Game Changer: Impact of a Reading Intervention on Cognitive and Non-Cognitive Skills[*]

Micole De Vera[†] Javier Garcia-Brazales[‡] Luz Rello[§]

September 12, 2024

Abstract

We evaluate a holistic reading intervention involving 600 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 to increase student's interest and parental involvement. On one side, we find an improvement in performance on a nation-wide, standardized language test by 13% of a standard deviation. On the other side, we find that 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, and to have a more internal locus-of-control. Our results show that cognitive and non-cognitive skills can be changed through a short, light-touch, and cost-effective education technology intervention.

JEL CODES: I24, I31.

KEYWORDS: Field experiment, Computer-based reading intervention, Non-cognitive skills, Chile, Dyslexia.

[*] 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.

[†] Banco de España and IZA

[‡] Corresponding author: University of Exeter and LEAP (Bocconi). E-mail: j.garcia-brazales2@exeter.ac.uk. Address: Rennes Dr, Exeter EX4 4PU, United Kingdom.

[§] IE University

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 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 and Masterov, 2007; Resnjanskij et al., 2024). Second, institutions may be unprepared to support students with learning disorders that lead them to leak out of the educational pipeline. For instance, dyslexia, which affects 15-20% of the population (American Psychiatric Association et al., 2013), is an important predictor of low motivation and academic self-worth, grade repetition, and drop out, despite not being correlated with intelligence (Singer, 2008; Cortiella and Horowitz, 2014). As such, interventions that have a strong chance to improve education need to be holistic in involving parents, teachers and students. However, developing holistic 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" which is implemented by Fundación Piñera Morel (FPM) and 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 involves parents in the reading development of their children. 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 on common reading difficulties among Spanish-speaking dyslexic children.[2] 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

[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] 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.

spelling, reading speed, and vocabulary of the participants, irrespective of their initial ability.

Dytective is played during 45-minute school visits that are facilitated by an educational psychologist[3] 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 towards improving reading, it also contains features that may inadvertently strengthen non-cognitive 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 non-cognitive 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 elements that have the potential to strengthen a *growth mindset*[4] is a key feature of the program, and the main focus of this paper. 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 La Ferrara (2021).

Our evaluation comprises 600 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 just by chance 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 widely employed by Chilean governmental institutions, size, and location. This research design relies on potential outcomes of students being the same within the match-pair. We show that, within the

[3] The educational psychologists posses tertiary education training in both pedagogy and psychology and are part of FPM's workforce.

[4] 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).

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 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. To explore potential mechanisms, we then study the effects on non-cognitive skills and perceptions elicited through an ad hoc survey that we designed and distributed before and after the intervention.

We detect gains in reading performance of about 13% of a standard deviation for students in treated schools. These effects are mediated by improved self-perceptions about academic aspirations, performance relative to peers, and the difficulty of courses. We also find suggestive improvements in locus of control and well-being. These impacts are present both for students that are at risk of having dyslexia and those that are not, but do not seem to be complemented by higher investments by the child (study time) nor by parents (i.e., caring about the child's academics and time devoted to helping the child with homework).

Acknowledging the possible limitations of our study design, we take steps to assuage concerns about the validity and robustness of our conclusions. First, with a small number of clusters, we may be concerned that standard errors computed based on asymptotic approximations may lead to incorrect statistical inference. Hence, we rely on inference based on wild cluster bootstrap procedures which have been shown to provide reliable inference in settings with as few as five clusters (Cameron et al., 2008). Second, though we argue that our setting is similar to a randomized match-pairs design, it is technically not. As such, we might be concerned that selection could affect our results. As mentioned, we control for a battery of individual controls to make conditional independence more plausible. To further probe the robustness of our results to selection on unobservables, we perform analyses following Oster (2019) and Masten et al. (2024). For the effects on non-cognitive outcomes, we find that selection on unobservables needs to be substantially larger than selection on our observed covariates to overturn our conclusions. The effects on reading performance are more suggestive.

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 non-cognitive 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),[5] we evaluate a program that provides a novel holistic approach by combining a reading-enhancement element with two other components that involve children's non-cognitive skills and parental investments. This program therefore tackles the reading problem along multiple fronts, arguably offering better chances to have an impact. In a similar vein, 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).

Relative to the burgeoning literature on how to use education technology to improve learning in early years, we make two contributions.[6] *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 non-cognitive skills and perceptions. We find gains in aspirations, self-confidence and locus-of-control, which are 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).[7] *Second*, we build upon existing evidence showing that personalised 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 not only can address learning of individuals throughout the whole ability distribution, but also goes one step further for those students that are constrained by the innate condition of dyslexia. 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), as dyslexic students typically do. To the best of our knowledge, this is the first time that experimental evidence on the cognitive and non-cognitive impact of education technologies is obtained jointly for both the dyslexic and non-dyslexic collectives.[8] By finding that both groups benefit from the program, our work offers valuable policy lessons to promote inclusive

[5] 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.

[6] 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) or Lafortune et al. (2024).

[7] 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 La 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 tutor (in our case, the educational psychologist).

[8] Galuschka et al. (2014) and Galuschka et al. (2020) provide meta-analyses of the limited existing experimental evidence on dyslexia-related interventions, including computerized approaches. The scarce work available exclusively evaluates the impact of the intervention on dyslexic individuals and does not extend to a wide range of non-cognitive skills.

growth in human capital.

Relative to the existing literature on the malleability of non-cognitive skills during early life (e.g., Ashraf et al., 2020; Alan and Mumcu, 2024), we provide novel evidence of how a short, light-touch, and low-cost intervention can jointly improve cognitive and non-cognitive scores among dyslexic students, a sizable subpopulation that disproportionately suffers from low self-confidence and aspirations as well as from higher rates of academic failure. Moreover, by showing that the effects are also present among not-at-risk students, we strengthen recent results by Alan et al. (2019) that, unlike previous consensus (Sisk et al., 2018), it is possible to design interventions that benefit individuals throughout the whole distribution. A limitation of the present paper is that we are not able to isolate the impact of Dytective from that of its auxiliary programs — i.e., the mobile library and the text messages to the 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 indicates that separating the relative contributions of each element of the intervention would, if anything, allow us to 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 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

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

application, with a focus on weaker linguistic areas and general cognitive abilities, e.g., working memory and attention. Each challenge takes around 20 minutes 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 minutes 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).

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 Figure A1 provides an example of how the text messages look like. Appendix Figure A2 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 the various reading and cognitive skills dimensions.

[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).

[11] Appendix C.2 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 minutes per week. A Leer Jugando was implemented for half of the 90-minute blocks on three different days.

[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.

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. As such, we were not able to design the evaluation through a fully randomized controlled experiment. Fortunately, discussions with our implementing partner highlighted that the five schools had been chosen primarily for convenience and independently of potential gains from the program. 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.[14] Matched schools serve as controls for the treated schools.

To encourage the control schools to allow us to collect data and distribute 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 means 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 essentially random.

Description of program protocols. We do not alter any of the elements of A Leer Jugando to avoid randomization biases (Heckman, 2020). As per the program's design, *all students* in grade 3 of the treated schools were subject to the Dytective and the mobile

[14] 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.

library components of the intervention during regular school time, while no student in the control schools had access to either.[15] 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 teacher-parents meeting prior to the start of the program.[16] 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 non-cognitive 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.

Non-cognitive 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 non-cognitive 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.[17] Given the young age of the

[15] 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.

[16] 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 minutes 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.

[17] 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 Section 4.2, this is unlikely to be a concern.

respondents, during survey collection, students were aided by the enumerators to make sure that questions were clearly understood and responses were properly recorded. 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 responded.

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 towards 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 minutes," "16–30 minutes," "31 minutes–1 hour," and "over an hour."

10. **Effort on weekends.** Time devoted to studying on a normal weekend. Same options as for weekdays.

11. **Parental investment.** As an additional outcome, and to help us better understand potential mechanisms, we look into measures of parental investment in the child. For this, we exploit information on how much students report that their parents help them with school work and 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."

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 859 students across control and treatment schools were enrolled in third grade. 715 of them (83%) 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 non-cognitive 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 La 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.

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 B.1.[18] 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. This fraction represents the estimated fraction of dyslexic individuals worldwide (e.g., Shaywitz, 1998). 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 Section 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 Subsection 4.2 we also show robustness to this choice by estimating models that employ nonparametric propensity score matching. Thus, our

[18] The counterpart for the full sample available at baseline (i.e., regardless of attrition at endline) is provided in Appendix Table B.2. 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 ≈ 26%.

baseline regressions are of the form:

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

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 and use the 6-point bootstrap weight distribution proposed by Webb (2023).

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 (Guarino et al., 2015; O'Neill et al., 2016).

Given the design of the program, our estimates are best interpreted as intention-to-treat 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.

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 to these observed characteristics. Table 1 reports these comparisons. We find that the difference in average characteristics are not statistically significant and are generally small in economic magnitude.[19] This provides further evidence towards plausible as-good-as-random assignment of treatment between schools within match-pairs.

Table 1: Balance checks

		(1)		(2)		(1)-(2)
		Control		**Treatment**		**Pairwise t-test**
Variable	N	Mean/(SD)	N	Mean/(SD)	N	Beta/[Wild bootstrapped p-value]
Male	254	0.457 (0.499)	273	0.436 (0.497)	527	0.013 [0.464]
Repeater	254	0.098 (0.298)	273	0.073 (0.261)	527	-0.030 [0.336]
Screening score	254	0.193 (0.074)	273	0.200 (0.071)	527	0.004 [0.708]
Index: aspirations	254	0.042 (0.979)	273	0.107 (0.940)	527	0.160 [0.266]
Index: perceived performance relative to peers	254	-0.011 (0.997)	273	0.069 (1.004)	527	-0.039 [0.744]
Index: finds courses easy	254	0.047 (0.905)	273	0.206 (0.882)	527	0.070 [0.428]
Index: like school courses	254	-0.014 (0.969)	273	0.136 (0.924)	527	0.085 [0.872]
Index: grit	254	0.001 (1.003)	273	0.202 (1.037)	527	0.214 [0.346]
Index: locus of control	254	-0.005 (0.997)	273	0.098 (0.969)	527	0.153 [0.214]
Index: individual well-being	254	0.068 (0.990)	273	0.064 (1.036)	527	-0.018 [0.698]
Index: social well-being	254	0.035 (0.987)	273	0.184 (0.968)	527	0.230 [0.388]
Index: study workdays	254	0.032 (1.010)	273	0.128 (0.980)	527	-0.055 [0.790]
Index: study weekends	254	0.063 (1.019)	273	0.028 (0.957)	527	-0.141 [0.366]
Index: parental investment	254	0.004 (0.950)	273	-0.089 (0.972)	527	-0.081 [0.524]
Reading test score	150	59.356 (19.245)	218	56.445 (21.596)	368	-0.682 [0.880]

*Notes: 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*

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

[19] Table B.3 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.

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 to focus on the qualitative conclusion that, at the bare minimum, 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 non-cognitive 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 the intervention has successfully improved 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 non-cognitive outcomes and educational investments. Table 2's Panels (b)–(e) report the estimates of the impact of the program on non-cognitive 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 8.8, 16.8 and 30.8% 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. Columns (2) and (3) in Table 2 report

Table 2: Estimated effects on reading performance and non-cognitive outcomes

		By risk of Dyslexia		By gender	
		Not-at-risk	At-risk	Female	Male
	(1)	(2)	(3)	(4)	(5)
Panel (a): Reading performance					
Average reading performance	2.594**	1.951	5.125	4.712*	0.540
	[0.038]	[0.106]	[0.466]	[0.096]	[0.804]
# Observations	368	315	53	212	156
Panel (b): Self perceptions					
Aspirations	0.088*	0.048	0.246	0.183	0.049
	[0.054]	[0.362]	[0.540]	[0.488]	[0.576]
Performance rel. peers	0.168**	0.095**	0.331	0.225	0.146
	[0.010]	[0.038]	[0.190]	[0.424]	[0.436]
Courses are easy	0.308**	0.287*	0.340	0.337***	0.348
	[0.014]	[0.050]	[0.186]	[0.004]	[0.144]
Like school	0.159	0.233	-0.144	0.278	0.080
	[0.368]	[0.304]	[0.528]	[0.348]	[0.832]
# Observations	527	455	72	292	235
Panel (c): Personality traits					
Grit	0.248	0.272	0.532	0.218	0.241
	[0.548]	[0.514]	[0.154]	[0.654]	[0.590]
LOC	0.194*	0.119*	0.710**	0.169	0.265
	[0.064]	[0.064]	[0.034]	[0.396]	[0.126]
# Observations	527	455	72	292	235
Panel (d): Subjective well-being					
Individual well-being	0.048	0.137	-0.410	0.060	0.020
	[0.374]	[0.156]	[0.154]	[0.408]	[0.736]
Social well-being	0.203	0.225*	-0.071	0.418**	0.003
	[0.124]	[0.094]	[0.824]	[0.012]	[0.994]
# Observations	527	455	72	292	235
Panel (e): Time and investment					
Study workdays	0.012	-0.017	0.167	-0.072	0.126
	[0.728]	[0.684]	[0.342]	[0.488]	[0.120]
Study weekends	0.093	0.097	0.455***	0.038	0.121*
	[0.308]	[0.334]	[0.004]	[0.856]	[0.064]
Parental investment	0.083	0.133	-0.174	-0.063	0.267
	[0.292]	[0.342]	[0.230]	[0.384]	[0.430]
# Observations	527	455	72	292	235

*Notes: Regressions estimate Equation 1. The set of individual controls includes the baseline value of all the eleven indices listed in Section 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* the effects for the not-at-risk and at-risk groups, respectively. We find that the effects documented in the previous paragraph are present among those who are not-at-risk. Moreover, we have suggestive evidence that the program had larger effects on at-risk students, but the small size of this subpopulation leaves us with little power to make precise conclusions.

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 non-cognitive 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 B.4 and B.5, 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. Though there may be heterogeneous effects along some dimensions, larger sample sizes would be needed to draw firm conclusions regarding their magnitude.

4.2 Robustness

We acknowledge that due to budgetary and logistical constraints, our study design has limitations that makes fall short of the ideal environment for causal identification. However, in this section, we argue that the gains described above for certain dimensions are unlikely to be spurious. For this, we show the robustness of our results to the two most salient threats to internal validity: selective attrition and selection on unobservables.

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 expected in an environment with high rates of absenteeism. 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. Following Muralidharan et al. (2019), we re-estimate our main

Table 3: Robustness of estimated effects on cognitive and non-cognitive outcomes

	Att (1)	No controls (2)	PSM (3)
Panel (a): Reading performance			
Average reading performance	1.963**	3.197*	2.620
	[0.022]	[0.080]	
# Observations	368	416	368
Panel (b): Self perceptions			
Aspirations	0.100**	0.143	0.051
	[0.026]	[0.364]	
Performance rel. peers	0.154**	0.202	0.193
	[0.018]	[0.434]	
Courses are easy	0.307**	0.328*	0.302
	[0.018]	[0.050]	
Like school	0.162	0.177	0.168
	[0.360]	[0.500]	
# Observations	527	527	527
Panel (c): Personality traits			
Grit	0.248	0.266	0.235
	[0.566]	[0.472]	
LOC	0.194*	0.306**	0.262
	[0.064]	[0.038]	
# Observations	527	527	527
Panel (d): Subjective well-being			
Individual well-being	0.046	0.141*	0.035
	[0.464]	[0.058]	
Social well-being	0.208*	0.254	0.139
	[0.068]	[0.338]	
# Observations	527	527	527
Panel (e): Time and investment			
Study workdays	0.003	-0.054	0.024
	[0.912]	[0.400]	
Study weekends	0.084	0.049	0.044
	[0.310]	[0.614]	
Parental investment	0.111	0.140	0.118
	[0.260]	[0.386]	
# Observations	527	527	527

*Notes: "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. Column (3) provides the estimates under propensity score matching based on Masten et al. (2024). 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* 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.

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.

Table 4: Robustness to selection on unobservables

Outcome variable	(1)	(2)
	Oster (2019)'s δ	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

Notes: 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.

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.[20] 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

[20] Further elaboration on the method is available in Appendix D.1.

on unobservables. 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 non-cognitive 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 non-parametric selection as proposed by Masten et al. (2024).[21] 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. To benchmark whether the breakdown points we find are large, we report in Appendix Tables B.6 and B.7 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 B.6, the breakdown point is only higher than 3 of the 17 variables. Consistent with the previous analysis, we find that our results on non-cognitive outcomes are likely to be robust to selection on unobservables while the result on cognition is less likely to be so. The breakdown points for self-perceptions and personality traits are greater than most of the 90th percentile. In particular, the ones corresponding to self-perception that courses are easy and locus of control are higher than 11 and 13 (of 16) of the reference values, respectively.

Alternative estimator: Propensity score matching. A byproduct of the analysis of Masten et al. (2024) is that we also obtain estimates using a different estimation method under the same conditional ignorability assumption. More specifically, their

[21] Further elaboration on the method is available in Appendix D.2.

analysis is based on nonparametric propensity score matching rather than the linear regression adjustment that we perform in our baseline specification. The treatment effect estimates based on propensity score matching are reported in Column (3) of Table 3. We find that the point estimates are very close to those we had obtained from linear regression adjustment in Table 2.

4.3 Discussion

Though we find a sizable and significant effect of the intervention on reading performance, the above analyses suggest that this result may not be robust to selection on unobservables. We, however, find that the economically and statistically significant effects of the intervention on non-cognitive outcomes, particularly on self-perceptions (i.e., perception that one is doing better than others and perception that courses are easy) and personality traits (i.e., locus of control), may be robust to selection on unobservables. Our results complement existing evidence in other settings that non-cognitive skills may be more malleable than cognitive skills (Carneiro et al., 2007; Cunha and Heckman, 2008; Almlund et al., 2011). As the estimation of the dynamics of cognitive and non-cognitive development by Cunha et al. (2010) suggests, non-cognitive skills help foster cognitive skills, but not necessarily the other way around. Thus, we might expect cognitive skills of these children to improve at later stages, mediated by the changes in non-cognitive outcomes we observe. Further study needs to be done to measure the long-run effects of our intervention.

5 Cost-effectiveness and Scalability

Cost-effectiveness. The implementation of the program that we evaluate costs €100 per student. Given the plethora of alternative interventions a policy maker could choose from, it is important to compare the gains per monetary unit spent in our intervention with those of other attractive options. We consider both the cognitive and the non-cognitive margins when evaluating the cost-effectiveness of the program.

In terms of cognitive gains, which are typically the center of attention in educational interventions, we find in Table 2 that our program generates an improvement of 13% of the control group's standard deviation. We compare this gain to two recent successful interventions by Carlana and La Ferrara (2021) and Gortazar et al. (2024). Their respective costs per participant are €50 and €300. To make their estimated gains comparable to ours when accounting for implementation costs, we obtain their impact per €100 — the per student cost of our intervention. Both Carlana and La Ferrara (2021) and Gortazar et al. (2024) quantify the effect of individualized tutoring sessions in mathematics to be of about 26% of a standard deviation. This means that Gortazar et al. (2024) find a gain of 8.7% of a standard deviation per each €100 spent, which they highlight as being a fairly attractive return relative to existing options. Carlana and La Ferrara (2021)'s intervention is able to substantially improve on Gortazar et al. (2024), as they generate a gain of 52% of a standard deviation for each €100 spent. The returns of our intervention are in-between these two. While this is therefore an already attractive result, one should keep in mind that our results are for reading, a domain for which it is typically harder to achieve gains than in mathematics (Dietrichson et al., 2021).

Moreover, we highlight that arguably the main benefits from our intervention come in terms of non-cognitive skills. We estimate improvements of between 10% to 30% of a standard deviation in a wide range of skills. Carlana and La Ferrara (2021) find comparable effects to us per €100 (e.g., 28% of a standard deviation in grid and 34% in well-being). Gortazar et al. (2024) find gains of about 4% of a standard deviation for aspirations. As such, when we factor in these gains, our intervention emerges as a cost-effective option.

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 towards Dytective. Moreover, although further research is needed to evaluate whether having professional educational psychologists implement the program strengthens the effectiveness of the intervention, in the event of the scarcity of these professionals, regular class teachers should be able to set up the sessions — this is being done, for instance, in the majority of schools in Madrid where Dytective is in place. 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, the library could instead use the electronic version of the books to eliminate the need for a person to visit the schools.

6 Conclusion

Closing educational gaps is still a major challenge for policymakers around the world. We evaluate a reading intervention that relies on computer-generated adaptive exercises that not only target the whole distribution of initial reading abilities (which tends to have a large support even among students in the same classroom), but also the deficiencies that typically constrain the performance of individuals with dyslexia — a sizable subpopulation often documented to struggle academically despite not possessing lower cognitive abilities. Given the multiplicity of factors that curtail learning in developing countries, the intervention also features complementary programs aiming at fostering participants' non-cognitive skills as well as parental involvement in the learning of their children.

This evaluation is, to the best of our knowledge, the first exploration of the cognitive and non-cognitive impacts for both dyslexic and non-dyslexic children of a reading intervention centered around an education technology. In line with general results in the literature, we find suggestive evidence that the program improves academic performance. The education technology's personalized learning feature is likely behind the fact that the effects are present throughout the ability distribution. Importantly, unlike existing work, which typically does not study non-cognitive outcomes or finds little effects in them, we show that our intervention meaningfully improves a range of non-cognitive skills and perceptions. We find evidence that self-confidence, locus-of-control and aspirations are enhanced for students at-risk of dyslexia. This is a group for whom there is evidence that these characteristics tend to be lacking, which matters because these characteristics are predictive of low academic performance and higher risk of grade repetition. Our findings are encouraging not only because the intervention seems to provide both cognitive and non-cognitive benefits for most of the students in the class at a relatively low cost, but also because it is able to affect the less-studied and harder-to-reach group of at-risk-of-dyslexia students.

References

Alan, Sule, Teodora Boneva, and Seda Ertac (2019) "Ever Failed, Try Again, Succeed Better: Results from a Randomized Educational Intervention on Grit*," *The Quarterly Journal of Economics*, 134 (3), 1121–1162, 10.1093/qje/qjz006. (return to page 5)

Alan, Sule and Ipek Mumcu (2024) "Nurturing Childhood Curiosity to Enhance Learning: Evidence from a Randomized Pedagogical Intervention," *American Economic Review*, 114 (4), 1173–1210, 10.1257/aer.20230084. (return to page 5)

Ally, Mohamed, Venkataraman Balaji, Anwar Abdelbaki, and Ricky Cheng (2017) "Use of Tablet Computers to Improve Access to Education in a Remote Location," *Journal of Learning for Development*, 4 (2), 10.56059/jl4d.v4i2.219. (return to page 21)

Almlund, Mathilde, Angela Lee Duckworth, James Heckman, and Tim Kautz (2011) "Chapter 1 - Personality Psychology and Economics," in Hanushek, Eric A., Stephen Machin, and Ludger Woessmann 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. (return to page 4, 20)

American Psychiatric Association, DSMTF, American Psychiatric Association et al. (2013) *Diagnostic and statistical manual of mental disorders: DSM-5*, 5: American Psychiatric Association Washington, DC. (return to page 1)

Anderson, Michael L. (2008) "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American Statistical Association*, 103 (484), 1481–1495, 10.1198/016214508000000841. (return to page 9)

Andrabi, Tahir, Jishnu Das, Asim Ijaz Khwaja, and Tristan Zajonc (2011) "Do Value-Added Estimates Add Value? Accounting for Learning Dynamics," *American Economic Journal: Applied Economics*, 3 (3), 29–54, 10.1257/app.3.3.29. (return to page 3, 12)

Ashraf, Nava, Natalie Bau, Corinne Low, and Kathleen McGinn (2020) "Negotiating a Better Future: How Interpersonal Skills Facilitate Intergenerational Investment*," *The Quarterly Journal of Economics*, 135 (2), 1095–1151, 10.1093/qje/qjz039. (return to page 5)

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukherji, Marc Shotland, and Michael Walton (2016) "Mainstreaming an effective intervention: Evidence from randomized evaluations of "Teaching at the Right Level" in India,"Technical report, National Bureau of Economic Research. (return to page 4)

Banerjee, Abhijit V., Shawn Cole, Esther Duflo, and Leigh Linden (2007) "Remedying Education: Evidence from Two Randomized Experiments in India*," *The Quarterly Journal of Economics*, 122 (3), 1235–1264, 10.1162/qjec.122.3.1235. (return to page 4)

Beg, Sabrin, Waqas Halim, Adrienne M. Lucas, and Umar Saif (2022) "Engaging Teachers with Technology Increased Achievement, Bypassing Teachers Did Not," *American*

Economic Journal: Economic Policy, 14 (2), 61–90, 10.1257/pol.20200713. (return to page 4)

Blackwell, Lisa S., Kali H. Trzesniewski, and Carol Sorich Dweck (2007) "Implicit Theories of Intelligence Predict Achievement Across an Adolescent Transition: A Longitudinal Study and an Intervention," *Child Development*, 78 (1), 246–263, https://doi.org/10.1111/j.1467-8624.2007.00995.x. (return to page 2)

Bruhn, Miriam and David McKenzie (2009) "In Pursuit of Balance: Randomization in Practice in Development Field Experiments," *American Economic Journal: Applied Economics*, 1 (4), 200–232, 10.1257/app.1.4.200. (return to page 7)

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller (2008) "Bootstrap-Based Improvements for Inference with Clustered Errors," *The Review of Economics and Statistics*, 90 (3), 414–427, 10.1162/rest.90.3.414. (return to page 3, 12)

Carlana, Michela and Eliana La Ferrara (2021) "Apart but connected: Online tutoring and student outcomes during the COVID-19 pandemic." (return to page 2, 4, 11, 20, 21)

Carneiro, Pedro, Claire Crawford, and Alissa Goodman (2007) "The impact of early cognitive and non-cognitive skills on later outcomes," Centre for Economics of Education Working Paper. (return to page 20)

Carneiro, Pedro, Emanuela Galasso, Italo Lopez Garcia, Paula Bedregal, and Miguel Cordero (2024) "Impacts of a Large-Scale Parenting Program: Experimental Evidence from Chile," *Journal of Political Economy*, 132 (4), 1113–1161, 10.1086/727288. (return to page 4)

Cortiella, Candace and Sheldon H Horowitz (2014) "The state of learning disabilities: Facts, trends and emerging issues," *New York: National Center for Learning Disabilities*, 25 (3), 2–45. (return to page 1)

Cuevas-Ruiz, Pilar, Luz Rello, Ismael Sanz, and Almudena Sevilla (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. (return to page 1)

Cunha, Flavio and James J. Heckman (2008) "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Journal of Human Resources*, 43 (4), 738–782, 10.3368/jhr.43.4.738. (return to page 20)

Cunha, Flavio, James J. Heckman, and Susanne M. Schennach (2010) "Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Econometrica*, 78 (3), 883–931, https://doi.org/10.3982/ECTA6551. (return to page 4, 20)

Dietrichson, Jens, Trine Filges, Julie K. Seerup, Rasmus H. Klokker, Bjørn C. A. Viinholt, Martin Bøg, and Misja Eiberg (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 Systematic Reviews*, 17 (2), e1152, https://doi.org/10.1002/cl2.1152. (return to page 21)

Dillon, Moira R, Harini Kannan, Joshua T Dean, Elizabeth S Spelke, and Esther Duflo (2017) "Cognitive science in the field: A preschool intervention durably enhances intuitive but not formal mathematics," *Science*, 357 (6346), 47–55, 10.1126/science.aal4724. (return to page 4)

Duflo, Esther, Rachel Glennerster, and Michael Kremer (2007) "Chapter 61 Using Randomization in Development Economics Research: A Toolkit," in Schultz, T. Paul and John A. Strauss eds. *Handbook of Development Economics*, 4, 3895–3962: Elsevier, https://doi.org/10.1016/S1573-4471(07)04061-2. (return to page 14)

Escueta, Maya, Andre Joshua Nickow, Philip Oreopoulos, and Vincent Quan (2020) "Upgrading Education with Technology: Insights from Experimental Research," *Journal of Economic Literature*, 58 (4), 897–996, 10.1257/jel.20191507. (return to page 1, 4)

Galuschka, Katharina, Ruth Görgen, Julia Kalmar, Stefan Haberstroh, Xenia Schmalz, and Gerd Schulte-Körne (2020) "Effectiveness of spelling interventions for learners with dyslexia: A meta-analysis and systematic review," *Educational Psychologist*, 55 (1), 1–20, 10.1080/00461520.2019.1659794. (return to page 4)

Galuschka, Katharina, Elena Ise, Kathrin Krick, and Gerd Schulte-Körne (2014) "Effectiveness of treatment approaches for children and adolescents with reading disabilities: A meta-analysis of randomized controlled trials," *PloS one*, 9 (2), e89900, https://doi.org/10.1371/journal.pone.0089900. (return to page 4)

Glewwe, P. and K. Muralidharan (2016) "Chapter 10 - Improving Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications," in Hanushek, Eric A., Stephen Machin, and Ludger Woessmann eds. *Handbook of the Economics of Education*, 5, 653–743: Elsevier, https://doi.org/10.1016/B978-0-444-63459-7.00010-5. (return to page 1)

Gortazar, Lucas, Claudia Hupkau, and Antonio Roldán-Monés (2024) "Online tutoring works: Experimental evidence from a program with vulnerable children," *Journal of Public Economics*, 232, 105082, https://doi.org/10.1016/j.jpubeco.2024.105082. (return to page 4, 20, 21)

Graham, Jimmy and Sean Kelly (2019) "How effective are early grade reading interventions? A review of the evidence," *Educational Research Review*, 27, 155–175, https://doi.org/10.1016/j.edurev.2019.03.006. (return to page 4)

Guarino, Cassandra M., Mark D. Reckase, and Jeffrey M. Wooldridge (2015) "Can Value-Added Measures of Teacher Performance Be Trusted?," *Education Finance and Policy*, 10 (1), 117–156, 10.1162/EDFP_a_00153. (return to page 12)

Heckman, James J and Dimitriy V Masterov (2007) "The productivity argument for investing in young children." (return to page 1)

Heckman, James Joseph (2020) *Randomization and Social Policy Evaluation Revisited*: National Bureau of Economic Research Cambridge, MA. (return to page 7)

Kautz, Tim, James J Heckman, Ron Diris, Bas ter Weel, and Lex Borghans (2014) "Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success," Working Paper 20749, National Bureau of Economic Research, 10.3386/w20749. (return to page 2)

Kim, Young-Suk G., Hansol Lee, and Stephanie S. Zuilkowski (2020) "Impact of Literacy Interventions on Reading Skills in Low- and Middle-Income Countries: A Meta-Analysis," *Child Development*, 91 (2), 638–660, https://doi.org/10.1111/cdev.13204. (return to page 4)

Lafortune, Jeanne, Todd Pugatch, José Tessada, and Diego Ubfal (2024) "Can gamified online training make high school students more entrepreneurial? Experimental evidence from Rwanda," *Economics of Education Review*, 101, 102559, https://doi.org/10.1016/j.econedurev.2024.102559. (return to page 4)

Masten, Matthew A., Alexandre Poirier, and Linqi Zhang (2024) "Assessing Sensitivity to Unconfoundedness: Estimation and Inference," *Journal of Business & Economic Statistics*, 42 (1), 1–13, 10.1080/07350015.2023.2183212. (return to page 3, 17, 18, 19, A8, A9, A11, A12, A13)

Muralidharan, Karthik, Abhijeet Singh, and Alejandro J. Ganimian (2019) "Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India," *American Economic Review*, 109 (4), 1426–60, 10.1257/aer.20171112. (return to page 4, 11, 16)

O'Neill, Stephen, Noémi Kreif, Richard Grieve, Matthew Sutton, and Jasjeet S. Sekhon (2016) "Estimating causal effects: considering three alternatives to difference-in-differences estimation," *Health Services and Outcomes Research Methodology*, 16 (1), 1–21, 10.1007/s10742-016-0146-8. (return to page 12)

Oster, Emily (2019) "Unobservable Selection and Coefficient Stability: Theory and Evidence," *Journal of Business & Economic Statistics*, 37 (2), 187–204, 10.1080/07350015.2016.1227711. (return to page 3, 18, 19, A11, A12, A13)

Pritchett, Lant (2013) *The rebirth of education: Schooling ain't learning*: CGD Books. (return to page 1)

Rello, Luz, Ricardo Baeza-Yates, Abdullah Ali, Jeffrey P Bigham, and Miquel Serra (2020) "Predicting risk of dyslexia with an online gamified test," *Plos one*, 15 (12), e0241687, https://doi.org/10.1371/journal.pone.0241687. (return to page 6)

Rello, Luz, Arturo Macias, Mariía Herrera, Camila de Ros, Enrique Romero, and Jeffrey P. Bigham (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, 10.1145/3132525.3134773. (return to page 6)

Resnjanskij, Sven, Jens Ruhose, Simon Wiederhold, Ludger Woessmann, and Katharina Wedel (2024) "Can Mentoring Alleviate Family Disadvantage in Adolescence? A Field Experiment to Improve Labor Market Prospects," *Journal of Political Economy*, 132 (3), 1013–1062, 10.1086/726905. (return to page 1)

Scammacca, Nancy K., Garrett J. Roberts, Eunsoo Cho, Kelly J. Williams, Greg Roberts, Sharon R. Vaughn, and Megan Carroll (2016) "A Century of Progress: Reading Interventions for Students in Grades 4–12, 1914–2014," *Review of Educational Research*, 86 (3), 756–800, 10.3102/0034654316652942, PMID: 28529386. (return to page 4)

Shaywitz, Sally E (1998) "Dyslexia," *New England Journal of Medicine*, 338 (5), 307–312, 10.1056/NEJM199801293380507. (return to page 11)

Singer, Elly (2008) "Coping with academic failure, a study of Dutch children with dyslexia," *Dyslexia*, 14 (4), 314–333, https://doi.org/10.1002/dys.352. (return to page 1)

Singh, Abhijeet (2015) "Private school effects in urban and rural India: Panel estimates at primary and secondary school ages," *Journal of Development Economics*, 113, 16–32, https://doi.org/10.1016/j.jdeveco.2014.10.004. (return to page 3, 12)

——— (2020) "Learning More with Every Year: School Year Productivity and International Learning Divergence," *Journal of the European Economic Association*, 18 (4), 1770–1813, 10.1093/jeea/jvz033. (return to page 1)

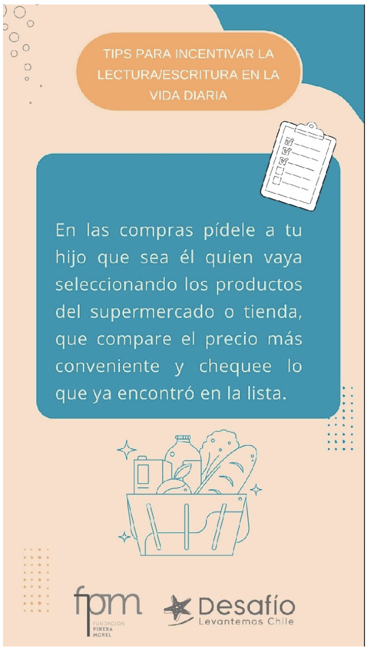
Sisk, Victoria F., Alexander P. Burgoyne, Jingze Sun, Jennifer L. Butler, and Brooke N. Macnamara (2018) "To What Extent and Under Which Circumstances Are Growth Mind-Sets Important to Academic Achievement? Two Meta-Analyses," *Psychological Science*, 29 (4), 549–571, 10.1177/0956797617739704, PMID: 29505339. (return to page 5)

Todd, Petra E. and Kenneth I. Wolpin (2003) "On the Specification and Estimation of the Production Function for Cognitive Achievement," *The Economic Journal*, 113 (485), F3–F33, https://doi.org/10.1111/1468-0297.00097. (return to page 3, 12)

Webb, Matthew D. (2023) "Reworking wild bootstrap-based inference for clustered errors," *Canadian Journal of Economics/Revue canadienne d'économique*, 56 (3), 839–858, https://doi.org/10.1111/caje.12661. (return to page 12, 13, 15, 17, A7)

A Online Appendix: Figures

Figure A1: Sample text message sent to parents



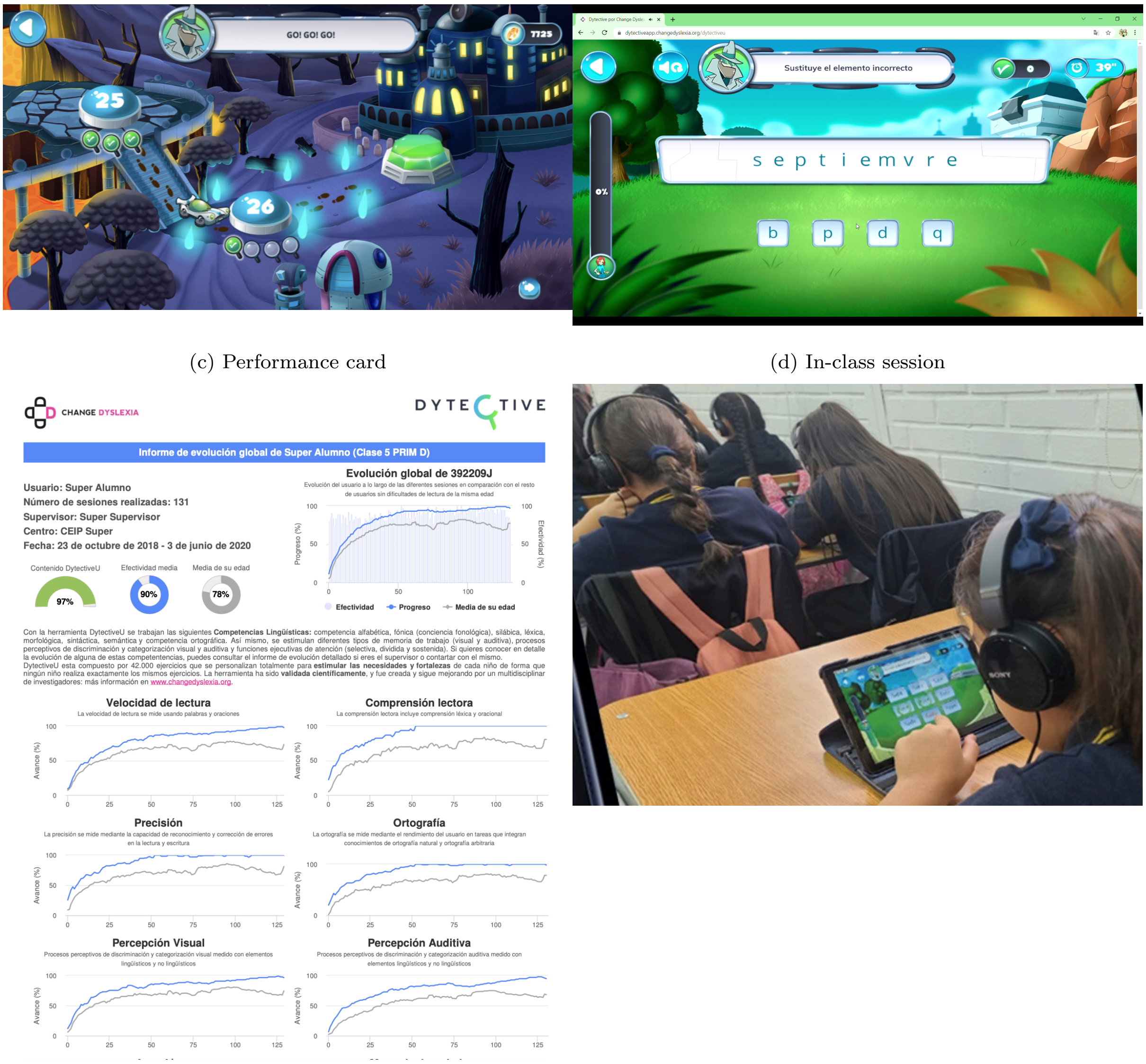
TIPS PARA INCENTIVAR LA LECTURA/ESCRITURA EN LA VIDA DIARIA

En las compras pídele a tu hijo que sea él quien vaya seleccionando los productos del supermercado o tienda, que compare el precio más conveniente y chequee lo que ya encontró en la lista.

fpm
FUNDACIÓN PIÑERA MOREL
Desafío
Levantemos Chile

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."

Figure A2: Contextual details



(a) Game interface

(b) Exercise interface

(c) Performance card

(d) In-class session

GO! GO! GO!
7725
25
26
Dytective por Change Dyslex
dytectiveapp.changedyslexia.org/dytectiveu
Sustituye el elemento incorrecto
0
39"
s e p t i e m v r e
0%
b
p
d
q
SONY

CHANGE DYSLEXIA

DYTECTIVE

Informe de evolución global de Super Alumno (Clase 5 PRIM D)

Usuario: Super Alumno
Número de sesiones realizadas: 131
Supervisor: Super Supervisor
Centro: CEIP Super
Fecha: 23 de octubre de 2018 - 3 de junio de 2020

Con la herramienta DytectiveU se trabajan las siguientes **Competencias Lingüísticas:** competencia alfabética, fónica (conciencia fonológica), silábica, léxica, morfológica, sintáctica, semántica y competencia ortográfica. Así mismo, se estimulan diferentes tipos de memoria de trabajo (visual y auditiva), procesos perceptivos de discriminación y categorización visual y auditiva y funciones ejecutivas de atención (selectiva, dividida y sostenida). Si quieres conocer en detalle la evolución de alguna de estas competentencias, puedes consultar el informe de evolución detallado si eres el supervisor o contartar con el mismo.
DytectiveU esta compuesto por 42.000 ejercicios que se personalizan totalmente para **estimular las necesidades y fortalezas** de cada niño de forma que ningún niño realiza exactamente los mismos ejercicios. La herramienta ha sido **validada científicamente**, y fue creada y sigue mejorando por un multidisciplinar de investigadores: más información en www.changedyslexia.org.

Contenido DytectiveU
97%
Efectividad media
90%
Media de su edad
78%
Evolución global de 392209J
Evolución del usuario a lo largo de las diferentes sesiones en comparación con el resto de usuarios sin dificultades de lectura de la misma edad
Progreso (%)
Efectividad (%)
0
50
100
Efectividad
Progreso
Media de su edad
Velocidad de lectura
La velocidad de lectura se mide usando palabras y oraciones
Comprensión lectora
La comprensión lectora incluye comprensión léxica y oracional
Precisión
La precisión se mide mediante la capacidad de reconocimiento y corrección de errores en la lectura y escritura
Ortografía
La ortografía se mide mediante el rendimiento del usuario en tareas que integran conocimientos de ortografía natural y ortografía arbitraria
Percepción Visual
Procesos perceptivos de discriminación y categorización visual medido con elementos lingüísticos y no lingüísticos
Percepción Auditiva
Procesos perceptivos de discriminación y categorización auditiva medido con elementos lingüísticos y no lingüísticos
Avance (%)
0
25
50
75
100
125

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 Online Appendix: Tables

Table B.1: 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
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
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

Notes: 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 B.2: 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
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
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 B.1 employing everybody who is available at baseline, irrespective of whether they are eventually observed at endline.

Table B.3: Balance check: full sample

	(1)		(2)		(1)-(2)	
	Control		**Treatment**		**Pairwise t-test**	
Variable	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]
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]
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.

Table B.4: Heterogeneity by at-risk status

	(1) Aspirations	(2) Performance rel. peers	(3) Courses are easy	(4) Like school	(5) Grit	(6) 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) Indiv. well-being	(8) Social well-being	(9) Study workdays	(10) Study weekends	(11) Parental investment	(12) 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 B.5: Heterogeneity by gender

	(1) Aspirations	(2) Performance rel. peers	(3) Courses are easy	(4) Like school	(5) Grit	(6) 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) Indiv. well-being	(8) Social well-being	(9) Study workdays	(10) Study weekends	(11) Parental investment	(12) 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 B.6: Variation in leave-out-variable-k change in propensity scores, average reading performance

	Quantiles			
	p50	p75	p90	max
Variable Name	(1)	(2)	(3)	(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	0.052	0.093
Ever repeater	0.019	0.031	0.056	0.165
Study workday	0.023	0.041	0.062	0.149
Study weekend	0.024	0.044	0.073	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 B.7: Variation in leave-out-variable-k change in propensity scores, non-cognitive outcomes

	Quantiles			
	p50	p75	p90	max
Variable Name	(1)	(2)	(3)	(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.

C Online 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 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.

D Online 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 of 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^2_{max}, 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^2_{max}. We set $R^2_{\text{max}} = 1.5 \times R^2_u$ 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 nonparametric 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 $X = x$,

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

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 B.6 and B.7.