SERIEs
https://doi.org/10.1007/s13209-026-00352-1

ORIGINAL ARTICLE

Game changer? Impact of a multi-component reading intervention on cognitive and socioemotional skills

Javier Garcia-Brazales[1] · Micole De Vera[2] · Luz Rello[3]

Received: 4 March 2026 / Accepted: 25 August 2026
© The Author(s) 2026

Abstract

We evaluate a holistic reading intervention involving 527 third-grade students in Chilean schools catering to disadvantaged populations. The intervention features an adaptive computer game designed to identify and improve weaknesses in literacy and cognitive skills, and is complemented by a mobile library and advice to parents aimed at fostering student engagement and parental involvement. Leveraging a matched-pair design under the staggered implementation of the intervention, in which we retrospectively match five schools already scheduled to receive the treatment with five schools from the yet-to-be-selected pool, we assess its impacts on both cognitive and socioemotional outcomes. Treated students are 10–30% of a standard deviation more likely to have higher academic aspirations, to believe that their performance is better than that of their peers, and that courses are easy. We also find suggestive evidence of a more internal locus of control and of improvements in performance on a nation-wide, standardized language test, indicating that a short, light-touch, and cost-effective bundled program can improve socioemotional outcomes and, potentially, reading performance.

Keywords Reading intervention · Educational technology · Socioemotional skills · Chile

JEL Classification I24 · I31

We thank Samuel Bentolila for being the catalyst for the collaboration that led to this project. We thank Fundación Piñera Morel for their integral contribution in designing and implementing the intervention studied in this paper. The authors declare the following potential conflict of interest: Rello is one of the creators of Dytective, which is part of the subject of evaluation in this paper. The research design, data processing, and estimation were undertaken solely by De Vera and Garcia-Brazales. The views expressed in this paper are those of the authors and do not necessarily reflect the position of the Banco de España or the Eurosystem.

✉ Javier Garcia-Brazales
j.garcia-brazales2@exeter.ac.uk

[1] University of Exeter and LEAP (Bocconi), Streatham Court, Rennes Drive, Exeter EX4 4PU, UK

[2] Banco de España, Madrid, Spain

[3] IE University, Madrid, Spain

Published online: 04 October 2026

1 Introduction

Low academic progress is a worldwide concern, and governments around the world continue to spend large amounts of resources to address this issue (Pritchett 2013; Singh 2020). This is particularly challenging for developing and emerging countries (Glewwe and Muralidharan 2016). For instance, in Chile, 60% of second-grade students lag behind their expected reading level by at least 6 months despite being consistently ranked among Latin America's top performers in standardized tests such as PISA.[1]

There are several reasons why strengthening academic progress may not be achieved solely by addressing institutional constraints like teacher quality or underprovision of educational materials. First, issues such as lack of student motivation or aspirations and limited parental investment are additional deterrents to academic performance (e.g., Heckman et al. 2007; Resnjanskij et al. 2024). Second, institutions may be unprepared to teach increasingly diverse classrooms, where students differ in linguistic backgrounds, learning needs, and socioemotional profiles. Such heterogeneity can reduce instructional effectiveness and contribute to students leaking out of the educational pipeline.[2] As such, interventions that have a strong chance to improve education need to be holistic in involving parents, teachers, and students, and effective across a variety of learners. However, developing holistic and inclusive interventions at scale is challenging. A source of renewed hope has been the introduction of education technology (Escueta et al. 2020).

In this paper, we evaluate the impact of a multifaceted and scalable intervention called "A Leer Jugando," implemented by Fundación Piñera Morel (FPM), which aims to improve the reading skills of Chilean third graders from relatively disadvantaged backgrounds. The intervention not only intends to directly enhance reading skills, but also strives to cultivate a joy for reading among students and encourages the involvement of parents in the reading development of their children. To achieve this goal, the intervention is innovative in that it integrates an adaptive computer-based reading game with complementary activities designed to promote both children's reading interest and parents' engagement in the process.

At the core of the program there is Dytective, a gamified language educational application that draws from a collection of over 42,000 linguistic exercises developed based

[1] These numbers are based on the report "Radiografía de la Lectura en Segundo Básico: Resultados de Evaluación Muestral de la Región Metropolitana 1er Semestre 2023" by researchers from the Pontificia Católica, Chile and Andes universities. For more details, see https://gobierno.uc.cl/noticias/el-60-de-estudiantes-de-segundo-basico-estan-bajo-los-niveles-de-comprension-lectora-esperados-para-su-edad/.

[2] More broadly, low academic progress is a multifaceted issue that also reflects factors such as teachers' low content knowledge (Bold et al. 2019), low parental education (Dickson et al. 2016), or chronic parental underinvestment (Dizon-Ross 2019).

on common reading difficulties among Spanish-speaking dyslexic children.[3] The learning program in Dytective is personalized wherein student-specific "challenges"—composed of a number of exercises—are generated based on past performance in the app and adapted to improve previous weaknesses in certain linguistic areas and cognitive skills. The application was designed to enhance, among others, the spelling, reading speed, and vocabulary of the participants, irrespective of their initial ability or whether they were at risk of dyslexia. Dytective is played during 45-minute school visits that are facilitated by an educational psychologist[4] three times a week for three months. In these sessions, which take place during regular Spanish language classes, the facilitator distributes tablets to the students for them to access their individual Dytective profiles and supervises their work.

To increase the students' interest in reading and parental involvement, the intervention features two auxiliary programs: a weekly mobile library where students may borrow books and other reading-related games, and weekly text messages to the parents with tips on how to take advantage of daily life situations to encourage their child to practice their reading and/or writing. Therefore, though the program is geared toward improving reading, it also contains features that may inadvertently strengthen socioemotional skills such as concentration, grit, and self-confidence, especially as reading ability strengthens. For instance, as children are encouraged to use their reading ability to help in chores such as grocery shopping, they may develop better confidence in themselves. This is attractive because recent evidence indicates that such socioemotional skills can be as important as cognitive skills, if not more, in predicting academic, health, and labor market outcomes at mid- and late-life (e.g., Kautz et al. 2014).

The combination of *adaptive low-cost computer-based learning* with complementary elements that have the potential to strengthen a *growth mindset*[5] is a key feature of the program. This integrated approach distinguishes it from most computer-based interventions, which typically focus exclusively on cognitive skill acquisition and standardized test outcomes (Muralidharan et al. 2019; Escueta et al. 2020), as well as from traditional socioemotional programs, which generally lack personalized and

[3] This application has been developed by Change Dyslexia founded by Luz Rello. Change Dyslexia is a decade-long project that has received multiple awards and grants, has reached more than 400,000 individuals in over 130 countries, and has recently signed an agreement to be present in all public and charter schools in Madrid, Spain, funded by the European Commission's Horizon 2020 and the Spanish Ministry of Science and Innovation. The agreement with the Community of Madrid to extend the use of Dytective to all its public and charter schools (around 1,250) can be found in the following link: https://www.comunidad.madrid/noticias/2023/10/22/comunidad-madrid-extiende-todos-centros-educativos-sostenidos-fondos-publicos-su-programa-ayuda-dislexia. A preliminary evaluation of an earlier stage of expansion with 107 schools in Madrid by Cuevas-Ruiz et al. (2021) suggests gains in English and Spanish for girls and in English for boys, although the authors caution against a causal interpretation due to the non-random allocation of the program across schools. To minimize potential conflicts of interest, the design, data processing, and estimation tasks of the present paper were conducted exclusively by De Vera and Garcia-Brazales.

[4] The educational psychologists possess tertiary education training in both pedagogy and psychology and were already part of FPM's workforce. The implementation of the program does not require specialized clinical or pedagogical skills. A key advantage of technology-based learning tools is that they minimize reliance on highly specialized personnel.

[5] Growth mindset refers to the belief that abilities can be acquired and that success can be achieved through effort. It has been shown to be predictive of, for instance, educational achievement (Blackwell et al. 2007).

adaptive academic instruction. Moreover, our intervention is easily scalable and its cost-effectiveness is comparable to the most attractive programs currently available, such as the one in Carlana and Ferrara (2021).

Our evaluation comprises 527 third graders in ten schools in the Chilean Metropolitan Region. At the time of partnering up with FPM, the program was already scheduled to be implemented in five schools for the second semester of 2023. To quantify the overall impact of A Leer Jugando, our preferred specification takes advantage of the staggered implementation of the program: For each treated school, we identify a similar corresponding control school that was regarded by FPM as equally attractive for program participation, but had not been selected for implementation during our study period given budgetary constraints.

To perform the pairing, we search over the pool of available schools and find the best match along the three key school-level characteristics employed by FPM to determine program participation: an educational vulnerability index that is highly predictive of school segregation and widely employed by Chilean governmental institutions for determining access to public support, school size, and school location at the commune level. This research design relies on potential outcomes of students being the same within the match pair. Although this assumption is fundamentally untestable, we show that within the resulting match pairs, treatment and control participants are indeed balanced across a wide range of predetermined characteristics and baseline measures of outcomes that were not used in the matching, which suggests that participants are plausibly balanced in unobservables. We further condition on the baseline values of the outcomes to make our conditional ignorability assumption more plausible. This choice has the additional advantage of making our specification coincide with value-added models commonly used in estimating human capital production functions and in the evaluation of education interventions (Todd and Wolpin 2003; Andrabi et al. 2011; Singh 2015). We first study the effects on socioemotional skills and perceptions elicited through an ad hoc survey that we designed and distributed before and after the intervention. We then quantify the impact of the intervention on reading ability as measured by a standardized national reading test held three times per year by the Chilean Education Quality Assurance Agency.

Main results. We find that the bundled intervention generates improvements in students' self-perceptions, most notably their academic aspirations, their sense of performing relative to peers, and their perceptions of course difficulty. We also detect suggestive gains in locus of control and some dimensions of well-being. These psychological and socioemotional changes appear to precede and potentially enable improvements in academic outcomes. Consistent with this interpretation, we find suggestive evidence that students in treated schools subsequently experience reading gains of about 13% of a standard deviation. However, these estimates are obtained from a smaller matched sample and are less robust to selection on unobservables. Notably, these impacts do not seem to be accompanied by higher investments by either the child (study time) or by parents (caring about the child's academics or time devoted to helping with homework).

Robustness. Acknowledging the possible limitations of our study design, we take steps to assuage concerns about the validity and robustness of our conclusions. We focus on two aspects: internal validity and inference. In terms of *internal validity*, the main threats are non-random attrition and selection on unobservables. Regarding non-random attrition, we follow Muralidharan et al. (2019) and re-estimate our main specifications weighting each observation by the inverse of the probability of remaining in the sample at endline. Our results remain quantitatively similar. Regarding selection on unobservables, we conduct analyses following Oster (2019) and Masten et al. (2024) to assess robustness to possible selection on unobservables. These analyses measure the extent to which selection on unobservables has to be relative to selection on observables to turn the findings into null results. The observables we include in these analyses include baseline characteristics that are highly predictive of the outcomes we measure. We find that the level of selection on unobservables needed to overturn our results for self-perception and locus of control would have to be significantly larger than the selection based on the observables we consider. The result on reading is less robust to selection on unobservables.

In terms of *inference*, we are concerned about two particular issues: (i) Standard errors based on asymptotic approximations may lead to incorrect statistical inference in our setting with a small number of clusters, and (ii) our results could arise spuriously from the multiple treatment effects we measure across different outcomes. Regarding statistical inference, as suggested by Mackinnon et al. (2023a), we report multiple methods of inference to assess the robustness of the statistical significance of our main results. In our main analyses, we report p-values from a wild cluster bootstrap procedure that has been shown to provide reliable inference in setting with as few as five clusters Cameron et al. (2008). We further report clustered standard errors with degrees of freedom adjustment and cluster-robust jackknife-based standard errors (Mackinnon et al. 2023b). We find that our conclusions remain consistent across different methods of inference. Regarding the issue on multiple hypothesis testing, we estimate treatment effects on general composite indices of the socioemotional outcomes we consider. We find that participation in the program improves self-perceptions.

Contributions to related literature. Our work naturally connects with three strands of the literature: (1) evaluation of reading interventions, (2) measurement of impacts of the use of education technologies, and (3) determinants and malleability of socioemotional skills.

Relative to the extensive literature on reading interventions (for recent reviews see Scammacca et al. 2016; Graham and Kelly 2019; Kim et al. 2020),[6] this paper contributes along two directions. *First*, our analysis goes beyond exclusively measuring effects on student outcomes as we explicitly quantify the evolution of parental time investments on children, a crucial input in human capital production traditionally overlooked in this literature (Cunha et al. 2010; Carneiro et al. 2024). *Second*, we evaluate

[6] The main conclusion of these meta-analyses is arguably that reading interventions are generally effective in improving the various components behind the reading process (e.g., phonological awareness or vocabulary building) but these results might not always translate into meaningful gains in reading ability and that more research is needed to identify which complementary elements of the learning process (e.g., teacher quality, students' disabilities) are important in facilitating the success of reading interventions.

a program that provides a *holistic* approach by combining a reading-enhancement element with two other components that involve children's socioemotional skills and parental investments. The existing literature has focused on programs that feature these components individually: for instance, improving access to reading material (e.g., Anger et al. 2024) or urging family involvement through text messages (e.g., Bergman et al. 2018; Bergman 2021). The program we evaluate aims to tackle the reading problem along multiple fronts by combining multiple approaches, arguably offering better chances to have an impact.

Relative to the burgeoning literature on how to use education technology to improve learning in early years, we make two contributions.[7] *First*, unlike most existing literature on technology-driven interventions (e.g., Banerjee et al. 2007; Muralidharan et al. 2019), we go beyond the traditional exploration of the impacts on cognitive abilities, which is an outcome more easily observable to policy makers, and purposefully focus on an intervention that has a large potential to impact socioemotional skills and perceptions. We find gains in aspirations and self-confidence, as well as suggestive gains in locus of control, dimensions generally considered malleable at young ages (e.g., Almlund et al. 2011) but hard to change through education technologies (e.g., Escueta et al. 2020; Gortazar et al. 2024).[8] *Second*, we build upon existing evidence showing that personalized learning that teaches "at the right level" has the greatest potential to promote learning (e.g., Banerjee et al. 2016), and study the impacts of a tool that adaptively addresses reading difficulties of individuals throughout the whole ability distribution. This is particularly important because existing work suggests that tutoring programs tend to be most effective for those students starting from low initial levels (e.g., Beg et al. 2022).

Relative to the existing literature on the malleability of socioemotional skills during early life (e.g., Ashraf et al. 2020; Alan and Mumcu 2024), we provide complementary evidence of how a short, light-touch, and low-cost bundled program can improve socioemotional outcomes. We also find suggestive evidence that the program improves reading performance and that its benefits extend to students across the ability distribution. The latter is consistent with recent results by Alan et al. (2019) showing that, contrary to previous consensus (Sisk et al. 2018), it is possible to design interventions that help individuals throughout the whole distribution. However, the small samples available for the heterogeneity analysis prevent us from drawing firm conclusions. A further limitation of the present paper is that we are not able to isolate the impact of Dytective from that of the program's auxiliary components, namely the mobile library and the text messages to parents. Having said this, the fact that we, as discussed later, find that the program is very cost-effective even as a bundle of various elements indi-

[7] Note that the contributions described in what follows are made to the literature on the use of *general* education technologies. For examples of *gamified* interventions, which are a subset of education technologies see, for instance, Dillon et al. (2017) and Lafortune et al. (2024).

[8] A recent online tutoring intervention during the COVID-19 pandemic in Italy that was able to generate gains in aspirations, grit, locus of control, and well-being is Carlana and Ferrara (2021). This is conceptually a very different program from ours since, among other reasons, it offers individual tutoring for course-specific material while our student-specific tailoring is done through the app's algorithm and there is much less of a mentorship relationship with the invigilator.

cates that separating the relative contributions of each element could help design an even more cost-effective program.

Outline of paper. The rest of the paper is organized as follows. Section 2 explains the context and the intervention in more detail. Section 3 describes our data and empirical approach. Section 4 reports our main results and assesses their robustness. Section 5 discusses the cost-effectiveness and scalability of the intervention. Section 6 concludes.

2 Context and intervention

2.1 A Leer Jugando

We evaluate the impact of the bundled program A Leer Jugando implemented by Fundación Piñera Morel. This program is targeted at third-grade Chilean students enrolled in schools catering to disadvantaged families (as measured by the Chilean Government's Educational Vulnerability Index—IVE by its Spanish initials). The program provides students with access to Dytective, an online gamified educational platform that offers over 42,000 linguistic exercises designed using natural language processing techniques to provide individualized training to improve the reading and writing skills of participants. The pool of exercises that the program draws from was developed over a decade of iterative design and field testing using identified patterns in reading and writing mistakes of dyslexic individuals.[9] The platform is displayed as a game where the main character needs to complete exercises to progress in their quest. Exercises are grouped into linguistic challenges, which are *adaptively* generated based on the player's past performance in the application, with a focus on weaker linguistic areas and general cognitive abilities, e.g., working memory and attention. Each challenge takes around 20 min to complete.

Although Dytective was initially created to aid dyslexic individuals, non-dyslexic students can also benefit by helping them build their vocabulary, improve their spelling, memory, and reading speed, and by strengthening their ability to pay attention to and focus on reading tasks. Dytective also features a back-end for school therapists that helps them monitor the progress of the student along three main executive functions (simultaneous attention, activation and attention, and sustained attention) and seven performance measures (error correction, reading comprehension, reading speed, natural spelling, arbitrary spelling, writing speed, and error recognition). It also provides a screening test that allows to get a fairly accurate prediction of the likelihood of having dyslexia within just 15 min and at a very low cost.[10] For more details on the characteristics of Dytective and the screening test, the reader may refer to Rello et al. (2017, 2020).

[9] For more details on the personalization of the challenges, see Appendix Sect. 1.

[10] The screening test integrated in the tool is a machine learning-based model that predicts risk of reading difficulties in general, not specifically of dyslexia (Rello et al. 2020). However, dyslexia is the most frequent reading disorder (and its formal diagnosis still has to be done by a professional).

A Leer Jugando has been active since 2022. In our evaluation, we focus on an implementation of the program during the second semester of the academic year 2023.[11] For three months, an educational psychologist visits students in their school three times per week during regular class hours, provides them with individual tablets to access their personal Dytective profile, and guides their use of Dytective throughout 45-minute-long sessions.[12] This is regarded as a regular school activity so all students participate in it. In a typical session, students work on their personalized challenges independently, and the facilitator is around to provide encouragement and to solve questions, if needed. Following Dytective's recommendations, students aim to complete two challenges per session.[13]

A Leer Jugando also features a mobile library that allows students to borrow books and reading-related games to take home once per week as well as a parental support component through which FPM offers tips, via short text messages shared weekly in a WhatsApp group, on how parents could take advantage of daily life situations to help and motivate their children with their reading and writing. Appendix Fig. 1 provides an example of how the text messages look like. Appendix Fig. 2 shows how students work individually in class with Dytective, and displays the appearance of both the interface of the game and the back-end recording the evolution of the participant along various reading and cognitive skills dimensions.

2.2 Study design

We evaluate the impact of the program on students enrolled in third grade at five schools in high-vulnerability areas of the Chilean Metropolitan Region that participated during the second semester of 2023. The Metropolitan Region, which includes Santiago, agglomerates most commercial and administrative centers of the country and is home to around 40% of the country's population. Though our implementing partner would like to extend the program to all schools catering to vulnerable children, the program is implemented in small batches due to financial and logistic constraints. Each batch tends to be locally clustered to optimize on FPM's resources (e.g., the facilitator's commuting time).

At the time that we initiated our research collaboration with FPM, the five schools that were to receive the treatment during the second semester of 2023 had already been identified.[14] As such, we were not able to design the evaluation through a fully randomized controlled experiment. Discussion with our implementing partner made clear that their objective is to eventually offer the intervention to as many schools as they possibly can, so the order in which schools get access to the program primarily

[11] Appendix 1 provides more details on the timeline of the intervention.

[12] The amount of Spanish language classes for Grade 3 students in Chile is regulated to be of four blocks of 90 min per week. A Leer Jugando was implemented for half of the 90-minute blocks on three different days. One potential concern is that the Dytective sessions crowd-out regular instructional time. We note that, if anything, this would make it less likely for us to detect positive impacts of the program on reading performance.

[13] Occasionally, some students finish the two challenges before the end of the session. In such situation, they are encouraged by the facilitator to do silent reading.

[14] Appendix 1 provides further details on school selection.

reflects practical considerations rather than anticipated gains. This allows us to estimate treatment effects on the five schools through an approach that mimics a matched pairs design (Bruhn and McKenzie 2009). For each of the five schools that were to be treated during the second semester of 2023, we searched across the full pool of schools in the Metropolitan Region to identify another school that resembled each treated school the most in terms of the educational vulnerability index, size (number of students enrolled), and location (same or nearby communes). These three school-level characteristics are the same dimensions set by our implementing partner as criteria to determine program participation.[15] The educational vulnerability index is a particularly attractive dimension to consider since the Chilean government relies extensively on it to determine access to public support (e.g., the Preferential School Subsidy, "Ley 20.248") and is very predictive of student sorting Valenzuela et al. (2014), which helps us to account for differences in family characteristics across schools. Matched schools serve as controls for the treated schools.

To encourage the control schools to participate in the data collection process by distributing our surveys, FPM committed to including them in the subsequent implementation batch of A Leer Jugando. This strategy proved successful, as all the five schools that we approached to act as controls agreed. These control schools were equally convenient for our implementation partner as the treated ones, but had not been selected for the implementation in the second semester of 2023 for budgetary reasons. Conceptually, this supports our identification assumption that, in other states of the world, the control schools would have had the same chance as the treated schools to be selected during the time frame we are interested in, therefore making treatment status within each match pair be plausibly random. Still, we highlight once again that our causal estimates are contingent on the assumption that there are no unobservable school-level differences—such as institutional readiness or leadership quality—that determined the timing of the rollout.

Description of program protocols. We do not alter any of the elements of A Leer Jugando, so that we evaluate the exact program that is offered outside our study. As per the program's design, *all students* in grade 3 of the treated schools were subject to the Dytective and the mobile library components of the intervention during regular school time, while no student in the control schools had access to either.[16] However, whether parents received the text messages with tips depended on them voluntarily joining the WhatsApp group after having been informed about its existence in a regular

[15] Given the features of the intervention, we do not have ex-ante priors on which schools catering to vulnerable populations may benefit the most from such a program. In fact, we expect that a large number of schools beyond the ones in our study would equally benefit from such an intervention.

[16] We verify that this is indeed the case by reviewing the profile of each student in the Dytective application. Every student in the treatment schools that remained enrolled until the end of the intervention completed multiple challenges throughout the three-month period, whereas none of the students in the control schools completed any. More specifically, the average number of challenges completed by students in treated schools is 32, the 10th percentile is 17, and the 90th is 49. Moreover, our implementing partner reported a high turnover of materials from the mobile library.

teacher–parents meeting prior to the start of the program.[17] As stated later, we will interpret our estimates of the impact of the program as capturing the intention-to-treat effects of the intervention.

3 Data, identification, and estimation

3.1 Data

We combine data from three sources, merging them at the student-level. First, we obtain secondary data reported by the school administration on academic performance in standardized tests. Second, we collect primary data on noncognitive skills, attitudes, and beliefs through an in-school survey. Lastly, we obtain a measure of each student's risk of dyslexia through the screening test developed in Dytective.

Reading performance. We obtain measures of reading performance from the "Diagnóstico Integral de Aprendizajes" (DIA), a standardized testing tool crafted by the Education Quality Assurance Agency (Agencia de Calidad de la Educación) of the Ministry of Education of the Chilean government that aims to capture third graders' ability to locate and interpret information, and to reflect on a text's content. This standardized test is distributed three times per year, including both in August and late November/early December, hence conveniently offering pre- and post-intervention measurements. The score is in a 0–100 scale.

Socioemotional outcomes: Skills, attitudes, and beliefs. To complement our measure of reading performance, we designed computer-based surveys to be distributed to all students in treatment and control schools both before the intervention and at its conclusion. The goal of this survey is to measure a wide range of socioemotional skills and attitudes of students that we expected to be malleable after an intervention of this kind (e.g., self-confidence and taste for school). These surveys were filled up during class time by all students present at school on the day of the delivery.[18]

Socioemotional skills are difficult to measure. In this paper, we use self-reported measures to elicit various socioemotional aspects we are interested in. We acknowledge that self-reported measures may suffer from issues of validity (whether the tools sufficiently capture the dimension of interest) and reliability, especially when respondents are younger. As in our case of 8-year-olds, children may be more prone to acquiescence and social desirability biases, as well as cognitive fatigue, all of which

[17] Schools organize parent–teacher meetings regularly throughout the school year. In the meeting just before the implementation of A Leer Jugando, our implementing partner had 15 min to present the program to the attendants and requested their voluntary inclusion into the WhatsApp group. FPM's records show that 60% of the children had one caretaker that belonged to the group.

[18] The facilitators (acting also as enumerators) only had access to the control schools once at baseline and once at endline, whereas they were allowed to return to treatment schools twice at baseline and at endline. This naturally leads to treated students being more likely to be observed at endline. As we discuss in Sect. 4.2, this is unlikely to be a concern for selective attrition. However, it could still be a threat if increased familiarity between the students and the facilitators in the treatment schools led to encouragement effects. We argue in Sect. 4.3 that this is also unlikely to be a major concern in our context.

can introduce measurement error. Moreover, heterogeneity in reading comprehension and in the abstract reasoning required to assess internal states may reduce the reliability and stability of self-reported responses. To mitigate these concerns, we rely on instruments that have been used in similar primary school-aged children by large and long-standing surveys such as the OECD's PISA and Oxford's Young Lives study (Organisation 2023; Sanchez et al. 2016). In addition, given the young age of the respondents, during survey collection, students were aided by the enumerators to make sure that questions were clearly understood and responses were properly recorded. In particular, the enumerators explained each survey item to the whole class and solved any doubts, and then requested the students to submit their response before proceeding to the next item. As such, the quality of responses is very high and, conditional on a student being present at school on the fielding day, all survey items were completed.

Most of the survey questions elicited agreement with statements on a Likert scale. The options in the five-point scale were "not at all," "a bit," "somewhat," "quite a bit," and "a lot." When appropriate, we reverse the scale to make the individual items within a family point toward the same direction (i.e., increasing values reflect better outcomes). We then build indices following Anderson (2008) to better capture latent characteristics and to deal with the measurement error in any individual item. The final outcome is an index for each family that is standardized to have a mean of 0 and a standard deviation of 1 for the control students. In the case of "families" of only one element, we simply standardize that variable to have a mean of 0 and a standard deviation of 1 for the control students. We focus on eleven families of primary outcomes, and use the following variables for their construction:

1. **Academic aspirations.** One question: Up to which academic level would you like to study? The options provided were: "until completing middle school," "until completing high school," and "until completing university."
2. **Self-perceived performance relative to peers.** Three questions asked: If you compare yourself to your classmates in math/language/reading, how well do you think you perform? Answers are on a five-point scale with options: "much worse," "a bit worse," "about the same," "a bit better," and "much better."
3. **Perceived easiness of courses.** Three questions: How much do you agree that math/language/reading is hard?
4. **Taste for academic subjects and for school.** Four questions in total. Three questions asked about how much the respondent likes math/language/reading. Respondents were also asked if they like attending school (answers followed the same categories as when asking for the level of agreement with a statement).
5. **Grit.** Four questions asking the level of agreement with the following statements: I like that homework is challenging even if that means that I make mistakes; I give up easily if I cannot reach my objectives; if I think I am going to lose in a game I prefer not to continue playing; and if I do not know how to do something it is a waste of time to keep trying.
6. **Locus of control.** Five questions asking the level of agreement with the following statements: If I try enough, I can improve my academic performance; no matter how much I have studied for an exam, if I have bad luck I will perform poorly; whenever I set goals for myself I feel confident I will reach them; I like to make

plans about my future; and I usually think about my future goals and in the steps needed to achieve them.

7. **Individual well-being.** Seven questions in total. Respondents state their level of agreement with the following statements: I feel happy; many things worry me; I feel sad; I get angry easily; oftentimes I do not feel like doing anything; oftentimes I feel I do things wrong; oftentimes I have problems focusing.
8. **Social well-being.** Three questions in total. Respondents state their level of agreement with the following statements: I feel lonely; my classmates treat me with respect; I feel safe at school.
9. **Effort on weekdays.** Time devoted to studying on a normal weekday. Options were: "no time," "1–15 min," "16–30 min," "31 min–1 h," and "over an hour."
10. **Effort on weekends.** Time devoted to studying on a normal weekend. Same options as for weekdays.
11. **Parental investment.** Two questions in total. As an additional outcome, and to help us better understand potential mechanisms, we look into measures of parental investment in the child. We exploit information on how much students report that their parents (i) help them with school work and (ii) worry about their academic performance. Answers were, once again, elicited on a 5-point scale: "nothing at all," "a bit," "somewhat," "quite a lot," and "a lot." We follow the same approach as for the main outcomes to construct an index of "parental investment."

A potential concern with these indices is that they may reflect idiosyncratic responses or measurement error rather than meaningful latent constructs. In Appendix 1, we assess the psychometric validity of the constructed indices through Cronbach's alpha and average inter-item correlations. Overall, the psychometric diagnostics are reassuring particularly for those indices for which we later find meaningful and robust effects.

Risk of dyslexia. We distributed Dytective's screening test to obtain a pre-intervention measure of the risk of dyslexia of each student. As mentioned, although Dytective is equipped to also help non-dyslexic students, particularly at low levels of reading ability, it was originally designed for children with dyslexia. As such, we expect at least some of the effects to be more pronounced among at-risk-of-dyslexia individuals. To explore this hypothesis, we employ the score in the screening test—a continuous measure theoretically ranging from 0 to 100—to investigate heterogeneity in treatment effects.

Response coverage. According to official school census records, at the start of the intervention, a total of 868 students across control and treatment schools were enrolled in third grade. 715 of them (82%) completed our baseline survey. This is a high proportion and aligns well with the fact that around 15-20% of Chilean students are flagged by the Ministry of Education as high-absenteeism students (i.e., attend less than 85% of the classes). A total of 527 students also completed the endline survey and can therefore be used to quantify the socioemotional impacts of the intervention. The fact that our surveys were completed during school hours helped to keep the attrition rate at comparable levels to those faced by successful interventions in similar contexts (e.g., Muralidharan et al. 2019; Carlana and Ferrara 2021). The reading test

scores from DIA are available at both baseline and endline for 368 out of the 527 students in our main estimating sample (70%). The main reason why this information is not available for some students is, once again, absences on the day of the exam.[19] Reassuringly for our analysis, we show below that attrition does not lead to unbalanced characteristics between treatment and control at baseline.

Descriptive statistics. Descriptive statistics of the main variables of interest for the sample employed in our estimations of the treatment effects are provided in Appendix Table 5.[20] For instance, in terms of background characteristics, 45% of the sample are males and 9% have repeated at least a grade level. The average score in the screening test is 0.197. For the study of heterogeneity in treatment effects, we will employ a measure of high risk of dyslexia that involves being in the top 15% of the continuous score of risk delivered by the test.[21] 72 students are identified as high-risk.

The table also highlights in bold the indices of interest (which are centered at a mean of zero and have a standard deviation of 1 for the control group at baseline prior to sample selection) and, below them, we show the descriptive statistics of the raw variables used to construct each of them. For example, we see that the average agreement to the statement that “math is easy" is 3.59 on a 1–5 scale.

3.2 Identification and estimation

Given our study design, the natural identifying assumption is that of conditional ignorability. More precisely, within each match pair, the students in treated and untreated schools have the same potential outcomes. In the absence of treatment, endline outcomes are similar across schools so that the students in untreated schools can serve as valid controls. Though our design makes conditional ignorability most plausible at the school level, we estimate our treatment effects using individual data to obtain more precise estimates. To maximize the plausibility of the conditional ignorability assumption at the individual level, we further condition on the pre-treatment level of all the outcomes that we measure (allowing for dynamics in the evolution of the outcomes), and other individual covariates that may determine potential outcomes. In Sect. 4.2, we study the robustness of our conditional unconfoundedness assumption.

Under our main identification assumption, we choose to estimate the effects based on regression adjustment—in Sect. 4.2 we also show robustness to this choice by estimating models using propensity score matching. Thus, our baseline regressions are of the form:

[19] The availability of the survey responses and the DIA test scores may not necessarily coincide because the survey and DIA test were conducted on different days.

[20] The counterpart for the full sample available at baseline (i.e., regardless of attrition at endline) is provided in Appendix Table 6. As one can see, the descriptive statistics are very similar, and the attrition rate of the 715 students that completed the baseline survey was 1 - 527/715 $\approx$ 26%.

[21] This fraction represents the estimated fraction of dyslexic individuals worldwide (e.g., Shaywitz 1998). Although a recent meta-analysis by Yang et al. (2022) reports a somewhat lower prevalence, using the higher figure makes our cutoff deliberately conservative. Increasing the threshold would reduce the likelihood of detecting a difference, not increase it.

$$y_{i2} = \beta \times \text{treated}_{s(i)} + \theta \times y_{i1} + X_i'\gamma + \delta_{p(s(i))} + \varepsilon_i, \quad (1)$$

where an outcome y for individual i at school s, measured at the end of the intervention (as indicated by the subscript 2), is regressed on an indicator of the school being in the treatment group, the baseline measure of the outcome variable (indicated by the subscript 1), and match pair fixed effects δ (indexed by p and only dependent on which school the individual goes to). Vector X_i contains the baseline values of the other outcomes of interest and other individual-level characteristics that likely determine potential outcomes: gender, repeater status, age, initial risk of dyslexia, and month of survey completion. We report wild cluster bootstrapped p-values which Cameron et al. (2008) show may provide reliable inference even with as few as five clusters. We cluster at the school-level. We further implement the 6-point bootstrap weight distribution proposed by Webb (2023) to obtain more precise p-values. As recommended by Mackinnon et al. (2023a), we report tests based on alternative inference procedures in Sect. 4.2 to show robustness.

This regression adjustment specification coincides with the so-called value-added specification that is common in models of human capital production and in the evaluation of educational interventions (Todd and Wolpin 2003; Andrabi et al. 2011; Singh 2015). By controlling for the outcome measured at baseline, we not only allow for imperfect balance in these characteristics between control and treatment schools which improves precision, but also allow for dynamics in learning. Moreover, these value-added specifications are found to estimate robust treatment effects in frameworks with non-random assignment or alternative dynamics and, in particular, difference-in-differences settings where there is non-random selection (Guarino et al. 2015; O'Neill et al. 2016).

Given the design of the program, our estimates are best interpreted as intention-to-treat (ITT) effects of the intervention. Conditional on school attendance, there is full compliance among treated participants in the main component of the intervention, Dytective, as these sessions were done in class and the protocol for the students' usage of Dytective was standardized and supervised by a facilitator. However, there may be variation in the take-up of the two complementary programs. First, though the mobile library is open to all students, usage is dependent on the students' willingness to borrow items. Second, not all of the parents signed up to receive the text messages with tips for helping their children with their reading. Consequently, our coefficient of interest, β, identifies the ITT effect of offering the multi-component intervention on the outcomes of interest.

Match pair balance in observables. Our identification assumption that potential outcomes are similar within each match pair is fundamentally untestable. We can, however, show that there are no within match pair differences in our primary outcomes and key background controls for our estimating sample at baseline. This is suggestive that, within a match pair, our conditional unconfoundedness assumption is plausible inasmuch as the unobservables that determine potential outcomes are correlated with these observed characteristics. Table 1 reports these comparisons. We find that the differences in average characteristics are not statistically significant (except

Table 1 Balance checks

	(1) Control		(2) Treatment		(1)-(2) Pairwise t-test	
Variable	N	Mean/(SD)	N	Mean/(SD)	N	Beta/[Wild bootstrapped *p*-value]
Male	254	0.457	273	0.436	527	0.013
		(0.499)		(0.497)		[0.464]
Repeater	254	0.098	273	0.073	527	−0.030
		(0.298)		(0.261)		[0.336]
Screening score	254	0.193	273	0.200	527	0.004
		(0.074)		(0.071)		[0.708]
Age (years)	254	8.594	273	8.548	527	−0.130*
		(0.587)		(0.548)		[0.080]
Index: aspirations	254	0.042	273	0.107	527	0.160
		(0.979)		(0.940)		[0.266]
Index: perceived performance relative to peers	254	−0.011	273	0.069	527	−0.039
		(0.997)		(1.004)		[0.744]
Index: finds courses easy	254	0.047	273	0.206	527	0.070
		(0.905)		(0.882)		[0.428]
Index: like school courses	254	−0.014	273	0.136	527	0.085
		(0.969)		(0.934)		[0.872]
Index: grit	254	0.001	273	0.202	527	0.214
		(1.003)		(1.037)		[0.346]
Index: locus of control	254	−0.005	273	0.098	527	0.153
		(0.997)		(0.969)		[0.214]
Index: individual well-being	254	0.068	273	0.064	527	−0.018
		(0.990)		(1.036)		[0.698]
Index: social well-being	254	0.035	273	0.184	527	0.230
		(0.987)		(0.968)		[0.388]
Index: study workdays	254	0.032	273	0.128	527	−0.055
		(1.010)		(0.980)		[0.790]

Table 1 continued

	(1) Control		(2) Treatment		(1)-(2) Pairwise t-test	
Variable	N	Mean/(SD)	N	Mean/(SD)	N	Beta/[Wild bootstrapped *p*-value]
Index: study weekends	254	0.063	273	0.028	527	−0.141
		(1.019)		(0.957)		[0.366]
Index: parental investment	254	0.004	273	−0.089	527	−0.081
		(0.950)		(0.972)		[0.524]
Reading test score	150	59.356	218	56.445	368	−0.682
		(19.245)		(21.596)		[0.880]

The table documents, for the main predetermined variables, indices, and the reading score their mean and standard deviation (SD) separately for the treatment and control subsamples. "N" stands for the number of individual observations. The last column reports the difference in means (after controlling for strata and date of survey fixed effects). In brackets, we report wild cluster bootstrapped (clustering at the school-level) p-values using Webb (2023)' s 6-point bootstrap weight distribution. *** p<0.01, ** p<0.05, * p<0.1

for age at the 10% level) and are generally small in economic magnitude.[22] This provides further evidence toward plausible as-good-as-random assignment of treatment between schools within match pairs. We acknowledge, however, that the absence of statistically significant differences may partly be a consequence of low power to detect these differences. Therefore, all our regressions control for the baseline value of the variables in Table 1 except the reading test score (to maximize the size of the samples used in the main analyses).

Internal and external validity. Under the maintained identification assumptions from above, we obtain valid average treatment effect estimates for the five treated schools. For our estimates to be externally valid—that is, for our average treatment effect to be a valid estimate of the treatment effect across a larger population, say, students in all schools in the Chilean Metropolitan Region catering to vulnerable populations—we would need to make the additional assumption that the five schools in our study are representative of some larger population. Generalization is a common issue for interpreting any exercise in causal inference, including randomized controlled trials (Duflo et al. 2007). Though the treated schools in our analysis were chosen by mere convenience and we do not suspect treatment effects to be specific to these schools, it could still be that students in locally clustered schools share similar characteristics that make them more or less responsive to the treatment. In light of this, a conservative interpretation of our results would be the qualitative conclusion that, at the bare min-

[22] Table 7 replicates the analysis for all the observations available at baseline irrespective of their future attrition status. We find a consistent picture of the lack of initial differences between the treatment arms. Table 8 proceeds similarly but focuses on the subsample for which information on reading performance is available. The same take-aways hold.

imum, our intervention is beneficial for our subpopulation of students and, given its cost-effectiveness, more financial resources and time should be invested to studying its impacts on more general populations.

4 Results

This section reports the main results of the paper. We first quantify the treatment effects of our intervention on reading and on socioemotional performance. We then probe the robustness of our results to potential threats to identification. We end this section with a discussion on our findings.

4.1 Treatment effects

Impact on academic reading performance. We first examine the effectiveness of the reading intervention in improving reading performance. Column (1) in Table 2's Panel (a) shows that, for the subsample for which reading scores are available, there is a positive and statistically significant effect of the intervention on reading performance. In particular, treated students obtain 2.59 points more in the test on average, a sizable effect of $2.59/19.25 \approx 13\%$ of the control group's standard deviation.

Impact on socioemotional outcomes and educational investments. Table 2's Panels (b)–(e) report the estimates of the impact of the program on socioemotional outcomes, time allocations, and parental investments. We detect clear gains in self-perception, particularly in our indices for aspirations, performance relative to peers and the perception that courses are easy. In particular, treated individuals display 10.8, 17.6 and 30.5% of a standard deviation higher values on average in these three dimensions, respectively. We find suggestive evidence that there are also gains in terms of personality traits (locus of control) and social well-being. We do not detect sizable increases in investments, whether by the child or by the parents.

Heterogeneity. We explore whether the effects of the intervention differ depending on the risk that the student is dyslexic and gender. As described below, we find some heterogeneity along these dimensions, but we caution on the interpretation of the results as, given our design, some subpopulations may be small which leaves us with little power to make precise conclusions. Future study with larger sample sizes would be needed to draw firm conclusions.

Columns (2) and (3) in Table 2 report the effects for the not-at-risk and at-risk groups, respectively. We find that the effects documented for the pooled sample are present among those who are not at risk. Moreover, we have suggestive evidence that the program had larger effects on at-risk students, though they are not precisely estimated.

Table 2 Estimated effects on reading performance and socioemotional outcomes

		By risk of Dyslexia		By gender	
	(1)	Not-at-risk (2)	At-risk (3)	Female (4)	Male (5)
Panel (a): Reading performance					
Average reading performance	2.594**	1.951	5.125	4.712*	0.980
	[0.038]	[0.106]	[0.466]	[0.096]	[0.804]
# Observations	368	315	53	212	156
Panel (b): Self-perceptions					
Aspirations	0.108*	0.072	0.246	0.183	0.091
	[0.050]	[0.220]	[0.540]	[0.488]	[0.238]
Performance rel. peers	0.176***	0.105**	0.331	0.225	0.162
	[0.008]	[0.040]	[0.190]	[0.424]	[0.428]
Courses are easy	0.305**	0.283*	0.340	0.337***	0.341
	[0.014]	[0.054]	[0.186]	[0.004]	[0.182]
Like school	0.160	0.234	−0.144	0.278	0.083
	[0.370]	[0.304]	[0.528]	[0.348]	[0.830]
# Observations	527	455	72	292	235
Panel (c): Personality traits					
Grit	0.247	0.271	0.532	0.218	0.240
	[0.550]	[0.514]	[0.154]	[0.654]	[0.594]
LOC	0.203*	0.129*	0.710**	0.169	0.282
	[0.060]	[0.066]	[0.034]	[0.396]	[0.144]
# Observations	527	455	72	292	235
Panel (d): Subjective well-being					
Individual well-being	0.034	0.119	−0.410	0.060	−0.011
	[0.586]	[0.204]	[0.154]	[0.408]	[0.884]
Social well-being	0.193	0.213*	−0.071	0.418**	−0.018
	[0.122]	[0.098]	[0.824]	[0.012]	[0.838]
# Observations	527	455	72	292	235
Panel (e): Time and investment					
Study workdays	0.025	−0.001	0.167	−0.072	0.154*
	[0.540]	[0.970]	[0.342]	[0.488]	[0.088]
Study weekends	0.100	0.105	0.455***	0.038	0.136*
	[0.312]	[0.328]	[0.004]	[0.856]	[0.066]
Parental investment	0.103	0.158	−0.174	−0.063	0.309

Table 2 continued

		By risk of Dyslexia		By gender	
	(1)	Not-at-risk (2)	At-risk (3)	Female (4)	Male (5)
	[0.284]	[0.334]	[0.230]	[0.384]	[0.404]
# Observations	527	455	72	292	235

Regressions estimate Equation (1). The set of individual controls includes the baseline value of all the eleven indices listed in Sect. 3.1, together with controls for gender, repeater status, age, initial risk of dyslexia, and month of survey. Reading performance is measured on a 0–100 scale. All other outcomes are z-scores. In brackets, we report wild cluster bootstrapped (clustering at the school-level) p-values using Webb (2023) 's 6-point bootstrap weight distribution. *** p<0.01, ** p<0.05, * p<0.1

In Columns (4) and (5) of Table 2, we report the effects depending on the gender of the student. In certain dimensions, such as reading performance, we find suggestive evidence that the effects are larger for female students. In other dimensions, the estimated effects are similar between male and female students. This is most evident among some of the socioemotional dimensions for which we had found sizable effects based on Column (1) of the same table: self-perception that courses are easy and locus of control.

In Appendix Tables 9 and 10, we complement the previous analysis by reporting a more parsimonious specification in which we interact the treatment indicator with an indicator of being at risk of having dyslexia and of being a male, respectively. The key takeaways remain. The coefficients corresponding to the treatment indicator are of similar magnitude to the estimates we find in Column (1) of Table 2. Moreover, we find that the interactions for some of the dimensions are economically significant but are not statistically significant. This suggests that the effects we find are likely common among at risk and not at risk as well as among males and females.

4.2 Robustness

We acknowledge that due to budgetary and logistical constraints, our study design has limitations that makes it fall short of the ideal environment for causal identification. However, in this section, we argue that the positive treatment effects described above for certain dimensions are unlikely to be spurious. First, we show the robustness of our results to the two most salient threats to internal validity: selective attrition and selection on unobservables. Then, we address the robustness of the results to issues regarding inference, including multiple hypothesis testing.

Attrition. Unrelated to our best efforts to keep high quality of the survey responses with the help of the enumerators, some attrition in the completion of the endline survey is to be expected. This is particularly so given that Chile has documented high levels of student absenteeism (e.g., Arbour et al. 2023) and that our endline surveys were not pre-announced. That said, and as explained above, out of the 715 students who completed the baseline survey, 527 (74%) also completed the endline survey comprising

Table 3 Robustness of estimated effects on cognitive and socioemotional outcomes

	Att (1)	No controls (2)	Degrees of freedom correction (3)	CV_3 (4)
Panel (a): Reading performance				
Average reading performance	1.963**	3.197*	2.594	2.594
	[0.022]	[0.080]	(1.448)	{2.438}
# Observations	368	416	368	368
Panel (b): Self-perceptions				
Aspirations	0.123**	0.157	0.108**	0.108**
	[0.022]	[0.362]	(0.026)	{0.033}
Performance rel. peers	0.163**	0.208	0.176***	0.176**
	[0.014]	[0.434]	(0.022)	{0.077}
Courses are easy	0.303**	0.325*	0.305***	0.305**
	[0.018]	[0.050]	(0.030)	{0.091}
Like school	0.163	0.178	0.160	0.160
	[0.362]	[0.500]	(0.199)	{0.541}
# Observations	527	527	527	527
Panel (c): Personality traits				
Grit	0.244	0.266	0.247	0.247
	[0.566]	[0.474]	(0.257)	{0.545}
LOC	0.200*	0.312**	0.203***	0.203***
	[0.068]	[0.038]	(0.033)	{0.042}
# Observations	527	527	527	527
Panel (d): Subjective well-being				
Individual well-being	0.030	0.141*	0.034	0.034
	[0.680]	[0.058]	(0.120)	{0.248}
Social well-being	0.197*	0.254	0.193*	0.193
	[0.064]	[0.338]	(0.080)	{0.276}
# Observations	527	527	527	527
Panel (e): Time and investment				
Study workdays	0.018	−0.054	0.025	0.025
	[0.672]	[0.400]	(0.064)	{0.081}
Study weekends	0.092	0.049	0.100	0.100
	[0.310]	[0.614]	(0.076)	{0.134}
Parental investment	0.133	0.140	0.103	0.103

Table 3 continued

	Att (1)	No controls (2)	Degrees of freedom correction (3)	CV_3 (4)
	[0.260]	[0.386]	(0.107)	{0.307}
# Observations	527	527	527	527

"Att" replicates the analysis in Table 2's Column (1) weighting each observation by the inverse of the probit-based propensity to remain in the sample at endline. "No controls" also replicates the same column but excludes the set of individual controls. The number of observations in Panel (a) increases because we no longer require that the baseline cognitive score is available. In Column (3), we reproduce the point estimates in Table 2's Column (1) but report in parentheses bias-corrected standard errors with a degrees of freedom adjustment. In Column (4), we reproduce the point estimates in Table 2's Column (1) but report in curly braces cluster-robust jackknife-based (CV3) standard errors following Mackinnon et al. (2023a). Reading performance is measured on a 0–100 scale. All other outcomes are z-scores. In brackets, we report wild cluster bootstrapped (clustering at the school level) p-values using Webb (2023)'s 6-point bootstrap weight distribution. *** $p<0.01$, ** $p<0.05$, * $p<0.1$

socioemotional outcomes, which is a considerably high response rate. Matching rates do fall down to 368 (51%) for those who completed both the endline survey and the two rounds of the reading test. We find that students in treated schools are 7.3 percentage points less likely to attrite than students in control schools, though the difference is not statistically significant (wild cluster bootstrapped p-value = 0.168). The presence of such a gap is unsurprising because, as previously stated, the enumerators were able to visit the treated schools multiple times during the implementation period, while visits to control schools were limited to once at baseline and once at endline.[23] Following Muralidharan et al. (2019), we re-estimate our main specifications weighting each observation by the inverse of its probability of remaining in the sample at endline. Column (1) in Table 3 reports these inverse attrition probability weighted estimates. We find quantitative and qualitatively similar results.

We also perform a bounding exercise following Lee (2009). We report the lower and upper bounds in Appendix Table 11. Typical of this exercise, the bounds are quite wide. Still, we find that the lower bounds for the effects on the self-perceived easiness of the subjects and on locus of control are negative but very small in absolute value. This suggests that less conservative assumptions on the nature of attrition would at worst lead to null results.

A related concern is that the cognitive and socioemotional effects are estimated on different samples (the reading test is available for 368 students, whereas the socioemotional outcomes are available for 527). To assess whether our socioemotional findings are sensitive to this difference, in Table 12 we re-estimate the full socioemotional analysis on the restricted sample of 368 students for whom the reading score is also available. The results preserve the broad pattern of our main findings. The effects on course easiness and locus of control remain positive, although less precisely estimated

[23] The more intensive interaction between FPM and treated schools, together with the regular presence of facilitators, primarily served logistical purposes—namely, tracking student attendance and reducing survey attrition—rather than addressing implementation challenges or managing specific incidents.

Table 4 Robustness to selection on unobservables

Outcome variable	(1) Oster (2019)'s δ	(2) Masten et al. (2024)'s c_{BP}
Panel (a): Reading performance		
Average	0.508	0.025
Panel (b): Self-perceptions		
Aspirations	18.710	0.024
Performance rel. peers	3.319	0.034
Courses are easy	8.855	0.063
Like school	2.705	0.033
Panel (c): Personality traits		
Grit	− 3.057	0.044
LOC	1.819	0.050
Panel (d): Subjective well-being		
Individual well-being	− 0.817	0.007
Social well-being	− 8.115	0.028
Panel (e): Time and investment		
Study workdays	− 0.508	0.004
Study weekends	22.789	0.008
Parental investment	6.984	0.018

The first column reports the estimated δ following Oster (2019). The second column reports the estimated breakdown point of Masten et al. (2024)'s c-dependence analysis. Details on these two methodologies are provided in appendices 1 and 1.

in the smaller sample, while the estimates for study time and parental investment remain close to zero. The effect on individual well-being is positive and marginally significant, providing further suggestive evidence of gains in this dimension.

Selection on unobservables. Our main identifying assumption is that of conditional ignorability. A key concern is that there are unobservable factors that drive potential outcomes and that we are unable to properly control for. We perform two exercises to examine the robustness of our results to selection on unobservables.

First, we follow Oster (2019) and probe the robustness of our results based on linear regression adjustment. This analysis formalizes the concept of coefficient stability that is often seen as a robustness exercise. We detail the key aspects of this method in Appendix 1. In the spirit of assessing coefficient stability, we first report estimates of Equation (1) without student controls in Column (2) of Table 3. We find that the estimates generally have a similar magnitude as in the baseline specification but are estimated less precisely. However, as Oster (2019) argues, coefficient stability to addition of observable controls is not sufficient to argue for robustness against selection on unobservables (since the included controls could potentially add very little explanatory power to the regression). We report Oster (2019)'s δs in Column (1) of Table 4. These δs measure the factor by which selection on unobservables must be relative to selection on observables to lead to a point estimate of zero (i.e., an exact null

result). As a rule of thumb, a δ above 1 in magnitude would suggest robustness of our results as selection on unobservables must be as large as the selection on observables for our results to be zero. We find that the δ corresponding to the effects on reading performance is less than 1, suggesting that the result may not be robust to selection. Conversely, this analysis strengthens our conclusions on socioemotional outcomes. The δ corresponding to both self-perceptions and personality traits are well above 1 in absolute value.

Second, we expand on Oster (2019)'s analysis based on a linear framework to consider one that allows for nonparametric selection as proposed by Masten et al. (2024). We detail the key aspects of this method in Appendix 1. Masten et al. (2024) introduce the concept of conditional c-dependence that relaxes the conditional independence assumption we relied on for our main analyses. It is governed by a parameter $c \in [0, 1]$ such that conditional independence only holds partially for $c > 0$. In particular, we allow for the observed propensity score to deviate from the propensity score that allows for unobservables by at most c in absolute value. By relaxing conditional independence in this way, the treatment effects are only partially identified and we are able to obtain bounds. The breakdown points c_{BP} are the largest value of c such that the identified set still allows us to conclude about the sign of the treatment effects (i.e., bounds do not contain zero).

The estimated breakdown points are reported in Column (2) of Table 4.[24] To benchmark whether the breakdown points we find are large, we report in Appendix Tables 13 and 14 features of the distribution of changes in the propensity score when we leave covariates out one at a time for cognitive and noncognitive outcomes, respectively. The idea is that we want to compare the breakdown point c_{BP} to the observed changes from leaving out observed variables. We take the 90th percentile of these distributions as reference values. Comparing the c_{BP} corresponding to reading performance (0.025) to the reference values in Appendix Table 13, the breakdown point is higher than the reference value of only 3 of the 17 variables. Conversely, and consistently with the previous analysis, we find that our results on socioemotional outcomes are likely to be robust to selection on unobservables. The breakdown points for self-perceptions and personality traits are greater than most of the 90th percentile reference values. In particular, the ones corresponding to self-perception that courses are easy and to locus of control are higher than 13 and 11 (of 16) of the reference values, respectively.

In both analyses, we use selection on observables as a benchmark to assess the potential biases from selection on unobservables. This analysis, therefore, is more informative if the observables we include are able to explain a large portion of the variation in the outcomes. In Appendix Table 15, we report the R^2 from the regression without observable controls (restricted regression) and the regression that includes all the observable controls (unrestricted regression). We find that the added controls do indeed add substantial explanatory power without largely changing the estimated treatment effects. This provides additional confidence to the high threshold that selec-

[24] A byproduct of the analysis of Masten et al. (2024) is that we also obtain estimates using nonparametric propensity score matching under the same conditional ignorability assumption. The treatment effect estimates based on propensity score matching are reported in Appendix Table 17. We find that the point estimates are quantitatively very close to those we had obtained from linear regression adjustment in Table 2, which reassures us that the results are not dependent on the estimation method.

tion on unobservables needs to be relative to the selection on observables to render our results null.

Alternative inference. Mackinnon et al. (2023a) suggests that in cases where there a few number of clusters, inference based on multiple methods should be reported to assess robustness. In Table 2, we have reported wild cluster bootstrapped p-values using Webb (2023)'s 6-point bootstrap weight distribution. In Column (3) of Table 3, we present bias-corrected clustered standard errors with a degrees of freedom adjustment. In Column (4) of Table 3, we report cluster-robust jackknife-based standard errors described in Mackinnon et al. (2023b) (denoted CV3 in Mackinnon et al. 2023a). Across all the inference methods we report, our results on self-perception and locus of control are statistically significant. Our preferred method for inference reported in the main analyses is conservative in most outcomes we consider.

Multiple hypothesis testing. Though we have already made a first attempt of reducing the number of outcomes by considering indices in the main analyses, one may worry that some of the effects we find arise spuriously as a consequence of a large number of tests across indices. To address this, we aggregate the components of Panels (b)–(e) in Table 2 and construct four general composite indices of socioemotional outcomes corresponding to the four panels. We estimate the treatment effect on each of these four general composite indices. The results are reported in Table 16 and reinforce the finding that participation in the program primarily improves self-perceptions, and that there are weaker but significant gains in well-being.

4.3 Discussion

This section discusses the main findings of the paper and situates them within the broader objectives of the study. We interpret the magnitude and robustness of the estimated effects, examine potential behavioral mechanisms, and consider why the intervention appears to have stronger impacts on socioemotional than cognitive outcomes. We also reflect on limitations and implications for future research.

Main findings. Our baseline analysis finds economically sizable and statistically significant effects of the intervention on reading performance and on self-perceptions and personality traits. However, the evidence for reading performance should be interpreted with some caution, as it is estimated on a smaller sample and may not be robust to selection on unobservables. By contrast, we find that the economically and statistically significant effects of the intervention on socioemotional outcomes, particularly on self-perceptions (i.e., perception that one is doing better than others and perception that courses are easy), appear robust to selection on unobservables.

Behavioral responses. We do not find corresponding changes in the time investments of students or parents. The absence of changes in the *quantity* of time investment does not necessarily imply the absence of behavioral changes. Our measures capture how much time is devoted to studying or helping with schoolwork, but they do not measure

the *quality* of that time. For students, with better socioemotional skills, the time spent studying may have been more effective. For parents, the intervention may have shifted how they interacted with their children around reading. In fact, this is precisely the type of behavioral margin targeted by the weekly text messages, which encouraged parents to use everyday situations as opportunities to support reading.

Cognitive vs. socioemotional skills. We find stronger evidence that our reading intervention changes socioemotional skills, most notably self-perception, but not necessarily reading performance. While this may be surprising, there are two compelling reasons why such a pattern may arise. First, existing evidence in other settings shows that socioemotional skills are often more malleable than cognitive skills in the short run (Carneiro et al. 2007; Cunha and Heckman 2008; Almlund et al. 2011). It is plausible that the improvements in socioemotional dimensions that we find serve as an early channel through which cognitive gains may materialize at later stages. Hence, further study needs to be done to measure the long-run effects of our intervention. Second, as mentioned previously, because the Dytective sessions were conducted during regular class time, the intervention could have displaced regular instructional time. To the extent that the displaced instruction would otherwise have improved reading performance, such crowding out would bias our estimates downward. Under this interpretation, our estimates may understate the program's cognitive impact relative to an implementation that does not displace regular instruction.

Potential adverse effects of gamification. While the preceding results focus on the potential benefits of the intervention, gamified educational programs may also generate unintended adverse effects. Prior research has highlighted concerns that gamification can negatively affect well-being (Hanus and Fox 2015) and contribute to attention problems (Skulmowski and Kate Man 2022). So far, we find no evidence of an average adverse effect of the bundled program on well-being, and some specifications suggest positive effects. Although we did not directly measure attention, the suggestive evidence on reading performance improvements indicates that any adverse effects operating through reduced attention are unlikely to dominate the intervention's overall impact.

Survey context and measurement of socioemotional outcomes. As previously discussed, the educational psychologists who facilitated the intervention also administered the surveys. While they visited treated schools multiple times during the implementation period, they only visited control schools twice, once for the baseline and once more for the endline survey. This differential contact raises the possibility that treated students became more comfortable with the facilitators or felt more inclined to provide responses that are perceived as socially desirable. While we cannot completely rule out this concern, several pieces of evidence suggest that our socioemotional findings are unlikely to be driven solely by these implementation features.

First, if differential contact induced a blanket social desirability or acquiescence bias, we would expect uniformly more favorable responses across the self-reported outcomes. Instead, the pattern is selective. We find the clearest effects on self-perceptions, particularly course perceptions, while the estimated effects on weekday and weekend

study time are small and statistically insignificant. Second, the self-reported socioemotional gains are accompanied by suggestive, positive gains in an objective cognitive outcome (the standardized reading test) which is not directly manipulable through socially desirable survey responses. Although the reading estimate is based on a smaller sample and is less robust to selection on unobservables, its positive direction provides complementary evidence that the survey results do not merely reflect a generalized tendency among treated students to give more favorable answers.

5 Cost-effectiveness and scalability

Cost-effectiveness. The multi-component program we evaluate costs €100 per student.[25] Whether this represents good value depends on how its benefits per euro compare with similar alternative interventions a policymaker could fund. We benchmark our program against the recent tutoring interventions studied by Carlana and Ferrara (2021) and Gortazar et al. (2024). The comparison is natural because all three interventions target disadvantaged students and are designed to complement, rather than replace regular schooling.

At the same time, two important differences should be kept in mind. First, both benchmark interventions rely on intensive, individualized *human* tutoring, whereas our program is delivered primarily through a self-guided and adaptive application. As a result, our intervention requires far less specialized labor and is therefore substantially easier to scale (as discussed below). Second, both benchmark interventions were implemented during the exceptional circumstances of the Covid-19 pandemic, with Gortazar et al. (2024) shortly afterward and Carlana and Ferrara (2021) taking place during the school closures themselves. Because control group students in those settings received limited in-person instruction, the resulting treatment–control comparison plausibly overstates the returns achievable under normal schooling conditions. By contrast, our estimates are obtained *on top of* regular in-person instruction—a more relevant policymaking context. We keep these similarities and differences in mind throughout the comparison, which we conduct under two scenarios.

Scenario 1: socioemotional gains only. The socioemotional improvements are the most robust benefits of our intervention. We estimate gains of 10–30% of a standard deviation across a range of skills, at a cost of €100 per student. On a per-€100-spent basis,[26] our estimated effects substantially exceed the 4% of a standard deviation increase in aspirations reported by Gortazar et al. (2024) and are broadly comparable to the 28% increase in grit and 34% increase in well-being reported by Carlana and Ferrara (2021). This comparison is particularly noteworthy given important differences

[25] This per-student cost includes both overhead and variable implementation costs and does not account for any tax exemptions that may be available to our implementing partner as a non-governmental organization. As a result, it may represent a conservative estimate of the effective cost of delivery.

[26] To facilitate comparisons, we rescale all reported effects to a common expenditure of €100 per student. For example, an intervention costing €50 per student is assigned an effect equal to twice its reported estimate. This normalization assumes a linear relationship between spending and impacts. If, as is often the case, marginal returns to educational investments are decreasing, the resulting comparisons are conservative: They may overstate the benefits that lower-cost interventions would achieve if additional resources were used to raise spending to €100 per student.

in delivery and context. Whereas both benchmark interventions rely on intensive one-to-one human tutoring, our effects are generated through a low-cost, largely self-guided intervention. Moreover, the estimates in Carlana and Ferrara (2021) were obtained during the exceptional circumstances of Covid-19 school closures, while our effects are achieved on top of regular face-to-face schooling. Taken together, these comparisons suggest that our intervention delivers socioemotional improvements that are highly competitive with the best-performing alternatives while relying on a model that appears substantially easier to scale.

Scenario 2: adding the suggestive reading gain. We do not find evidence that these socioemotional gains come at the expense of academic achievement. If anything, we find suggestive evidence of a reading improvement of 13% of a standard deviation, pointing to the possibility of both cognitive and socioemotional benefits from a single low-cost intervention. Such joint returns are relatively uncommon: Educational technologies often improve academic outcomes without affecting socioemotional skills, while stand-alone socioemotional programs frequently leave academic achievement unchanged. We emphasize, however, that the reading estimate should be interpreted with caution. Relative to our socioemotional outcomes, it is based on a smaller sample and is less robust to selection on unobservables (Sect. 4.2). Consequently, our assessment of the intervention's cost-effectiveness does not rely on this result. Nevertheless, taken at face value, the estimated reading effect lies between the per-€100 returns in academic performance reported by Gortazar et al. (2024) (8.7%) and Carlana and Ferrara (2021) (52%). This comparison is encouraging because the effects of the benchmark interventions are largely driven by improvements in mathematics, whereas our estimate pertains to reading, a domain that is generally considered more difficult to influence through educational interventions (Dietrichson et al. 2021). Thus, while tentative, the evidence suggests that the intervention may generate academically meaningful gains in addition to its more robust socioemotional benefits.

Scalability. We distinguish between two dimensions of scalability: *component scalability*—the ease with which the intervention's constituent elements can be deployed broadly—and *implementation scalability*—the extent to which program fidelity and participant engagement can be maintained under large-scale delivery. We discuss each in turn.

Regarding component scalability, our intervention is composed of three highly scalable elements. *First*, the core benefit of adaptive education technologies like Dytective is that all that is needed for their scalability is ensuring that students have access to electronic devices to play the game. The fact that the use of tablets at schools is becoming widespread, including in remote locations (e.g., Ally et al. 2017), facilitates the transition to adopt Dytective. Although further research is needed to assess whether having professional educational psychologists deliver the program enhances its effectiveness, our implementation relied on them primarily because they were already part of our partner's staff, not because the intervention required specialized expertise. Nothing in the design of the sessions appears to depend on skills unique to trained psychologists, and this is in fact a key strength of this technology. For instance, the large-scale rollout of Dytective in Spain is conducted by regular classroom teachers. *Second*, the text messages sent to parents have virtually no costs and could easily be streamlined—

for instance, through WhatsApp Business Automation platforms. *Finally*, the mobile library might be relatively harder to scale up as it is requires someone to physically bring the material to the schools on a weekly basis. Having said this, the reason why the library is rotating in our particular intervention is to reduce costs. With more generous budgets, the material could permanently remain in the school without the need of external visits to deliver it. Moreover, as long as students have access to electronic devices at home or school, the library could instead use the electronic version of the books to eliminate the need for a person to visit the schools (e.g., Anger et al. 2024).

Regarding implementation scalability, our evaluation was conducted in only five schools. This scale likely facilitated communication between FPM staff and participating schools, monitoring of program implementation, and the resolution of operational issues. These factors may have contributed to the high survey completion rates, low attrition, and overall program fidelity observed in the study. At the same time, several features of the intervention suggest that its implementation may scale more readily than that of many educational programs. The intervention relies on highly standardized components, including the same adaptive app, parental messages, and reading materials across schools, which helps preserve program fidelity during scale-up. Moreover, the program imposes relatively low demands on schools and teachers, not requiring extensive staff training. Finally, because the intervention depends only minimally on scarce specialized personnel, scaling up need not entail recruiting and monitoring large numbers of highly trained practitioners. Although these features are encouraging, it remains an open question whether similar levels of engagement and implementation fidelity can be maintained under large-scale deployment.

6 Conclusion

Leveraging adaptive technologies holds significant promise for fostering academic progress, particularly in heterogeneous groups. We study a bundled reading intervention that combines adaptive exercises designed to address linguistic difficulties with complementary components aimed at strengthening interest in reading and fostering parental engagement. We document meaningful improvements in students' self-confidence and academic aspirations, along with suggestive gains in locus of control. These improvements in self-confidence and aspirations are robust to a wide range of checks on the internal validity of our research design, most notably selection on unobservables, non-random attrition, and inference under few clusters. We also find suggestive evidence that these improvements may have translated into performance gains on standardized reading tests, although these results are less robust, most likely due to the smaller available sample for this outcome. Other behavioral responses, such as educational investments by students and parents, are not affected, although we cannot determine whether our particular intervention lacks the elements to make meaningful changes along these dimensions or whether such changes are possible but will take longer to manifest. These results show that a short, light-touch, and cost-effective bundled program can generate meaningful improvements in socioemotional outcomes, which may in turn be accompanied by gains in reading performance.

A Appendix: Additional figures

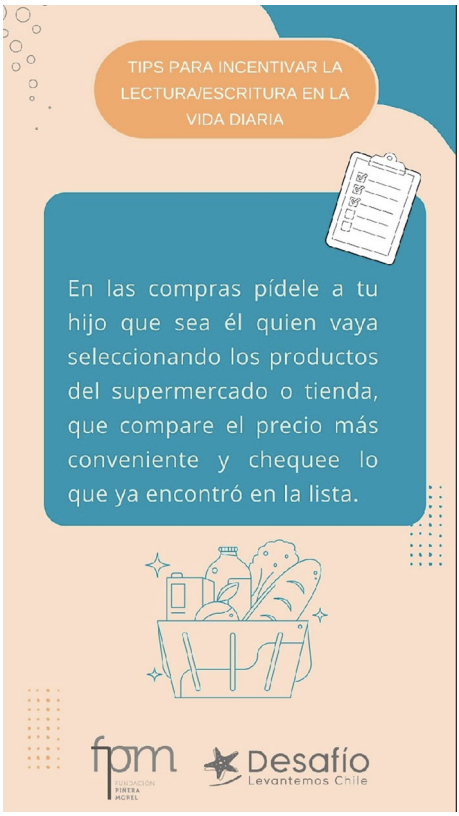
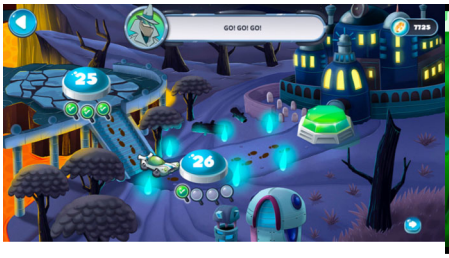


Fig. 1 Sample text message sent to parents. *Notes* Sample text message shared with parents in WhatsApp groups. The English translation is: “When shopping, ask your child to be the one that picks the products from the supermarket or store, ask her to compare prices, and ask her to check which items of the shopping list have already been added to the cart."

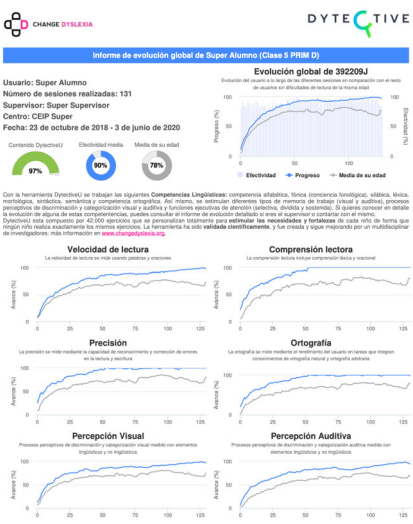
(a) Game interface



(b) Exercise interface



(c) Performance card



(d) In-class session



Fig. 2 Contextual details. *Notes* Panel **a** shows the video-game style of Dytective's interface. Panel **b** displays a sample exercise that a student could be exposed to in Dytective. Panel **c** provides a sample of the report card that educational psychologists can use to monitor the performance of a student. Dytective internally uses this information to personalize the challenges that it gives to the participants. Panel **d** shows how students work individually during a regular session of A Leer Jugando.

B Appendix: Additional tables

Table 5 Summary statistics

Variable	Mean	Std. Dev.	Min.	Max.	N
Male	0.446	0.498	0	1	527
Repeater	0.085	0.280	0	1	527
Screening score	0.197	0.072	0.067	0.494	527
Age (years)	8.524	0.571	8	10	527

Table 5 continued

Variable	Mean	Std. Dev.	Min.	Max.	N
Index: aspirations	0.076	0.959	− 1.607	0.621	527
Aspires to university (dummy)	0.755	0.430	0	1	527
Index: perceived performance relative to peers	0.031	1,000	− 2.931	1.424	527
Math relative to peers	3.624	1.213	1	5	527
Spanish relative to peers	3.719	1.185	1	5	527
Reading relative to peers	3.844	1.242	1	5	527
Index: finds courses easy	0.129	0.896	− 2.843	1.256	527
Math is easy	3.590	1.308	1	5	527
Spanish is easy	4.011	1.170	1	5	527
Reading is easy	4.180	1.215	1	5	527
Index: likes school courses	0.064	0.953	− 3.487	1.190	527
Likes school	4.142	1.151	1	5	527
Likes math	4.042	1.279	1	5	527
Likes Spanish	3.917	1.175	1	5	527
Likes reading	4.063	1.176	1	5	527
Index: grit	0.105	1.025	− 2.816	2.045	527
Likes hard tasks	3.269	1.427	1	5	527
Easily gives up (reversed)	3.218	1.510	1	5	527
Gives up if losing (reversed)	3.786	1.475	1	5	527
Wasted effort if not known (reversed)	3.368	1.595	1	5	527
Index: locus of control	0.049	0.983	− 2.662	1.613	527
Can improve if try	4.055	1.228	1	5	527
Luck matters in exams (reversed)	3.393	1.422	1	5	527
Can reach goals	3.934	1.295	1	5	527
Makes plans	4.006	1.339	1	5	527
Thinks of future	3.879	1.321	1	5	527
Index: individual well-being	0.066	1.014	− 3.322	2.086	527
Feeling happy	4.258	1.078	1	5	527
Many things worry me (reversed)	3.011	1.441	1	5	527
Feeling sad (reversed)	3.882	1.334	1	5	527
Easy to get mad (reversed)	3.332	1.469	1	5	527
Feel like doing nothing (reversed)	3.135	1.400	1	5	527
I do badly (reversed)	2.860	1.396	1	5	527
Hard to focus (reversed)	3.066	1.417	1	5	527

Table 5 continued

Variable	Mean	Std. Dev.	Min.	Max.	N
Index: social well-being	0.112	0.979	− 2.760	1.445	527
Feeling alone (reversed)	3.600	1.487	1	5	527
Classmates respect me	3.586	1.328	1	5	527
Feel safe at school	4.049	1.311	1	5	527
Index: study workdays	0.082	0.995	− 1.124	1.628	527
Hours study weekday	2.753	1.446	1	5	527
Index: study weekends	0.045	0.987	− 1.067	1.747	527
Hours study weekend	2.581	1.402	1	5	527
Index: parental investment	− 0.044	0.962	− 3.403	0.997	527
Parents help with homework	3.763	1.203	1	5	527
Parents care about school	4.463	0.838	1	5	527
Reading test score	57.632	20.693	0	100	368

Summary statistics at baseline of main predetermined variables, indices (and their individual elements), and of the reading score. The sample includes those individuals used in our main estimations (i.e., those that complete both the baseline and the endline survey). Indices were constructed to have a mean of 0 and a standard deviation of 1 for the control group prior to sample selection. All elements of the indices were elicited on 5-point scales (including time devoted to school work). The only exception is an indicator taking the value of 1 if the respondent aspires to reach university. “Reversed" indicates those variables whose scales have been inverted (relative to how they were originally posed to the respondents) to make higher values indicate better outcomes. The statistics are reported after implementing the change.

Table 6 Summary statistics (full sample)

Variable	Mean	Std. Dev.	Min.	Max.	N
Male	0.459	0.499	0	1	715
Repeater	0.094	0.292	0	1	715
Screening score	0.199	0.075	0.067	0.525	635
Age (years)	8.524	0.575	8	10	715
Index: aspirations	0.035	0.981	− 1.607	0.621	715
Aspires to university (dummy)	0.737	0.441	0	1	715
Index: perceived performance relative to peers	0.005	1.017	− 2.931	1.424	715
Math relative to peers	3.608	1.209	1	5	715
Spanish relative to peers	3.705	1.182	1	5	715
Reading relative to peers	3.800	1.267	1	5	715
Index: finds courses easy	0.071	0.954	− 2.843	1.256	715
Math is easy	3.550	1.343	1	5	715
Spanish is easy	3.959	1.219	1	5	715
Reading is easy	4.097	1.275	1	5	715

Table 6 continued

Variable	Mean	Std. Dev.	Min.	Max.	N
Index: like school courses	0.066	0.958	− 3.487	1.190	715
Likes school	4.145	1.149	1	5	715
Likes math	4.049	1.271	1	5	715
Likes Spanish	3.922	1.189	1	5	715
Likes reading	4.046	1.189	1	5	715
Index: grit	0.063	1.029	− 2.816	2.045	715
Likes hard tasks	3.255	1.437	1	5	715
Easily gives up (reversed)	3.190	1.524	1	5	715
Gives up if losing (reversed)	3.731	1.506	1	5	715
Wasted effort if not known (reversed)	3.319	1.615	1	5	715
Index: locus of control	0.036	1.007	− 3.625	1.613	715
Can improve if try	4.038	1.250	1	5	715
Luck matters in exams (reversed)	3.371	1.439	1	5	715
Can reach goals	3.924	1.300	1	5	715
Makes plans	3.997	1.345	1	5	715
Thinks of future	3.902	1.308	1	5	715
Index: individual well-being	0.021	1.019	− 3.322	2.086	715
Feeling happy	4.238	1.098	1	5	715
Many things worry me (reversed)	3.010	1.427	1	5	715
Feeling sad (reversed)	3.836	1.346	1	5	715
Easy to get mad (reversed)	3.302	1.492	1	5	715
Feel like doing nothing (reversed)	3.083	1.446	1	5	715
I do badly (reversed)	2.869	1.401	1	5	715
Hard to focus (reversed)	2.980	1.429	1	5	715
Index: social well-being	0.076	0.977	− 2.760	1.445	715
Feeling alone (reversed)	3.575	1.487	1	5	715
Classmates respect me	3.533	1.338	1	5	715
Feel safe at school	4.027	1.330	1	5	715
Index: study workdays	0.065	0.990	− 1.124	1.628	715
Hours study weekday	2.729	1.438	1	5	715
Index: study weekends	0.031	0.987	− 1.067	1.747	715
Hours study weekend	2.561	1.403	1	5	715
Index: parental investment	− 0.043	1.009	− 3.942	0.997	715
Parents help with homework	3.800	1.208	1	5	715
Parents care about school	4.435	0.906	1	5	715
Reading test score	55.637	20.93	0	100	552

Notes: Replication of Table 5 employing everybody who is available at baseline, irrespective of whether they are eventually observed at endline.

Table 7 Balance check: full sample

Variable	(1) Control		(2) Treatment		(1)-(2) Pairwise t-test	
	N	Mean/(SD)	N	Mean/(SD)	N	Beta/[Wild bootstrapped p-value]
Male	360	0.478	355	0.439	715	−0.010
		(0.500)		(0.497)		[0.530]
Repeater	360	0.094	355	0.093	715	−0.007
		(0.293)		(0.291)		[0.754]
Screening score	322	0.193	313	0.205	635	0.009
		(0.073)		(0.077)		[0.766]
Age (years)	360	8.578	355	8.470	715	−0.099
		(0.587)		(0.559)		[0.108]
Index: aspirations	360	−0.004	355	0.075	715	0.148
		(1.002)		(0.959)		[0.290]
Index: perceived performance relative to peers	360	−0.000	355	0.011	715	−0.057
		(1.003)		(1.032)		[0.460]
Index: finds courses easy	360	0.003	355	0.140	715	0.082
		(0.987)		(0.915)		[0.444]
Index: like school courses	360	−0.004	355	0.137	715	0.091
		(1.000)		(0.909)		[0.868]
Index: grit	360	−0.017	355	0.145	715	0.198
		(0.991)		(1.061)		[0.544]
Index: locus of control	360	−0.005	355	0.078	715	0.082
		(1.002)		(1.011)		[0.506]
Index: individual well-being	360	−0.004	355	0.046	715	0.058
		(1.001)		(1.038)		[0.642]
Index: social well-being	360	0.001	355	0.153	715	0.257
		(0.997)		(0.951)		[0.282]
Index: study workdays	360	0.003	355	0.128	715	−0.018
		(1.002)		(0.975)		[0.892]
Index: study weekends	360	0.006	355	0.057	715	−0.052
		(1.003)		(0.972)		[0.780]

Table 7 continued

Variable	(1) Control		(2) Treatment		(1)-(2) Pairwise t-test	
	N	Mean/(SD)	N	Mean/(SD)	N	Beta/[Wild bootstrapped p-value]
Index: parental investment	360	−0.001	355	−0.087	715	−0.104
		(1.005)		(1.014)		[0.396]
Reading test score	238	58.700	314	53.316	552	−2.369
		(19.492)		(21.701)		[0.480]

Notes: Replication of Table 1 employing everybody who is available at baseline, irrespective of whether they are eventually observed at endline. *** p<0.01, ** p<0.05, * p<0.1

Table 8 Balance check: sample matched with DIA

Variable	(1) Control		(2) Treatment		(1)-(2) Pairwise t-test	
	N	Mean/(SD)	N	Mean/(SD)	N	Beta/[Wild bootstrapped p-value]
Male	150	0.433	218	0.417	368	−0.000
		(0.497)		(0.494)		[0.964]
Repeater	150	0.107	218	0.073	368	−0.020
		(0.310)		(0.261)		[0.836]
Screening score	150	0.192	218	0.202	368	0.003
		(0.073)		(0.071)		[0.702]
Age (years)	150	8.653	218	8.445	368	−0.161*
		(0.567)		(0.525)		[0.088]
Index: aspirations	150	0.146	218	0.110	368	−0.011
		(0.916)		(0.939)		[0.908]
Index: perceived performance relative to peers	150	−0.066	218	0.072	368	−0.135
		(0.946)		(0.988)		[0.480]
Index: finds courses easy	150	0.110	218	0.229	368	0.078
		(0.866)		(0.875)		[0.730]
Index: like school courses	150	0.032	218	0.196	368	−0.009
		(0.938)		(0.888)		[0.990]

Table 8 continued

Variable	(1) Control		(2) Treatment		(1)-(2) Pairwise t-test	
	N	Mean/(SD)	N	Mean/(SD)	N	Beta/[Wild bootstrapped p-value]
Index: grit	150	0.101	218	0.201	368	0.185
		(1.031)		(1.049)		[0.602]
Index: locus of control	150	0.129	218	0.132	368	− 0.014
		(0.929)		(0.966)		[0.838]
Index: individual well-being	150	0.050	218	0.087	368	0.030
		(0.975)		(1.013)		[0.674]
Index: social well-being	150	0.104	218	0.211	368	0.182
		(0.904)		(0.960)		[0.522]
Index: study workdays	150	− 0.000	218	0.186	368	− 0.046
		(1.007)		(0.974)		[0.882]
Index: study weekends	150	0.017	218	0.008	368	− 0.229
		(0.994)		(0.948)		[0.326]
Index: parental investment	150	0.040	218	− 0.064	368	− 0.118
		(0.990)		(0.934)		[0.640]
Reading test score	150	59.356	218	56.445	368	− 0.682
		(19.245)		(21.596)		[0.880]

Notes: Replication of Table 1 employing the subsample of individuals who are matched with the DIA information at endline. *** p<0.01, ** p<0.05, * p<0.1

Table 9 Heterogeneity by at-risk status

	(1)	(2)	(3)	(4)	(5)	(6)
	Aspirations	Performance rel. peers	Courses are easy	Like school	Grit	LOC
Treated	0.096*	0.139**	0.302**	0.201	0.232	0.162**
	[0.068]	[0.010]	[0.032]	[0.354]	[0.614]	[0.028]
At-risk	−0.012	−0.169	−0.103	0.102	0.074	−0.262
	[0.958]	[0.730]	[0.792]	[0.652]	[0.830]	[0.318]
Treated × At-risk	0.073	0.217	0.005	−0.250	0.105	0.230
	[0.778]	[0.396]	[0.958]	[0.312]	[0.642]	[0.254]
Observations	527	527	527	527	527	527
R-squared	0.146	0.184	0.274	0.362	0.184	0.271
	(7)	(8)	(9)	(10)	(11)	(12)
	Indiv. well-being	Social well-being	Study workdays	Study weekends	Parental investment	Reading performance
Treated	0.087	0.210***	0.041	0.142	0.137	3.077**
	[0.234]	[0.008]	[0.486]	[0.230]	[0.352]	[0.018]
At-risk	0.393	0.187	0.229	0.574	−0.008	7.445
	[0.122]	[0.336]	[0.390]	[0.026]	[0.996]	[0.148]
Treated × At-risk	−0.292*	−0.092	−0.075	−0.204	−0.222	−0.970
	[0.084]	[0.528]	[0.452]	[0.220]	[0.516]	[0.702]
Observations	527	527	527	527	527	368
R-squared	0.334	0.278	0.136	0.161	0.176	0.533

Notes: Complementary analysis to that in Table 2's Columns (2) and (3) employing an interaction between an at-risk indicator and the treatment indicator instead of splitting the sample by at-risk status. In brackets, we report wild cluster bootstrapped (clustering at the school-level) p-values using Webb (2023)'s 6-point bootstrap weight distribution. *** p<0.01, ** p<0.05, * p<0.1

Table 10 Heterogeneity by gender

	(1)	(2)	(3)	(4)	(5)	(6)
	Aspirations	Performance rel. peers	Courses are easy	Like school	Grit	LOC
Treated	0.172*	0.079	0.209*	0.201	0.214	0.114
	[0.084]	[0.680]	[0.058]	[0.256]	[0.668]	[0.236]
Male	−0.115	−0.109	−0.204	0.053	−0.120	0.000
	[0.124]	[0.266]	[0.282]	[0.298]	[0.336]	[0.980]
Treated × Male	−0.120	0.181	0.180	−0.076	0.062	0.166
	[0.246]	[0.536]	[0.192]	[0.494]	[0.746]	[0.214]
Observations	527	527	527	527	527	527
R-squared	0.147	0.184	0.276	0.361	0.183	0.271
	(7)	(8)	(9)	(10)	(11)	(12)
	Indiv. well-being	Social well-being	Study workdays	Study weekends	Parental investment	Reading performance
Treated	0.071	0.307*	−0.111	0.043	−0.006	3.003
	[0.332]	[0.096]	[0.424]	[0.778]	[0.958]	[0.178]
Male	0.121*	0.251	−0.272**	−0.173*	0.072	1.733
	[0.072]	[0.352]	[0.016]	[0.058]	[0.566]	[0.626]
Treated × Male	−0.069	−0.214	0.255	0.108	0.203	−0.803
	[0.394]	[0.460]	[0.110]	[0.534]	[0.626]	[0.820]
Observations	527	527	527	527	527	368
R-squared	0.331	0.280	0.138	0.152	0.176	0.529

Notes: Complementary analysis to that in Table 2's Columns (4) and (5) employing an interaction between a male indicator and the treatment indicator instead of splitting the sample by at-risk status. In brackets, we report wild cluster bootstrapped (clustering at the school-level) p-values using Webb (2023)'s 6-point bootstrap weight distribution. *** p<0.01, ** p<0.05, * p<0.1

Table 11 Robustness: Lee (2009) bounds

	Lower bound (1)	Upper bound (2)
Panel (a): Reading performance		
Average reading performance	− 1.637	8.652
# Observations	610	610
Panel (b): Self-perceptions		
Aspirations	− 0.054	0.675
Performance rel. peers	− 0.040	0.369
Courses are easy	− 0.002	0.364
Like school	0.169	0.397
# Observations	715	715
Panel (c): Personality traits		
Grit	− 0.147	0.186
LOC	− 0.015	0.393
# Observations	715	715
Panel (d): Subjective well-being		
Individual well-being	− 0.245	0.116
Social well-being	0.051	0.256
# Observations	715	715
Panel (e): Time and investment		
Study workdays	− 0.322	0.381
Study weekends	− 0.220	0.483
Parental investment	− 0.052	0.283
# Observations	715	715

Notes: Robustness checks of the results in 2's Column (1) in which we provide the lower and upper bounds of the effects based on Lee (2009) bounds. Bounds are tightened employing five indicators for the five pairs of schools in our study.

Table 12 Socioemotional treatment effects, restricted to the sample with available reading scores ($N = 368$)

Outcome (Anderson index)	Treated	Wild-bootstrap p-value	R^2
Panel A. Self-perceptions and skills			
Aspirations	0.124	0.288	0.153
Perceived performance relative to peers	0.218	0.196	0.190
Finds courses easy	0.220*	0.066	0.259
Likes school courses	0.358	0.600	0.372
Grit	0.376	0.494	0.211
Locus of control	0.246*	0.066	0.302

Table 12 continued

Outcome (Anderson index)	Treated	Wild-bootstrap p-value	R^2
Panel B. Well-being, effort, and parental investment			
Individual well-being	0.312*	0.072	0.347
Social well-being	0.278	0.146	0.327
Study time (weekdays)	−0.032	0.646	0.146
Study time (weekends)	0.005	0.912	0.179
Parental investment	0.035	0.616	0.185

Each row reports the estimated treatment effect (*Treated*) from our preferred specification, replicated on the subsample of 368 students for whom both endline reading scores and endline survey responses are available—i.e., the exact sample used for the cognitive result. All outcomes are Anderson (2008) indices standardized to have mean zero and unit standard deviation in the control group. All regressions include match pair fixed effects and the full set of individual controls; the number of observations is 368 in every row. p-values are wild cluster bootstrapped and clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1

Table 13 Variation in leave-out-variable-k change in propensity scores, average reading performance

	Quantiles			
Variable Name	p50 (1)	p75 (2)	p90 (3)	max (4)
LOC	0.001	0.002	0.003	0.006
Screening test	0.002	0.003	0.005	0.012
Age = 9	0.006	0.012	0.020	0.253
Individual well-being	0.010	0.018	0.029	0.066
Age = 8	0.010	0.019	0.030	0.323
Parental investment	0.015	0.024	0.038	0.078
Male	0.019	0.029	0.039	0.068
Like school	0.013	0.026	0.047	0.106
Courses are easy	0.017	0.031	0.048	0.175
Performance rel. peers	0.016	0.031	0.052	0.108
Aspirations	0.021	0.034	0.053	0.093
Ever repeater	0.019	0.031	0.056	0.165
Study workday	0.023	0.041	0.063	0.149
Study weekend	0.024	0.044	0.072	0.184
Grit	0.023	0.046	0.075	0.144
Social well-being	0.030	0.054	0.084	0.273
Average reading performance	0.037	0.068	0.100	0.187

Notes: Features of the distribution of leave-out-variable-k changes in propensity scores used to benchmark c_{BP} as in Masten et al. (2024). Propensity scores estimated as logistic model. Columns (1)–(3) are the 50th, 75th, and 90th percentiles of the distribution. Column (4) reports the maximum of the support of the distribution. Variables ordered in increasing value of the 90th percentile

Table 14 Variation in leave-out-variable-k change in propensity scores, socioemotional outcomes

Variable Name	Quantiles p50 (1)	p75 (2)	p90 (3)	max (4)
Courses are easy	0.001	0.002	0.003	0.010
LOC	0.001	0.002	0.004	0.008
Male	0.005	0.008	0.010	0.015
Screening score	0.006	0.012	0.020	0.046
Repeater	0.008	0.014	0.021	0.067
Age = 9	0.006	0.012	0.021	0.234
Parental investment	0.011	0.018	0.028	0.055
Age = 8	0.011	0.018	0.031	0.325
Study weekends	0.012	0.021	0.033	0.081
Study workdays	0.014	0.024	0.037	0.076
Like school courses	0.014	0.028	0.044	0.152
Individual well-being	0.017	0.033	0.052	0.107
Aspirations	0.017	0.033	0.054	0.092
Perform rel. peers	0.025	0.046	0.068	0.148
Grit	0.034	0.069	0.095	0.200
Social well-being	0.039	0.068	0.103	0.257

Notes: Features of the distribution of leave-out-variable-k changes in propensity scores used to benchmark c_{BP} as in Masten et al. (2024). Propensity scores estimated as logistic model. Columns (1)–(3) are the 50th, 75th, and 90th percentiles of the distribution. Column (4) reports the maximum of the support of the distribution. Variables ordered in increasing value of the 90th percentile.

Table 15 Robustness to selection on unobservables: increases in R^2

Outcome variable	(1) Oster (2019)'s δ	(2) R^2 without controls	(3) R^2 with controls
Panel (a): Reading performance			
Average	0.508	0.085	0.199
Panel (b): Self-perceptions			
Aspirations	18.710	0.020	0.146
Performance rel. peers	3.319	0.019	0.182
Courses are easy	8.855	0.026	0.274
Like school	2.705	0.035	0.360
Panel (c): Personality traits			
Grit	− 3.057	0.026	0.183
LOC	1.819	0.049	0.269

Table 15 continued

Outcome variable	(1) Oster (2019)'s δ	(2) R^2 without controls	(3) R^2 with controls
Panel (d): Subjective well-being			
Individual well-being	−0.817	0.025	0.331
Social well-being	−8.115	0.053	0.277
Panel (e): Time and investment			
Study workdays	−0.508	0.003	0.135
Study weekends	22.789	0.006	0.152
Parental investment	6.984	0.028	0.174

The table reports the R^2 in the restricted (short) and unrestricted (long) regressions of the Oster (2019) analysis from Table 4. The restricted regression only controls for strata fixed effects. The unrestricted regression additionally includes for the maintained controls in Table 2.

Table 16 Robustness to the use of composite indices for socioemotional outcomes

	Self-perceptions (1)	Personality traits (2)	Subjective well-being (3)	Time and investment (4)
Treated	0.294*	0.292	0.145*	0.120
	[0.072]	[0.388]	[0.062]	[0.316]

Notes: Replication of Table 2's Column (1) where the outcomes are the Anderson (2008)-based indices constructed aggregating the outcomes in Panels (b)–(e). In brackets, we report wild cluster bootstrapped (clustering at the school-level) p-values using Webb (2023)'s 6-point bootstrap weight distribution. *** p<0.01, ** p<0.05, * p<0.1

Table 17 Robustness: Propensity score matching

	(1) Aspirations	(2) Performance rel. peers	(3) Courses are easy	(4) Like school	(5) Grit	(6) LOC
Treated	0.051	0.193	0.302	0.168	0.235	0.262
Observations	527	527	527	527	527	527
	(7)	(8)	(9)	(10)	(11)	(12)
	Indiv. well-being	Social well-being	Study workdays	Study weekends	Parental investment	Reading performance
Treated	0.035	0.139	0.024	0.044	0.118	2.620
Observations	527	527	527	368	527	368

The table provides the estimates under propensity score matching based on Masten and Poirier (2018).

C Appendix: Additional details

C.1 Details on Dytective's personalized challenges

Dytective receives as inputs: a) the age of the user, b) the number of sessions already completed, c) the performance of the user in each of the completed sessions. The age of each user is required so that the exercises/games presented to each user reflect the cognitive capabilities of users in this age. The number of sessions is used to understand if the user faces difficulties or not in a specific linguistic capability. The performance of the user in the past sessions is needed so that Dytective personalizes the future exercises/games. Specifically, users will face games with increasing difficulty if their performance is improving and they will face games that are aimed to improve cognitive abilities that have not improved in the past sessions.

C.2 School selection

To clarify the source of identifying variation, it is useful to describe how FPM selects schools for the program and how the rollout operates. FPM's objective is to eventually implement the intervention in all schools in the country. To this end, it issued a call for expressions of interest several years before the cohort studied here. The five treatment schools in our evaluation were located in Santiago and had applied to participate in earlier rounds but could not be accommodated at the time because of budgetary and logistical constraints. Instead, FPM had committed to including them in a subsequent rollout wave, namely the one evaluated in this paper. By the time we joined the project as external evaluators, these schools had already been scheduled for treatment. We therefore selected comparison schools from the pool of non-treated schools in Santiago, matching them as closely as possible along the three key dimensions that FPM uses to determine the feasibility of participation in a given round. Several features lend credibility to our identification strategy. First, FPM's stated objective of universal expansion makes it less likely that schools are selected based on their expected gains from the program, as the relevant consideration is primarily when, rather than whether, each school can enter the program. Second, schools are incorporated in small cohorts mainly on the basis of budgetary feasibility and logistical clustering, rather than expected school-specific trends or anticipated program gains. Lastly, the strong balance in observable characteristics between treatment and control schools further strengthens the credibility of our identifying assumption.

C.3 Timeline of the intervention

The evaluation team established a research collaboration with FPM in March 2023. The identification of and contact with control schools was done by July 2023. A Leer Jugando was in place between September and the end of November 2023. The implementing partner started fielding the baseline survey at the end of August. The main data collection took place during September but, due to logistic limitations, treated schools were prioritized. Some classes in control schools completed the survey

in the first half of October. Without the intervention, we expect the outcomes of interest among the control schools not to have drifted from what we would have observed had they been surveyed in September. Still, we account for this variation in our main specifications by controlling for the month of survey fielding. The endline survey was distributed at the end of November and the beginning of December 2023.

C.4 Psychometric validity of survey items

To address concerns regarding the measurement reliability of the indices we construct, we evaluated the internal consistency of each family using Cronbach's alpha and average inter-item correlations. While some scales (particularly, Grit and Parental investment) have Cronbach's alphas below the typical 0.7 rule of thumb, such values are not uncommon in developmental and education research. Moderate internal consistency coefficients are common in child socioemotional assessment instruments and do not necessarily imply poor construct validity (Schmitt 1996). Moreover, Cronbach's alpha depends not only on the degree of association among items but also on the number of items in a scale, implying that shorter scales will often exhibit lower alpha values even when they capture a coherent underlying construct. For this reason, we complement Chronbach's alpha with average inter-item correlation (AIC). Reassuringly, the AICs for our multi-item indices generally fall within or close to the range considered appropriate for broad psychological constructs (approximately 0.20–0.40, as in Briggs and Cheek 1986). Together, these psychometric statistics suggest that the indices exhibit acceptable internal consistency and alleviate concerns that the findings are driven by substantial measurement error. To the extent that remaining measurement error is classical, it would tend to attenuate estimated effects and reduce statistical power, implying that our estimates are more likely to be conservative than overstated (Table 18).

Table 18 Psychometric properties of socioemotional index families

Index family	No. of items	Cronbach's α	AIC
Self-perceptions			
Academic aspirations	1	–	–
Performance rel. peers	3	0.587	0.321
Courses are easy	3	0.581	0.316
Like school	4	0.629	0.298
Personality traits			
Grit	4	0.487	0.192
LOC	5	0.588	0.222

Table 18 continued

Index family	No. of items	Cronbach's α	AIC
Subjective well-being			
Individual well-being	7	0.724	0.272
Social well-being	3	0.425	0.198
Time and investment			
Study workdays	1	–	–
Study weekends	1	–	–
Parental investment	2	0.471	0.308

The table reports the number of items, Cronbach's α, and average inter-item correlation (AIC) for each socioemotional index family. Internal consistency metrics are not applicable for single-item indices

D Appendix: Summaries of Oster (2019) and Masten et al. (2024)

D.1 Summary of Oster (2019)

Consider the following data generating process:

$$Y = T\beta + X\gamma_X + Z\gamma_Z + \varepsilon, \tag{2}$$

where β is the treatment effect of interest. We observe (Y, T, X). A sufficient assumption for identification of the treatment effect is that T is orthogonal to (Z, ε) conditional on X (i.e., the conditional ignorability assumption we discuss in the text). We want to assess the robustness of our results to misspecification due to possible selection into treatment induced by Z conditional on X.

The two feasible regressions given observations (Y, T, X) are:

1. Restricted (short) regression: Y on $T \rightarrow \beta_r$ coefficient on T with R-squared R_r^2; and
2. Unrestricted (long) regression (which coincide with our reported estimates): Y on T and $X \rightarrow \beta_u$ coefficient on T with R-squared R_u^2.

Common practice would compare changes between β_r and β_u to argue how robust the results might be to the hypothetical inclusion of the unobserved Z.

Oster (2019) argues that looking at coefficient stability (i.e., looking at $\beta_u - \beta_r$) is not sufficient to conclude about the robustness of the results to selection on unobservables if the relationship between T and Z is unknown. We need to consider two additional quantities:

1. The difference in R^2 between the unrestricted and restricted regressions: $R_u^2 - R_r^2$. Coefficient stability is more informative of robustness if the observed covariates X explain a large portion of variation in the outcome.
2. The theoretical maximum explainable variance of the outcome: $R_{\max}$. This is necessary as a comparison for what can be considered a "large" change in the R^2.

Suppose that "selection on unobservables" is proportional to "selection on observables" in the following manner:

$$\delta \frac{\text{Cov}(X\gamma_X, T)}{\text{Var}(X\gamma_X)} = \frac{\text{Cov}(Z\gamma_Z, T)}{\text{Var}(Z\gamma_Z)}, \tag{3}$$

where δ is the relevant constant of proportionality. Then, the quantities $(\beta_u - \beta_r, R_r^2, R_u^2)$ are related to the unknown quantities:

- $\beta - \beta_u$, the bias in the treatment effect;
- δ, the proportionality between selection on observables and selection on unobservables; and
- R_{max}^2, the theoretical maximum proportion of the variation in the outcome that can be explained.

Oster (2019) characterizes these relationships. Thus, given two of the unknown quantities, one can solve for the third unknown quantity. This gives rise to a natural way to quantify robustness to selection on unobservables within this framework: to find the proportionality constant δ such that the bias due to unobservables drives the treatment effect to zero (i.e., $\beta = 0$) for some given R_{max}^2. We set $R_{\text{max}}^2 = 1.5 \times R_u^2$ which is more conservative than what is often used in the literature. Our estimates $\hat{\delta}$ are reported in Column (1) of Table 4.

D.2 Summary of Masten et al. (2024)

Masten et al. (2024) extends the analysis in Oster (2019) to a propensity score matching setting. We again start with the unconfoundedness assumption:

$$Y(0), Y(1) \perp T \mid X, \tag{4}$$

where $Y(0)$ and $Y(1)$ are the potential outcomes corresponding to the treatment indicator T. The propensity score to treatment is $\Pr(T = 1 \mid X = x)$.

Masten et al. (2024) consider a relaxation of the unconfoundedness assumption in the form of "conditional c-dependence." More precisely, T is conditionally c-dependent with $Y(t)$ conditional on X if for all $X = x$,

$$\sup_{y \in \text{supp}(Y(t)|X=x)} |\Pr(T = 1 \mid Y(t) = y, X = x) - \Pr(T = 1 \mid X = x)| \leq c. \tag{5}$$

When $c = 0$, then conditional c-dependence coincides with the unconfoundedness assumption. For $c > 0$, we allow for deviations from unconfoundedness by allowing the conditional probability $\Pr(T = 1 \mid Y(t) = y, X = x)$ to differ from the observed propensity score $\Pr(T = 1 \mid X = x)$ by at most c. Given a level of $c > 0$, the treatment effect is only partially identified and we can construct bounds.

A natural parameter to assess the sensitivity of our estimates to unconfoundedness, then, is to find a breakdown point c_{BP} which is the largest c such that the identified bounds do not contain 0 (i.e., a null treatment effect when the TE > 0). We report

these breakdown points in Column (2) of Table 4. Intuitively, these breakdown points tell us how much we can relax the unconfoundedness assumption before we become unsure of the sign of the treatment effects based on the identified bounds.

What is a level of c_{BP} for which we can say a conclusion is robust? This is context specific. Following Oster (2019), one can use selection on observables to provide a benchmark for the breakdown point. Masten et al. (2024) propose comparing the breakdown point relative to the distribution of leave-one-variable-out changes in the propensity score. The idea is to see the changes in the propensity score leaving out one of the observed variables as a benchmark for the values of c that might be reasonable. More precisely, for each component of X indexed by k, we consider the random variable

$$\Delta_k = \mid \Pr(T = 1 \mid X) - \Pr(T = 1 \mid X_{-k}) \mid, \tag{6}$$

where X_{-k} is the set of observables excluding the kth component. These are changes in the magnitude of the propensity score when one variable is excluded. In practice, these changes are estimated at the individual level which means that there is a distribution of these possible changes. A very conservative approach is to compare c_{BP} to the maximum of the support of Δ_k. But we can also compare it to other quantiles of the distribution, such as the 90th percentile, which is our preferred reference value. The idea is that if c_{BP} is greater than the reference values, the conclusion might be robust because the level of selection on unobservables needs to be at least comparable to the selection induced by the observables to render the analysis inconclusive. We report features of the distribution of leave-out-variable-k changes in the propensity scores in Appendix Tables 13 and 14.

Data Availability Data availability statement for "Game Changer? Impact of a Multi-component Reading Intervention on Cognitive and Socioemotional Skills" By Garcia-Brazales, De Vera, and Rello. For anonymity reasons, the data collected for this project cannot be made fully public. An anonymized version with additional restrictions may be made available upon request to Javier Garcia-Brazales (j.garcia-brazales2@exeter.ac.uk).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Alan S, Boneva T, Ertac S (2019) Ever failed, try again, succeed better: results from a randomized educational intervention on grit*. Q J Econ 134(3):1121–1162. https://doi.org/10.1093/qje/qjz006

Alan S, Mumcu I (2024) Nurturing childhood curiosity to enhance learning: evidence from a randomized pedagogical intervention. Am Econ Rev 114(4):1173–1210. https://doi.org/10.1257/aer.20230084

Ally M, Balaji V, Abdelbaki A, Cheng R (2017) Use of tablet computers to improve access to education in a remote location. J Learn Dev. https://doi.org/10.56059/jl4d.v4i2.219

Almlund M, Duckworth AL, Heckman J, Kautz T (2011) Chapter 1—personality psychology and economics. In Hanushek EA, Machin S, Woessmann L eds. Handbook of the economics of education, 4 of handbook of the economics of education, 1–181. Elsevier. https://doi.org/10.1016/B978-0-444-53444-6.00001-8

Anderson ML (2008) Multiple inference and gender differences in the effects of early intervention: a reevaluation of the abecedarian, perry preschool, and early training projects. J Am Stat Assoc 103(484):1481–1495. https://doi.org/10.1198/016214508000000841

Andrabi T, Das J, Khwaja AI, Zajonc T (2011) Do value-added estimates add value? Accounting for learning dynamics. Am Econ J Appl Econ 3(3):29–54. https://doi.org/10.1257/app.3.3.29

Anger S, Christoph B, Galkiewicz A, Margaryan S, Peter F, Sandner M, Siedler T (2024) A library in the palm of your hand? A randomized reading intervention with low-income children. Working papers 2024-018, Human Capital and Economic Opportunity Working Group. https://ideas.repec.org/p/hka/wpaper/2024-018.html

Arbour MC, Soto C, Alée Y, Atwood S, Muñoz P, Marzolo M (2023) Absenteeism prevention in preschools in Chile: impact from a quasi-experimental evaluation of 2011–2017. Ministry of Education data. In Frontiers in Education, 7, 975092. Frontiers Media SA

Ashraf N, Bau N, Low C, McGinn K (2020) Negotiating a better future: how interpersonal skills facilitate intergenerational investment*. Q J Econ 135(2):1095–1151. https://doi.org/10.1093/qje/qjz039

Banerjee A, Banerji R, Berry J, Duflo E, Kannan H, Mukherji S, Shotland M, Walton M (2016) Mainstreaming an effective intervention: evidence from randomized evaluations of Teaching at the Right Level in India. Technical report, National Bureau of Economic Research

Banerjee AV, Cole S, Duflo E, Linden L (2007) Remedying education: evidence from two randomized experiments in India*. Q J Econ 122(3):1235–1264. https://doi.org/10.1162/qjec.122.3.1235

Beg S, Halim W, Lucas AM, Saif U (2022) Engaging teachers with technology increased achievement, bypassing teachers did not. Am Econ J Econ Pol 14(2):61–90. https://doi.org/10.1257/pol.20200713

Bergman P (2021) Parent–child information frictions and human capital investment: evidence from a field experiment. J Polit Econ 129(1):286–322. https://doi.org/10.1086/711410

Bergman P, Edmond-Verley C, Notario-Risk N (2018) Parent skills and information asymmetries: experimental evidence from home visits and text messages in middle and high schools. Econ Educ Rev 66:92–103. https://doi.org/10.1016/j.econedurev.2018.06.008

Blackwell LS, Trzesniewski KH, Dweck CS (2007) Implicit theories of intelligence predict achievement across an adolescent transition: a longitudinal study and an intervention. Child Dev 78(1):246–263. https://doi.org/10.1111/j.1467-8624.2007.00995.x

Bold T, Filmer D, Molina E, Svensson J (2019) The lost human capital: teacher knowledge and student achievement in Africa. World Bank Policy Research Working Paper (8849)

Briggs SR, Cheek JM (1986) The role of factor analysis in the development and evaluation of personality scales. J Pers 54(1):106–148. https://doi.org/10.1111/j.1467-6494.1986.tb00391.x

Bruhn M, McKenzie D (2009) In pursuit of balance: randomization in practice in development field experiments. Am Econ J Appl Econ 1(4):200–232. https://doi.org/10.1257/app.1.4.200

Cameron AC, Gelbach JB, Miller DL (2008) Bootstrap-based improvements for inference with clustered errors. Rev Econ Stat 90(3):414–427. https://doi.org/10.1162/rest.90.3.414

Carlana M, La Ferrara E (2021) Apart but connected: online tutoring and student outcomes during the COVID-19 pandemic

Carneiro P, Crawford C, Goodman A (2007) The impact of early cognitive and non-cognitive skills on later outcomes. Centre for Economics of Education Working Paper

Carneiro P, Galasso E, Garcia IL, Bedregal P, Cordero M (2024) Impacts of a large-scale parenting program: experimental evidence from Chile. J Polit Econ 132(4):1113–1161. https://doi.org/10.1086/727288

Cuevas-Ruiz P, Rello L, Sanz I, Sevilla A (2021) Medidas educativas de refuerzo en lectoescritura: el caso del programa de Ayuda a la Dislexia en la Comunidad de Madrid. Cuadernos Económicos de ICE (102). https://doi.org/10.32796/cice.2021.102.7317

Cunha F, Heckman JJ (2008) Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. J Hum Resour 43(4):738–782. https://doi.org/10.3368/jhr.43.4.738

Cunha F, Heckman JJ, Schennach SM (2010) Estimating the technology of cognitive and noncognitive skill formation. Econometrica 78(3):883–931. https://doi.org/10.3982/ECTA6551

Dickson M, Gregg P, Robinson H (2016) Early, late or never? When does parental education impact child outcomes? Econ J 126(596):F184–F231. https://doi.org/10.1111/ecoj.12356

Dietrichson J, Filges T, Seerup JK, Klokker RH, Viinholt BCA, Martin Bøg, Eiberg M (2021) Targeted school-based interventions for improving reading and mathematics for students with or at risk of academic difficulties in Grades K-6: a systematic review. Campbell Syst Revi 17(2):e1152. https://doi.org/10.1002/cl2.1152

Dillon MR, Kannan H, Dean JT, Spelke ES, Duflo E (2017) Cognitive science in the field: a preschool intervention durably enhances intuitive but not formal mathematics. Science 357(6346):47–55. https://doi.org/10.1126/science.aal4724

Dizon-Ross R (2019) Beliefs about their children's academic ability implications for educational investments. Am Econ Rev 109(8):2728–65. https://doi.org/10.1257/aer.20171172

Duflo E, Glennerster R, Kremer M (2007) Chapter 61 using randomization in development economics research: a toolkit. In: Paul ST, Strauss JA eds. Handbook of Development Economics, 4, 3895–3962. Elsevier. https://doi.org/10.1016/S1573-4471(07)04061-2

Escueta M, Nickow AJ, Oreopoulos P, Quan V (2020) Upgrading education with technology: insights from experimental research. J Econ Lit 58(4):897–996. https://doi.org/10.1257/jel.20191507

Glewwe P, Muralidharan K (2016) Chapter 10—improving education outcomes in developing countries: evidence, knowledge gaps, and policy implications. In: Hanushek EA, Machin S, Woessmann L eds. Handbook of the economics of education, 5, 653–743. Elsevier. https://doi.org/10.1016/B978-0-444-63459-7.00010-5

Gortazar L, Hupkau C, Roldán-Monés A (2024) Online tutoring works: experimental evidence from a program with vulnerable children. J Public Econ 232:105082. https://doi.org/10.1016/j.jpubeco.2024.105082

Graham J, Kelly S (2019) How effective are early grade reading interventions? A review of the evidence. Educ Res Rev 27:155–175. https://doi.org/10.1016/j.edurev.2019.03.006

Guarino CM, Reckase MD, Wooldridge JM (2015) Can value-added measures of teacher performance be trusted? Educ Finance Policy 10(1):117–156. https://doi.org/10.1162/EDFP_a_00153

Hanus MD, Fox J (2015) Assessing the effects of gamification in the classroom: a longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance. Comput Educ 80:152–161. https://doi.org/10.1016/j.compedu.2014.08.019

Heckman JJ, Masterov DV (2007) The productivity argument for investing in young children

Kautz T, Heckman JJ, Diris R, ter Weel B, Borghans L (2014) Fostering and measuring skills: improving cognitive and non-cognitive skills to promote lifetime success. Working Paper 20749, National Bureau of Economic Research. https://doi.org/10.3386/w20749

Kim Y-SG, Lee H, Zuilkowski SS (2020) Impact of literacy interventions on reading skills in low- and middle-income countries: a meta-analysis. Child Dev 91(2):638–660. https://doi.org/10.1111/cdev.13204

Lafortune J, Pugatch T, Tessada J, Ubfal D (2024) Can gamified online training make high school students more entrepreneurial? Experimental evidence from Rwanda. Econ Educ Rev 101:102559. https://doi.org/10.1016/j.econedurev.2024.102559

Lee DS (2009) Training, wages, and sample selection: estimating sharp bounds on treatment Effects. Rev Econ Stud 76(3):1071–1102. https://doi.org/10.1111/j.1467-937X.2009.00536.x

MacKinnon JG, Ørregaard Nielsen M, Webb MD (2023a) Cluster-robust inference: a guide to empirical practice. J Econ 232(2):272–299. https://doi.org/10.1016/j.jeconom.2022.04.001

MacKinnon JG, Ørregaard Nielsen M, Webb MD (2023b) Fast and reliable jackknife and bootstrap methods for cluster-robust inference. J Appl Econ 38(5):671–694. https://doi.org/10.1002/jae.2969

Masten MA, Poirier A (2018) Identification of treatment effects under conditional partial independence. Econometrica 86(1):317–351. https://doi.org/10.3982/ECTA14481

Masten MA, Poirier A, Zhang L (2024) Assessing sensitivity to unconfoundedness: estimation and inference. J Bus Econ Stat 42(1):1–13. https://doi.org/10.1080/07350015.2023.2183212

Muralidharan K, Singh A, Ganimian AJ (2019) Disrupting education? Experimental evidence on technology-aided instruction in India. Am Econ Rev 109(4):1426–60. https://doi.org/10.1257/aer.20171112

O'Neill S, Kreif N, Grieve R, Sutton M, Sekhon JS (2016) Estimating causal effects: considering three alternatives to difference-in-differences estimation. Health Serv Outcomes Res Method 16(1):1–21. https://doi.org/10.1007/s10742-016-0146-8

Organisation for Economic Co-operation and Development (2023) PISA 2022 Results (Volume I): the state of learning and equity in education. OECD Publishing. https://doi.org/10.1787/53f23881-en

Oster E (2019) Unobservable selection and coefficient stability: theory and Evidence. J Bus Econ Stat 37(2):187–204. https://doi.org/10.1080/07350015.2016.1227711
Pritchett L (2013) The rebirth of education: schooling ain't learning. CGD Books
Rello L, Baeza-Yates R, Ali A, Bigham JP, Serra M (2020) Predicting risk of dyslexia with an online gamified test. PLoS ONE 15(12):e0241687. https://doi.org/10.1371/journal.pone.0241687
Rello L, Macias A, Herrera M, de Ros C, Romero E, Bigham JP (2017) DytectiveU: a game to train the difficulties and the strengths of children with dyslexia. In: Proceedings of the 19th international ACM SIGACCESS conference on computers and accessibility ASSETS '17, 319–320, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3132525.3134773
Resnjanskij S, Ruhose J, Wiederhold S, Woessmann L, Wedel K (2024) Can mentoring alleviate family disadvantage in adolescence? A field experiment to improve labor market prospects. J Polit Econ 132(3):1013–1062. https://doi.org/10.1086/726905
Sanchez A, Woldehanna T, Thuc Duc L, Boyden J, Penny M, Galab S (2025) Young lives: an international study of childhood poverty: round 5, 2016. https://doi.org/10.5255/UKDA-SN-8357-2
Scammacca NK, Roberts GJ, Cho E, Williams KJ, Roberts G, Vaughn SR, Carroll M (2016) A century of progress: reading interventions for students in grades 4–12, 1914–2014. Rev Educ Res 86(3):756–800. https://doi.org/10.3102/0034654316652942. (**PMID: 28529386**)
Schmitt N (1996) Uses and abuses of coefficient alpha. Psychol Assess 8(4):350–353. https://doi.org/10.1037/1040-3590.8.4.350
Shaywitz SE (1998) Dyslexia. N Engl J Med 338(5):307–312. https://doi.org/10.1056/NEJM199801293380507
Singh A (2015) Private school effects in urban and rural India: panel estimates at primary and secondary school ages. J Dev Econ 113:16–32. https://doi.org/10.1016/j.jdeveco.2014.10.004
Singh A (2020) Learning more with every year: school year productivity and international learning divergence. J Eur Econ Assoc 18(4):1770–1813. https://doi.org/10.1093/jeea/jvz033
Sisk VF, Burgoyne AP, Sun J, Butler JL, Macnamara BN (2018) To what extent and under which circumstances are growth mind-sets important to academic achievement? Two meta-analyses. Psychol Sci 29(4):549–571. https://doi.org/10.1177/0956797617739704. (**PMID: 29505339**)
Skulmowski A, Kate Man X (2022) Understanding cognitive load in digital and online learning: a new perspective on extraneous cognitive load. Educ Psychol Rev 34(1):171–196
Todd PE, Wolpin KI (2003) On the specification and estimation of the production function for cognitive achievement. Econ J 113(485):F3–F33. https://doi.org/10.1111/1468-0297.00097
Valenzuela JP, Bellei C, Danae de los Ríos (2014) Socioeconomic school segregation in a market-oriented educational system. The case of Chile. J Educ Policy 29(2):217–241
Webb MD (2023) Reworking wild bootstrap-based inference for clustered errors. Can J Econ/Revue canadienne d'économique 56(3):839–858. https://doi.org/10.1111/caje.12661
Yang L-P, Li C-B, Li X-M, Zhai M-M, Zhao J, Weng X-C (2022) Prevalence of developmental dyslexia in primary school children: a protocol for systematic review and meta-analysis. World J Pediatr 18(12):804–809. https://doi.org/10.1007/s12519-022-00572-y

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.