	T2 vs. Control	0.102	(0.081)	0.516***	(0.073)	168
	T1 vs. T2	0.010	(0.079)	0.573***	(0.073)	175
App log data‡	T1 vs. T2	0.067	(0.077)	0.677***	(0.061)	2282

Notes: Cluster-corrected standard errors are shown in parentheses. IDELA, SDQ, and the family questionnaire are estimated over the common analytic sample of children who completed both the Leo Test and IPAL (the sample used in Tables 6, 7, 8, and 10). The teacher questionnaire is estimated over the full sample of participating schools, since teacher level data collection was not tied to child level test completion. †All children in the common analytic sample were administered IDELA, so the completion indicator has no variation across arms; the regression is therefore mechanical (diff = 0, constant = 1 by construction) and we report it for completeness rather than as a balance test. ‡App log data are only observed for children assigned to Treatment 1 or Treatment 2, since children in the control group were never given access to the app; the comparison is therefore restricted to Treatment 1 vs. Treatment 2. ***p<0.01, **p<0.05, *p<0.10.

Table G2: Descriptive statistics and balance of children that answered Leo Test and IPAL/IDELA

	Control			Treatment 1			Treatment 2			Diff. T1 – Control			Diff. T2 – Control			Diff. T1 – T2		
Variable	Mean	s.d.	N	Mean	s.d.	N	Mean	s.d.	N	Diff.	s.e.	N	Diff.	s.e.	N	Diff.	s.e.	N
Panel A: School																		
Number of children enrolled (5 years old)	45.93	23.56	30	43.94	21.22	33	42.68	19.13	34	-1.57	4.45	63	-3.26	4.20	64	1.67	3.70	67
Total enrolled with behavior issues	2.03	2.87	30	1.64	2.13	33	1.18	1.70	34	-0.40	0.65	63	-0.86	0.60	64	0.47	0.47	67
Total enrolled with communicational difficulties	3.80	4.71	30	1.94	3.03	33	2.09	3.33	34	-1.85*	1.02	63	-1.71	1.04	64	-0.14	0.78	67
Total enrolled with learning disabilities	1.43	2.96	30	2.00	3.33	33	2.15	4.52	34	0.58	0.80	63	0.71	0.95	64	-0.10	0.89	67
Number of third year groups	2.60	1.04	30	2.67	1.08	33	2.50	0.93	34	0.09	0.20	63	-0.10	0.19	64	0.19	0.19	67
Number of third grade teachers	2.60	1.04	30	2.70	1.10	33	2.50	0.93	34	0.12	0.21	63	-0.10	0.19	64	0.22	0.19	67
Number of third grade schoolrooms	2.57	0.97	30	2.67	1.08	33	2.44	0.89	34	0.12	0.20	63	-0.13	0.19	64	0.24	0.19	67
Double shift	0.13	0.35	30	0.15	0.36	33	0.06	0.24	34	0.02	0.09	63	-0.07	0.08	64	0.09	0.08	67
Location = Monterrey	0.10	0.31	30	0.18	0.39	33	0.12	0.33	34	0.08	0.09	63	0.02	0.08	64	0.06	0.09	67
Panel B: Teacher																		
Female	1.00	0.00	57	0.99	0.12	68	1.00	0.00	63	-0.01	0.01	125	0.00	0.00	120	-0.01	0.01	131
Weekly hours of work	25.62	9.36	57	24.67	8.43	64	24.81	8.21	58	-0.99	2.07	121	-0.80	1.90	115	-0.12	1.59	122
Weekly hours at studied center	22.73	6.51	54	21.28	5.17	66	21.81	4.16	59	-1.40	1.05	120	-0.91	1.10	113	-0.48	0.75	125
Years of experience	15.10	9.11	53	16.91	7.75	66	12.48	7.46	62	1.88	2.05	119	-2.52	2.06	115	4.46***	1.62	128
Years of experience at studied center	8.91	8.47	53	8.15	6.20	66	7.21	5.52	61	-0.76	1.55	119	-1.73	1.64	114	0.95	1.13	127
Years of education	21.15	2.03	53	20.94	2.47	68	20.53	2.02	62	-0.14	0.44	121	-0.61	0.39	115	0.42	0.45	130
Panel C: Child																		
Female	0.49	0.50	1017	0.51	0.50	1112	0.50	0.50	1160	0.03	0.02	2129	0.01	0.02	2177	0.02	0.02	2272
Age	5.86	0.29	1007	5.88	0.29	1112	5.87	0.29	1143	0.02	0.01	2119	0.00	0.01	2150	0.01	0.01	2255
Completed second grade of preschool	0.86	0.35	1018	0.93	0.26	1114	0.79	0.41	1161	0.07	0.05	2132	-0.07	0.07	2179	0.14**	0.06	2275
Caregiver's years of education	13.67	2.70	932	13.76	2.72	1058	13.26	2.55	1015	0.09	0.38	1990	-0.43	0.39	1947	0.51*	0.28	2073
Caregiver's age	33.00	6.30	914	33.11	6.77	1020	32.71	6.44	902	0.11	0.59	1934	-0.31	0.55	1816	0.43	0.47	1922
Panel D: Household																		
Number of books index	1.40	0.79	213	1.36	0.77	269	1.40	0.77	230	-0.04	0.08	482	0.02	0.10	443	-0.06	0.10	499
Frequency of parental involvement: reading	1.36	0.85	213	1.36	0.84	269	1.37	0.81	230	0.00	0.07	482	0.00	0.07	443	-0.01	0.07	499
Frequency of parental involvement: games	2.77	0.96	213	2.76	0.89	269	2.78	0.88	230	-0.02	0.09	482	0.01	0.09	443	-0.03	0.09	499
Literacy at municipality level	0.99	0.01	30	0.99	0.01	33	0.99	0.01	34	0.00	0.00	63	0.00	0.00	64	0.00	0.00	67
Caregiver completed primary education (municipality)	0.99	0.00	30	0.99	0.00	33	0.99	0.00	34	0.00	0.00	63	0.00	0.00	64	0.00	0.00	67
Average years of schooling of caregiver (municipality)	10.35	0.64	30	10.23	0.40	33	10.41	0.53	34	-0.13	0.14	63	0.05	0.15	64	-0.18	0.11	67

Notes: Cluster-corrected standard errors are shown in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Table G3: Baseline balance among family questionnaire respondents

	Control			Treatment 1			Treatment 2			Diff. T1 – Control			Diff. T2 – Control			Diff. T1 – T2		
Variable	Mean	s.d.	N	Mean	s.d.	N	Mean	s.d.	N	Diff.	s.e.	N	Diff.	s.e.	N	Diff.	s.e.	N
Panel A: Child																		
Female	0.454	0.499	262	0.497	0.501	316	0.487	0.501	300	0.042	0.037	578	0.032	0.039	562	0.010	0.038	616
Age	5.851	0.304	258	5.898	0.296	317	5.863	0.291	295	0.046**	0.022	575	0.012	0.021	553	0.034	0.021	612
Completed second grade of preschool	0.870	0.337	262	0.931	0.255	317	0.863	0.344	300	0.060	0.046	579	-0.006	0.060	562	0.067	0.051	617
Caregiver's years of education	13.529	2.651	242	14.026	2.690	302	13.483	2.527	271	0.498	0.438	544	-0.057	0.431	513	0.555	0.383	573
Caregiver's age	32.344	6.238	240	33.566	6.884	294	33.538	6.682	240	1.223	0.768	534	1.185	0.746	480	0.076	0.683	534
Panel B: Household																		
Number of books index	1.429	0.781	84	1.505	0.862	107	1.452	0.715	93	0.076	0.125	191	0.033	0.112	177	0.041	0.139	200
Frequency of parental involvement: reading	1.548	0.884	84	1.430	0.891	107	1.355	0.761	93	-0.117	0.116	191	-0.204**	0.093	177	0.080	0.114	200
Frequency of parental involvement: games	2.964	0.828	84	2.710	0.901	107	2.849	0.779	93	-0.254**	0.112	191	-0.114	0.113	177	-0.136	0.134	200

Notes: Mean, standard deviation in parentheses, and number of observations in square brackets for each group. Sample restricted to children in the common Leo Test/IPAL analytic sample who also completed the endline family questionnaire (Panel A, N=262/317/300 in Control/T1/T2). Panel B is further restricted to the subset of these respondents who answered the household background items, which had lower coverage at baseline (N=84/107/93). Difference estimates and standard errors come from regressions of each characteristic on treatment indicators (one comparison at a time), including randomization strata fixed effects. Standard errors are cluster-robust at the school level. ***p<0.01, **p<0.05, *p<0.10.

Table G4: Descriptive statistics of the outcomes at follow-up. Control observations only.

	Mean	SD	Min	Max	N
Panel A: Leo Test					
Letter-name and letter-sound recognition	0.987	0.112	0	1	1022
Initial sound awareness	0.876	0.327	0	1	1022
Intermediate sound awareness	0.525	0.499	0	1	1022
Word reading	0.423	0.494	0	1	1022
Reading comprehension	0.259	0.438	0	1	1022
Word writing	0.109	0.312	0	1	1022
Number of subtests completed	5.429	2.019	1	9	1022
Leo score	4.692	2.380	0.00	14.30	1022
Panel B: SDQ					
Hyperactivity score	1.345	1.531	-3	7	1012
Peer problems score	-3.066	0.726	-7	0	996
Often complains of headaches, stomach aches, or nausea	1.067	0.295	1	3	990
Panel C: IPAL					
Correct answers	20.826	23.855	0	80	901
Letter name knowledge	7.455	10.530	0	69	1021
Letter sound knowledge	6.578	9.747	0	55	1020
Panel D: IDELA					
Correct numbers identified	12.065	6.325	0	20	1018
Correct arithmetic operations	2.416	0.766	0	3	1012
Panel E: Family questionnaire					
It is hard for my child to stop using screens	2.137	1.201	1	5	262
Screen entertainment is my child's unique motivation	2.145	1.281	1	5	262
My child is always thinking about screen entertainment	1.751	1.029	1	5	262
My child's screen use interferes with family activities	1.801	1.144	1	5	262
My child's screen use causes problems for the family	1.679	1.116	1	5	262
My child gets frustrated when not allowed to use screens	1.713	1.067	1	5	262
My child's desire to use screens is increasing	1.706	0.967	1	5	262
My child hides to use screens	1.221	0.658	1	5	262
My child turns to screens for comfort on bad days	1.408	0.833	1	5	262
Frequency of reading books with child	2.690	1.765	0	7	262
Frequency of playing writing games (letters/name)	4.694	1.719	0	7	262
Frequency of teaching child to write words	4.403	1.901	0	7	262
Panel F: Teacher questionnaire					
Groups children by level	6.167	3.010	1	10	12
Plans differentiated activities by child's ability level	8.083	2.429	2	10	12
Uses technology to reinforce language learning	8.083	2.151	2	10	12
Reviews individual language progress outside the classroom	7.500	2.067	3	10	12
Expects students to read fluently by age 8	19.333	5.033	10	30	12
Expects students to finish secondary school	22.000	5.985	8	30	12
Expects students to start university	15.000	6.537	1	25	12
Expects students to complete university	15.000	6.537	1	25	12
Evaluates the role of families in children's language learning	7.167	2.758	2	10	12
Considers parental help at home frequent	6.500	2.067	2	10	12
Believes technology can improve learning outside the classroom	8.083	3.029	2	10	12
Feels capable of ensuring all students communicate effectively	8.333	2.188	2	10	12
Motivated by receiving concrete feedback on student progress	8.917	2.314	2	10	12
Identifying students with greater literacy difficulties relative to peers	7.071	2.176	3	10	28
Differentiated instruction by student level	7.857	2.138	3	10	28

Notes: This table reports descriptive statistics for all outcome variables measured at follow-up.

Table G5: Robustness to Differential Outcome Completion: Inverse-Probability-Weighted Estimates of the Effects of the Intervention on Literacy

	Treatment 1 vs. Control		Treatment 2 vs. Control		Treatment 1 vs. Treatment 2	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Leo Test						
Overall score	0.301***	0.265***	0.086	0.130	0.223***	0.132*
	(0.090)	(0.080)	(0.102)	(0.081)	(0.074)	(0.074)
	[2139]	[2139]	[2087]	[2087]	[2182]	[2182]
Number of subtests taken	0.210***	0.189***	0.046	0.073	0.175***	0.117*
	(0.070)	(0.068)	(0.074)	(0.063)	(0.055)	(0.059)
	[2139]	[2139]	[2087]	[2087]	[2182]	[2182]
Panel B: IPAL Test						
Phonological awareness	0.276***	0.217***	0.116	0.106	0.146*	0.072
	(0.084)	(0.080)	(0.090)	(0.084)	(0.082)	(0.078)
	[1955]	[1955]	[1973]	[1973]	[2126]	[2126]
Letter name knowledge	0.173**	0.136*	0.049	0.033	0.078	0.064
	(0.081)	(0.077)	(0.074)	(0.068)	(0.067)	(0.069)
	[2136]	[2136]	[2181]	[2181]	[2275]	[2275]
Letter sound knowledge	0.153*	0.103	0.061	0.040	0.152**	0.110
	(0.085)	(0.082)	(0.080)	(0.076)	(0.068)	(0.066)
	[2131]	[2131]	[2183]	[2183]	[2274]	[2274]

Notes: This table reestimates Table 4 of the main paper weighting each regression by a stabilized inverse-probability-of-response weight. Weights are estimated in two steps. First, a LASSO logit model of joint response (completing both the Leo Test and IPAL), with the penalty parameter selected via 10-fold cross-validation, is estimated on five candidate baseline covariates (child's sex, child's age, child's education, caregiver's education, and caregiver's age), with treatment assignment and strata fixed effects forced into the model and left unpenalized; evaluator fixed effects are not included in this selection model. Predicted probabilities from this model form the weight denominator. Second, a standard logit of joint response on treatment assignment and strata fixed effects only, excluding all baseline covariates, is estimated; predicted probabilities from this model form the weight numerator. The stabilized weight is the ratio of the numerator to the denominator predicted probability, and is trimmed at the 1st and 99th percentiles to limit the influence of extreme values. Weights are estimated separately for each treatment comparison, using only the two arms relevant to that comparison. See Section 3.6 of the main paper for further detail. Cluster-corrected standard errors are shown in parentheses. Number of observations are reported in square brackets. Odd columns report the ITT coefficient without covariates. Even columns report the ITT coefficient from specifications controlling for baseline unbalanced controls: child's education, caregiver's education and teacher experience. All regressions include strata fixed effects and enumerator fixed effects. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Table G6: Pooled Estimates of Main Outcomes

	T1 vs. Control	T2 vs. Control	P-value ($T1 = T2$)	N
	(1)	(2)	(3)	(4)
Panel A: Leo Test				
Overall score	0.305***	0.076	0.005***	3204
	(0.097)	(0.107)		
Number of subtests taken	0.211***	0.039	0.003***	3204
	(0.073)	(0.076)		
Panel B: IPAL Test				
Phonological awareness	0.254***	0.116	0.083*	3027
	(0.094)	(0.095)		
Letter name knowledge	0.126	0.074	0.426	3296
	(0.087)	(0.080)		
Letter sound knowledge	0.121	0.021	0.126	3294
	(0.092)	(0.088)		
Panel C: Leo Subtests				
Letter-name and letter-sound recognition	0.003	-0.002	0.340	3204
	(0.005)	(0.006)		
Initial sound awareness	0.037**	0.009	0.059*	3204
	(0.015)	(0.017)		
Intermediate sound awareness	0.102***	0.036	0.010**	3204
	(0.030)	(0.032)		
Word reading	0.110***	0.040	0.012**	3204
	(0.031)	(0.034)		
Reading comprehension	0.071**	0.004	0.004***	3204
	(0.029)	(0.029)		
Word writing	0.028	-0.011	0.033**	3204
	(0.027)	(0.027)		
Panel D: IDELA				
Correct numbers identified	0.086	0.019	0.320	3280
	(0.076)	(0.085)		
Correct arithmetic operations	0.279**	0.092	0.018**	3235
	(0.118)	(0.089)		

Notes: This table reports pooled OLS estimates of the effects of Treatment 1 and Treatment 2 relative to the control group using the full sample. The estimated specification is: $Y_{is} = \beta_0 + \beta_1 T1_s + \beta_2 T2_s + \varepsilon_{is}$ where Y_{is} is the outcome of interest for child i in school s, and $T1_s$ and $T2_s$ are indicators for assignment to Treatment 1 and Treatment 2, respectively. Control schools are the omitted category. Columns 1 and 2 report coefficients relative to the control group. Column 3 reports the p-value from an F-test of equality between the Treatment 1 and Treatment 2 coefficients. Regressions include strata and evaluator fixed effects. Standard errors clustered at the school level are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Table G7: Teacher's Questionnaire Results. Coefficients in Standard Deviations.

	Treatment 1 vs. Control	Treatment 2 vs. Control	Treatment 1 vs. Treatment 2
	(1)	(2)	(3)
Plans differentiated activities by child's ability level (frequency)	0.156	0.049	0.197
	(0.289)	(0.305)	(0.235)
	[34]	[36]	[46]
Uses technology to reinforce language learning (frequency)	0.156	0.049	0.197
	(0.289)	(0.305)	(0.235)
	[34]	[36]	[46]
Reviews individual language progress outside the classroom (frequency)	0.035	-0.072	0.136
	(0.328)	(0.337)	(0.283)
	[34]	[36]	[46]
Expects students to read fluently by age 8 (percentage)	-0.039	0.002	-0.002
	(0.457)	(0.367)	(0.402)
	[34]	[36]	[46]
Expects students to finish secondary school (percentage)	0.195	-0.038	0.230
	(0.302)	(0.347)	(0.251)
	[34]	[36]	[46]
Expects students to start university (percentage)	0.361	-0.117	0.566*
	(0.335)	(0.340)	(0.291)
	[34]	[36]	[46]
Expects students to complete university (percentage)	0.361	-0.117	0.566*
	(0.335)	(0.340)	(0.291)
	[34]	[36]	[46]
Believes technology can improve learning outside the classroom	0.229	0.156	0.152
	(0.280)	(0.292)	(0.256)
	[34]	[36]	[46]
Feels capable of ensuring all students communicate effectively	0.282	0.229	0.187
	(0.281)	(0.303)	(0.240)
	[34]	[36]	[46]
Motivated by receiving concrete feedback on student progress	0.144	0.140	0.081
	(0.275)	(0.300)	(0.229)
	[34]	[36]	[46]
Frequency of reminders to students for home app use			0.295
			(0.214)
			[74]
Identifying students with greater literacy difficulties relative to peers	0.387	0.093	0.134
	(0.249)	(0.272)	(0.192)
	[71]	[61]	[75]
Differentiated instruction by student level	0.058	0.197	-0.180
	(0.245)	(0.263)	(0.234)
	[71]	[61]	[75]
Difficulty in allocating attention to Leo Leo due to other classroom priorities			0.013
			(0.239)
			[74]
School leadership support for the use of supplementary tools such as Leo Leo			0.574***
			(0.200)
			[73]

Notes: Cluster-corrected standard errors are shown in parentheses. Number of observations are reported in square brackets. All regressions include strata fixed effects. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Table G8: Baseline balance among children with observed app engagement data (Treatment 1 vs. Treatment 2)

	Treatment 1			Treatment 2			Diff. T1 – T2		
Variable	Mean	s.d.	N	Mean	s.d.	N	Diff.	s.e.	N
Panel A: Child									
Female	0.505	0.500	1231	0.513	0.500	1113	-0.008	0.020	2344
Age	5.873	0.291	1227	5.863	0.297	1106	0.010	0.011	2333
Completed second grade of preschool	0.902	0.298	1233	0.819	0.386	1113	0.083	0.064	2346
Caregiver's years of education	13.881	2.764	1153	13.491	2.466	951	0.395	0.317	2104
Caregiver's age	33.006	6.366	1091	32.818	6.381	881	0.187	0.458	1972
Panel B: Household									
Number of books index	1.373	0.788	306	1.386	0.789	277	-0.036	0.097	583
Frequency of parental involvement: reading	1.346	0.859	306	1.347	0.800	277	-0.005	0.072	583
Frequency of parental involvement: games	2.755	0.885	306	2.776	0.905	277	-0.021	0.082	583

Notes: Mean, standard deviation in parentheses, and number of observations in square brackets for each group. Sample restricted to children in Treatment 1 or Treatment 2 with observed Leo Leo app engagement data (i.e., who ever activated the app; see Section 4.4 of the main text and the notes to Table 10). Control is omitted because children in the control group never received access to the application and therefore have no engagement data. Difference estimates and standard errors come from a regression of each characteristic on a Treatment 1 indicator, including randomization strata fixed effects. Standard errors are cluster-robust at the school level. ***p<0.01, **p<0.05, *p<0.10.