IZA Institute of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16937

# Game Changer: Impact of a Reading Intervention on Cognitive and Non-cognitive Skills

Micole De Vera
Javier Garcia-Brazales
Luz Rello

APRIL 2024

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 16937

# Game Changer: Impact of a Reading Intervention on Cognitive and Non-cognitive Skills

**Micole De Vera**
*University College London and IZA*

**Javier Garcia-Brazales**
*University of Exeter and LEAP*

**Luz Rello**
*IE University*

APRIL 2024



ISSN: 2365-9793

**IZA – Institute of Labor Economics**

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

# ABSTRACT

## Game Changer: Impact of a Reading Intervention on Cognitive and Non-cognitive Skills*

We evaluate a 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. We first quantify the impact on non-cognitive skills and academic perceptions. We find that, after just three months of intervention, treated students are 20–30 percent of a standard deviation more likely to believe that their performance is better than that of their peers, to like school, to have stronger grit, and to have a more internal locus-of-control. Gains in aspirations and self-confidence are particularly large for students that we identify as at-risk-of-dyslexia. These improvements are reflected in better performance on a nation-wide, standardized language test. Our results show that non-cognitive skills, particularly of at-risk-of-dyslexia students, can be changed through a short, light-touch, and cost-effective education technology intervention.

**JEL Classification:** I24, I31

**Keywords:** field experiment, computer-based reading intervention, non-cognitive skills, Chile, dyslexia

**Corresponding author:**
Micole De Vera
University College London
Drayton House
30 Gordon St
London WC1H 0AN
United Kingdom
E-mail: micole.devera.work@gmail.com

* 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.

# 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, a country consistently ranked among Latin America's top performers in standardized tests such as PISA, 60% of second grade students lag behind their expected reading level by at least 6 months.[1] Strengthening academic progress is difficult because it is not enough to address institutional constraints like teacher training or quality of educational materials. Issues such as lack of student motivation or aspirations and limited parental investment are also serious deterrents to academic performance (e.g., Heckman and Masterov, 2007; Resnjanskij et al., 2024). Moreover, institutions may be unprepared to support students with learning disorders that lead them to leak out of the educational system. 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 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 cultivates 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, and student-specific "challenges" — composed of a number of exercises — are generated based on past performance in the app and adapted to improve previous weaknesses in certain linguistic areas and cognitive skills. The application was designed to enhance, among others, the spelling, reading speed, and vocabulary of the

[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.

participants.

Dytective is played during 45-minute school visits by a psychopedagogue three times a week for three months. In these sessions, which take place during regular Spanish language classes, the psychopedagogue 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. Though all the components of the program are geared towards improving reading, they also contain features that may inadvertently generate gains in non-cognitive skills such as concentration, grit, and self-confidence, especially as reading ability strengthens. 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*[3] is a key feature of our program.

This study 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, we take 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. 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 the Chilean governmental institutions, size, and location. This research design relies on potential outcomes of students being the same within the match-pair. Reassuringly, we find that, within the resulting match-pairs, treatment and control participants are indeed balanced across a wide range of predetermined characteristics and baseline measures of outcomes that were not used in the matching, which suggests that participants are plausibly balanced in unobservables. With this support for our identification assumption, we first explore the impact of the intervention on non-cognitive skills and perceptions elicited through an ad hoc survey that we designed and distributed before and after the intervention. We then evaluate the impact on reading ability as measured by a standardized national reading test held three times per year by the Chilean Education Quality Assurance Agency.

[3] 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).

We find that students in treated schools display higher self-perceived academic performance, more taste for school, stronger grit and a more internal locus-of-control. The effects are economically large, being of around 20–30% of a standard deviation, and map into higher overall self-reported well-being. These impacts are present both for students that are at risk of having dyslexia and those that are not, and are complemented by higher parental investments (i.e., care about the child's academics and time devoted to helping the child with homework). All these changes translate into an improvement in reading performance of about 18% of a standard deviation. Our results are robust to a wide range of checks, and provide relevant evidence to inform the debate on how to improve literacy, reading ability, and non-cognitive skills in constrained environments.

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 a recent detailed review see Scammacca et al., 2016), 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. *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 large gains in grit, locus-of-control and well-being, 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).[4] *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

[4]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 psychopedagogue).

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.[5] By finding that both groups benefit from the program, but that at-risk students experience larger improvements in self-perceived performance, perceived easiness of the school subjects, locus-of-control, and actual reading ability, our work offers valuable policy lessons to promote inclusive 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, 2022), 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.

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, shows their robustness, and discusses the cost-effectiveness of the intervention. Section 5 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

[5] 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 exclusively evaluates the impact of the intervention on dyslexic individuals and does not extend to a wide range of non-cognitive skills.

provide individualized psychopedagogic 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.[6] The platform is displayed as a game where the main character needs to complete exercises to progress in their quest. Exercises are grouped into linguistic challenges, which are *adaptively* generated based on the player's past performance in the application, with a focus on weaker linguistic areas and general cognitive abilities, e.g., working memory and attention. Each challenge takes around 20 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.[7] 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.[8] For three months, a psychopedagogue visits students in their school three times per week during regular class hours, equips them with individual tablets to access their personal Dytective profile, and guides their use of Dytective throughout 45-minute-long sessions.[9] 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 psychopedagogue is around to provide encouragement and to solve questions, if needed. Following Dytective's recommendations, students aim to complete two challenges per session.[10]

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

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

[7] 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 diagnosis still has to be done by a professional).

[8] Appendix C.2 provides more details on the timeline of the intervention.

[9] 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.

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

and motivate their children with their reading and writing. Appendix Figure A.1 provides an example of how the text messages look like. Appendix Figure A.2 shows how students work individually in class with Dytective, and displays the appearance of both the interface of the game and the back-end recording the evolution of the participant along the various reading and cognitive skills dimensions.

## 2.2 Study design

We evaluate the impact of the program on students enrolled in third grade at five schools in high-vulnerability areas of the Chilean Metropolitan Region that participated during the second semester of 2023. The Metropolitan Region, which includes Santiago, agglomerates most commercial and administrative centers of the country, and is home to around 40% of the country's population. Though our implementing partner would like to extend the program to all schools catering to vulnerable children, the program is implemented in small batches due to financial and logistic constraints. Each batch tends to be locally clustered to optimize on FPM's resources (e.g., the psychopedagogue'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.[11] 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. 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

[11] Based on 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 (relative to the ten schools in our study) would equally benefit from such an intervention.

frame we are interested in, therefore making treatment status within each match-pair be essentially random.

**Main identification assumption.** Identification based on our study design relies on the following assumption: within each match-pair, without spillovers or interference, the students in the treated and control schools have the same potential outcomes. That is, the endline outcomes are expected to be similar between the schools in the absence of the treatment. Moreover, expected endline outcomes should also be similar in the presence of the treatment. As the similarity of potential outcomes is fundamentally untestable, we aim to obtain match-pairs for which this assumption holds by matching based on baseline observable characteristics. We do this for the three school-level summary measures employed by FPM in determining eligibility for program participation: educational vulnerability index, size, and location. In the balance checks, we confirm that students in treatment and control schools also match along demographics and multiple dimensions of cognitive and non-cognitive skills, which were untargeted characteristics when matching. This strengthens the credibility of our identification assumption: within each match-pair, schools are not only similar in baseline matching characteristics, but are also similar in baseline characteristics not used in matching, so similarity in unobserved potential outcomes is highly plausible.

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

**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 library components of the intervention during regular school time, while no student in the control schools had access to either.[12] However, whether parents received the text messages with tips depended on them voluntarily joining the group after having been informed about its existence in a regular teacher-parents meeting prior to the start of the program.[13] 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 and empirical approach

## 3.1 Data

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

**Survey on skills, attitudes, and beliefs.** We designed computer-based surveys to be distributed to all students in treatment and control schools both before the intervention and at its conclusion. These surveys were filled up during class time by all students present at school on the day of the delivery.[14] Although the surveys rely on already-validated survey items, an attractive feature of the survey delivery given the young age of the respondents is that students were aided by psychopedagogues (who were acting as enumerators) to make sure that the questions were clearly understood.[15] 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. The goal of this survey was to measure a wide range of non-cognitive skills and attitudinal skills of students that we expected to be malleable after an intervention of this kind (e.g., self-confidence and taste for school).

Most of the questions asked elicited the level of agreement with a statement. Those answers were on a five-point scale with options "not at all," "a bit," "somewhat," "quite a bit," and "a lot." We reverse the scale, when appropriate, to make the individual items within a family point towards the same direction (i.e., increasing values reflect better

[12] 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.

[13] 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.

[14] Our 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 4.3, this is unlikely to be a concern.

[15] More specifically, psychopedagogues were present during survey completion to explain each survey item one by one and ensured that all students had completed the item at hand before moving onto the next one.

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."

**Complementary data sources: Risk of dyslexia and academic performance.** 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.

As a final source of data, we link students' survey responses to information on their performance in 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 is distributed three times per year, including both in August and late November/early December, hence nicely offering pre- and post-intervention measurements. We study performance in the three main competences tested (ability to locate information, ability to interpret information, and ability to reflect on a text's content) and in the overall score — all of them range from 0 to 100.

**Response coverage.** According to official school census records, at the start of the intervention, a total of 867 students across control and treatment schools were enrolled in third grade. 723 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 595 students (69% of the target population) also completed the endline survey and can therefore be used to quantify the 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 419 out of the 595 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.[16] 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) 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 3.56 on a 1–5 scale.

**Match-pair balance in observables.** To verify that treatment and control schools share similar characteristics, Table 1 reports the comparison of baseline values of all our primary outcomes and key background controls for our estimating sample. We find that the difference in average characteristics are generally not statistically significant and small in economic magnitude. Out of the fifteen dimensions explored, only the index for social well-being is statistically different between the two groups, which may plausibly arise naturally given the large number of comparisons explored.[17] In any case, our robustness checks show that controlling for the value of this dimension at baseline (and also for all the other indices to further increase precision) does not alter our results.

[16] The counterpart for the full sample available at baseline (i.e., regardless of attrition at the endline) is provided in Appendix Table B.2. As one can see, the descriptive statistics are very similar, and the attrition rate of the 723 students that completed the baseline survey was 1 - 595/723 $\approx$ 18%.

[17] 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.

Table 1: Balance checks

| | (1) | | (2) | | (1)-(2) | |
|---|---|---|---|---|---|---|
| | **Control** | | **Treatment** | | **Pairwise t-test** | |
| Variable | N | Mean/(SE) | N | Mean/(SE) | N | Beta/(SE) |
| Male | 287 | 0.474 | 308 | 0.425 | 595 | 0.011 |
| | | (0.095) | | (0.102) | | (0.018) |
| Repeater | 287 | 0.094 | 308 | 0.081 | 595 | 0.015 |
| | | (0.021) | | (0.021) | | (0.019) |
| Screening score | 257 | 0.193 | 274 | 0.200 | 531 | -0.004 |
| | | (0.007) | | (0.004) | | (0.004) |
| Index: aspirations | 287 | 0.062 | 308 | 0.071 | 595 | -0.067 |
| | | (0.121) | | (0.066) | | (0.072) |
| Index: perceived performance relative to peers | 287 | -0.024 | 308 | 0.058 | 595 | -0.001 |
| | | (0.095) | | (0.067) | | (0.073) |
| Index: finds courses easy | 287 | 0.016 | 308 | 0.184 | 595 | -0.097 |
| | | (0.074) | | (0.051) | | (0.070) |
| Index: like school courses | 287 | -0.035 | 308 | 0.132 | 595 | -0.078 |
| | | (0.137) | | (0.089) | | (0.168) |
| Index: grit | 287 | 0.001 | 308 | 0.149 | 595 | -0.144 |
| | | (0.134) | | (0.048) | | (0.105) |
| Index: locus of control | 287 | 0.006 | 308 | 0.094 | 595 | -0.083 |
| | | (0.113) | | (0.076) | | (0.065) |
| Index: individual well-being | 287 | 0.015 | 308 | 0.059 | 595 | -0.006 |
| | | (0.019) | | (0.089) | | (0.042) |
| Index: social well-being | 287 | 0.048 | 308 | 0.164 | 595 | -0.172** |
| | | (0.121) | | (0.109) | | (0.065) |
| Index: study workdays | 287 | -0.009 | 308 | 0.138 | 595 | 0.005 |
| | | (0.086) | | (0.081) | | (0.100) |
| Index: study weekends | 287 | 0.017 | 308 | 0.034 | 595 | 0.092 |
| | | (0.072) | | (0.074) | | (0.127) |
| Index: parental investment | 287 | -0.014 | 308 | -0.080 | 595 | 0.107 |
| | | (0.039) | | (0.072) | | (0.061) |
| Reading test score | 170 | 59.422 | 249 | 55.344 | 419 | 1.471 |
| | | (2.634) | | (3.851) | | (3.595) |

*Notes: The table documents, for the main predetermined variables and indices, their mean and standard error (SE) 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) and its statistical significance (*** p<0.01, ** p<0.05, * p<0.1).*

### 3.2 Regression framework

Under our identification assumption, we estimate the effects of our intervention using the following regression:

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

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 $\delta$ (indexed by $p$ and only dependent on which school the individual goes to). Including the match-pair fixed effects is crucial as it allows us to make within-pair comparisons of treated and non-treated students. The identification assumption for $\beta$ — that treatment assignment is as good as random within each match-pair — is strongly supported by the balance checks in Table 1. To increase precision, our preferred specification controls for $X$, a vector containing individual-level characteristics (gender, repeater status, age, initial risk of dyslexia[18]) and month of survey completion. In Section 4.3, we document the stability of our results both to the exclusion of the individual controls in $X$ and to the inclusion of additional controls to it. We cluster standard errors at the school level and document the significance of the treatment effects to alternative choices for inference.

Given the design of the program, we interpret $\beta$ as the 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. 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 text messages with tips for helping their children with their reading.

# 4 Results

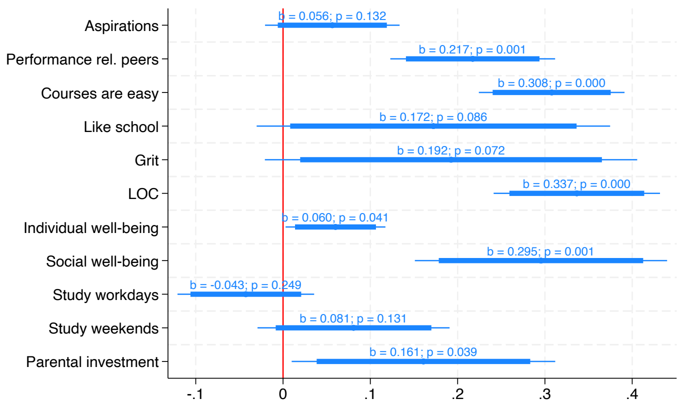
This section reports the main results of the paper. We first quantify the treatment effects of our intervention and provide a discussion on its cost-effectiveness and policy implications. Then, we show that our main findings are robust to a number of potential threats to identification and variations in our empirical choices.

[18] To not lose the 64 observations for which we have both baseline and endline indices but no measure of risk of dyslexia, we replace missing values for risk of dyslexia by a number outside of the variable's support (specifically, 99) and additionally control for an indicator signaling the observations for which such replacement has been implemented. We verify that the main results hold when we (1) do not control for risk of dyslexia and (2) control for risk of dyslexia but do not include the indicator for this variable being missing — i.e., we use the 531 observations for which we have both baseline and endline indices and the dyslexia risk measure.

## 4.1 Treatment effects

**Main results.** Figure 1 reports the estimates of the impact of the program on our primary outcomes of interest. The graphical representation makes it clear that there is a substantial and widespread rightwards shift in our non-cognitive and attitudinal measures among students in treated schools. For instance, participants display higher perceived performance relative to their peers and higher perceived easiness of school subjects. They also develop a stronger grit and taste for school, a more inner-based locus-of-control, and have higher well-being. The magnitudes of these effects are large, at around 20 to 30% of a standard deviation. We also find evidence that parents invest more on children. We do not detect substantial changes in effort — if anything, there may be a substitution between weekday and weekend time devoted to school work.

Figure 1: Main results



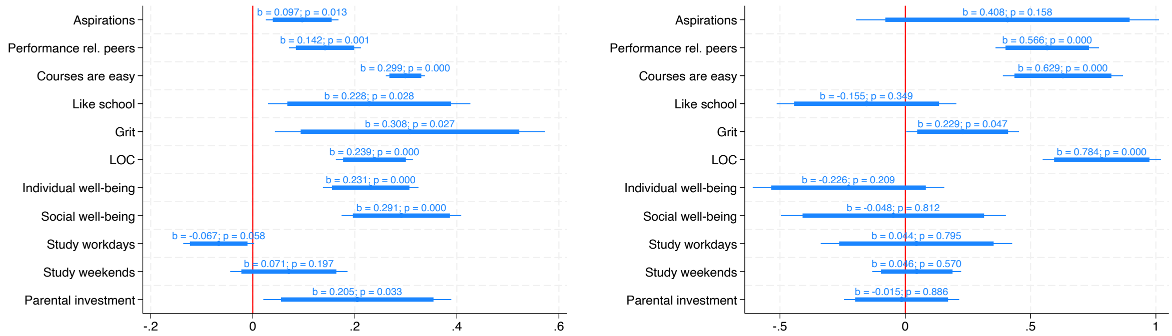

*Notes: The figure shows graphically the point estimate and the 90% (thin blue line) and 95% (thick blue line) confidence intervals for the treatment indicator in Equation 1. It also explicitly reports the point estimate and corresponding p-value for the reader's convenience. Sample size is 595.*

**Heterogeneity by risk of dyslexia.** As explained, Dytective also offers a dyslexia screening test. We split the sample according between those that we have identified as at-risk of dyslexia and those that we have not, and re-run the main analysis.

Figure 2 reports the results. Importantly, we find that the program is largely effective for both at-risk and not-at-risk students. Still, the impact on aspirations, perceived performance relative to peers, and locus-of-control is particularly large for the at-risk students — albeit the small number of at-risk individuals renders these comparisons somewhat imprecise. The impact on well-being is instead concentrated on the not-at-risk students.

Figure 2: Treatment effects by "at-risk status"

(a) Not-at-risk (b) At-risk




*Notes: The figure reports results from the same regressions as in Figure 1 separately for the subsamples of individuals classified as (a) not-as-risk and (b) at-risk of dyslexia based on their performance in Dytective's dyslexia screening test conducted prior to the intervention.*

**Heterogeneity by gender.** Appendix Figure A.3 shows that, while the non-cognitive skills of both males and females are positively affected, females' improvements in aspirations, taste for school and well-being seem to be larger.

**Impact on academic reading performance.** We now move on to explore whether the improvements in non-cognitive skills and perceptions documented above had tangible implications on the reading ability of the students. Table 2 shows that this is the case. Columns (2) and (3) report that participants in the program perform better at interpreting information and reflecting on the text that they read. This translates into a better overall performance. In particular, column (4) indicates that treated students score 3.45 points more in the test, for which the baseline mean score and standard deviation of the control group were 59.42 and 18.88, respectively. In columns (5) and (6), one appreciates that the treatment was beneficial both for at-risk and not-at-risk individuals, and that the effect seems to be larger for the at-risk subsample.

## 4.2 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 find returns to our program along both the cognitive and the non-cognitive margins, so both aspects should be considered when analyzing the cost effectiveness of the intervention.

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 3.45/18.88

Table 2: Impact on reading performance

| | (1) Locate Info. | (2) Interpret Info. | (3) Reflect | (4) Average | (5) Average | (6) Average |
|---|---|---|---|---|---|---|
| Treated | -2.430 | 3.452*** | 7.542*** | 3.454*** | 4.389** | 7.930** |
| | (2.780) | (0.810) | (1.336) | (0.600) | (1.577) | (3.353) |
| Baseline score | 0.475*** | 0.687*** | 0.287** | 0.698*** | 0.690*** | 0.864*** |
| | (0.041) | (0.059) | (0.108) | (0.062) | (0.062) | (0.111) |
| Observations | 419 | 419 | 419 | 419 | 317 | 54 |
| R-squared | 0.302 | 0.450 | 0.209 | 0.474 | 0.475 | 0.604 |
| Sample | Full | Full | Full | Full | Not-at-risk | At-risk |

*Notes: Regressions replicate those of Figure 1 for DIA reading performance. The mean (s.d.) of the outcomes in columns (1)–(4) are: 69.29 (26.48), 66.71 (22.28), 52.35 (31.04), and 59.42 (18.88). Sample size decreases because, relative to the sample selection for the main results, we additionally require that both the baseline and endline performance in the standardized reading tests are available. Columns (5) and (6) proceed separately for the not-at-risk and at-risk subsamples (sample size decreases because for 419-371 = 48 individuals we do not have the dyslexia risk measure and therefore cannot assign them to either subsample). "Locate information" refers to the ability of the student to navigate and extract information out of texts. "Interpreting information" refers to the ability of processing information to give a meaning to the text. "Reflect" refers to the ability to connect what has been read in the text with ideas that are external to the text to create critical thinking. "Average" refers to the (simple) average score across the three previous dimensions. All outcomes are on 0–100 scales. Baseline score refers to the value of the outcome at baseline (as expected, there is strong persistence in reading ability). Standard errors clustered at the school level in parentheses. p<0.01, ** p<0.05, * p<0.1*

≈ 18% of a 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 around 20% 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 psychopedagogues 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.

### 4.3 Robustness

**Attrition.** An important concern in longitudinal studies is selective attrition. Column (1) in Appendix Table B.4 reports that students from treated schools are 7.4 percentage points less likely to attrite than control students. This is not surprising because, as stated, our implementing partner was able to visit the treated schools multiple times per survey round, while it was granted access only once to control schools. We argue that this is unlikely to have a significant impact on our results for three reasons.

*First*, exploiting information on the date of survey completion, we construct an alternative measure of attrition that only considers the students that did the survey on the same day as the majority of their classmates. By only allowing one day of survey completion per class, we expect attrition rates to become fully comparable between treatment and control schools as long as the treatment had no impact on class attendance. Column (2) in Table B.4 shows that the likelihood of remaining in the sample under this alternative definition is indeed the same in the treatment and control groups. *Second*, following Muralidharan et al. (2019), we re-estimate our main specifications weighting each observation by the inverse of its probability of remaining in the sample at endline. Panel (a) in Figure 3 shows that the results remain largely unchanged. *Third*, in Appendix Figure A.5, we replicate the balance checks in Table 1 where the independent variable of interest, rather than being our treatment indicator, is an indicator taking the value of 1 if the person remains in the endline sample and 0 otherwise. This allows us to test for systematic differences at baseline between those who attrite and those who do not.

Consistently with the first argument, we find that the sole difference is that attritors are more likely to belong to the treated schools.

**Additional robustness checks.** We conduct three sets of additional robustness checks to probe the magnitude and statistical significance of our results. First, Figure 3's panel (b) shows that our estimated coefficients are largely unchanged after the joint inclusion as controls of the baseline measures of all the eleven families of primary outcomes. This strengthens our confidence that, within match-pairs, the treatment is plausibly as-good-as random. In a similar vein, Appendix Figure A.6 demonstrates the stability of the findings to the exclusion of the vector $X$ of individual controls.

Second, though we already attempt to reduce the number of outcomes that we explore through the construction of indices, one may still worry that some of the effects that we uncover are simply arising as a consequence of the large number of tests performed. We follow recent advancements in the literature on multiple hypothesis corrections by Viviano et al. (2023) who, in the case of multiple outcomes and a single treatment, propose to aggregate all outcomes to a single index and conduct inference on that single outcome. We obtain statistical weights to construct the index using principal component analysis. This procedure yields a first principal component that explains 24% of the total variation. Appendix Table B.5 reports the weights of each of the eleven variables employed. Appendix Table B.6 shows that, while our overall measure did not statistically differ at baseline between treatment and control groups, at endline the difference is of 31% of a standard deviation and it is significant at the 1% confidence level.

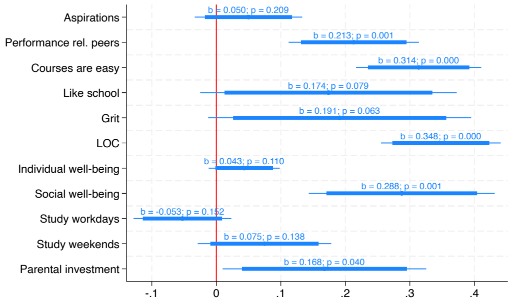
Third, Appendix Figure A.7 shows that the statistical significance of the main results is preserved when we cluster at the match-pair level. Employing wild-bootstrapped standard errors to account for the small number of clusters renders some of the coefficients statistically insignificant but the overall picture is preserved.

# 5 Conclusion

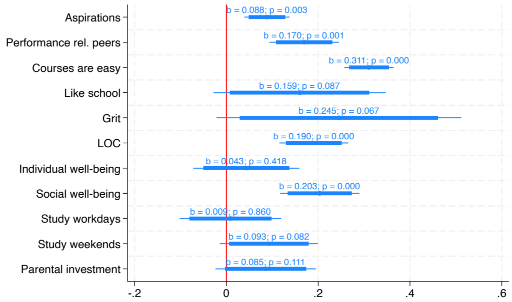
Closing educational gaps is still a major challenge for policymakers around the world. We evaluate a reading intervention that relies on computer-generated adaptive exercises that not only target the whole distribution of initial reading abilities (which tends to have a large support even among students 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.

Figure 3: Robustness of main results

(a) Inverse probability weighting




(b) Additional controls




*Notes: Robustness checks of the results in Figure 1. Panel (a) weights each observation by the inverse of the predicted probability of remaining in the sample at endline (based on a probit model using as predictors the maintained controls and the indices of academic aspirations and of enjoying academics). Panel (b) re-estimates the main specification additionally controlling for the baseline measures of all the indices explored as main outcomes.*

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 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 particularly 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. (return to page 4)

Alan, Sule and Ipek Mumcu (2022) *Nurturing Childhood Curiosity to Enhance Learning: Evidence from a Randomized Pedagogical Intervention*: Centre for Economic Policy Research. (return to page 4)

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). (return to page 17)

Almlund, Mathilde, Angela Lee Duckworth, James Heckman, and Tim Kautz (2011) "Personality psychology and economics," in *Handbook of the Economics of Education*, 4, 1–181: Elsevier. (return to page 3)

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. (return to page 9)

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. (return to page 4)

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 3)

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. (return to page 3)

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. (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. (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. (return to page 6)

Carlana, Michela and Eliana La Ferrara (2021) “Apart but connected: Online tutoring and student outcomes during the COVID-19 pandemic.” (return to page 3, 11, 16)

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 (3). (return to page 3)

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). (return to page 1)

Cunha, Flavio, James J Heckman, and Susanne M Schennach (2010) “Estimating the technology of cognitive and noncognitive skill formation,” *Econometrica*, 78 (3), 883–931. (return to page 3)

Dietrichson, Jens, Trine Filges, Julie K Seerup, Rasmus H Klokker, Bjørn CA 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. (return to page 16)

Duflo, Esther, Rachel Glennerster, and Michael Kremer (2007) “Chapter 61 Using Randomization in Development Economics Research: A Toolkit,” 4 of Handbook of Development Economics, 3895–3962: Elsevier. (return to page 7)

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. (return to page 1, 3)

Galuschka, K, R Görgen, J Kalmar, S Haberstroh, X Schmalz, and G Schulte-Körne (2020) “Effectiveness of spelling treatment approaches for learners with dyslexia–A meta-analysis and systematic review,” *Educational Psychologist*, 55, 1–20. (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. (return to page 4)

Glewwe, Paul and Karthik Muralidharan (2016) “Improving education outcomes in developing countries: Evidence, knowledge gaps, and policy implications,” in *Handbook of the Economics of Education*, 5, 653–743: Elsevier. (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. (return to page 3, 16)

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." (return to page 2)

Muralidharan, Karthik, Abhijeet Singh, and Alejandro J Ganimian (2019) "Disrupting education? Experimental evidence on technology-aided instruction in India," *American Economic Review*, 109 (4), 1426–1460. (return to page 3, 11, 17)

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. (return to page 5)

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*, 319–320. (return to page 5)

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), 000–000. (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 3)

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

Singer, Elly (2008) "Coping with academic failure, a study of Dutch children with dyslexia," *Dyslexia*, 14 (4), 314–333. (return to page 1)

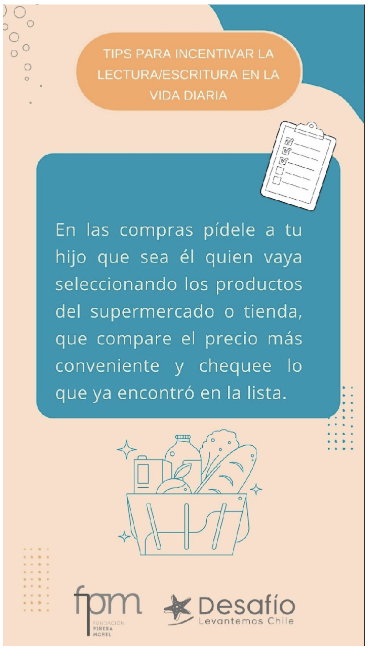
Singh, Abhijeet (2020) "Learning more with every year: School year productivity and international learning divergence," *Journal of the European Economic Association*, 18 (4), 1770–1813. (return to page 1)

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 mindsets important to academic achievement? Two meta-analyses," *Psychological science*, 29 (4), 549–571. (return to page 4)

Viviano, Davide, Kaspar Wuthrich, and Paul Niehaus (2023) "(When) should you adjust inferences for multiple hypothesis testing?" *arXiv preprint arXiv:2104.13367*. (return to page 18)

# A Online appendix: Figures

Figure A.1: Sample text message sent to parents



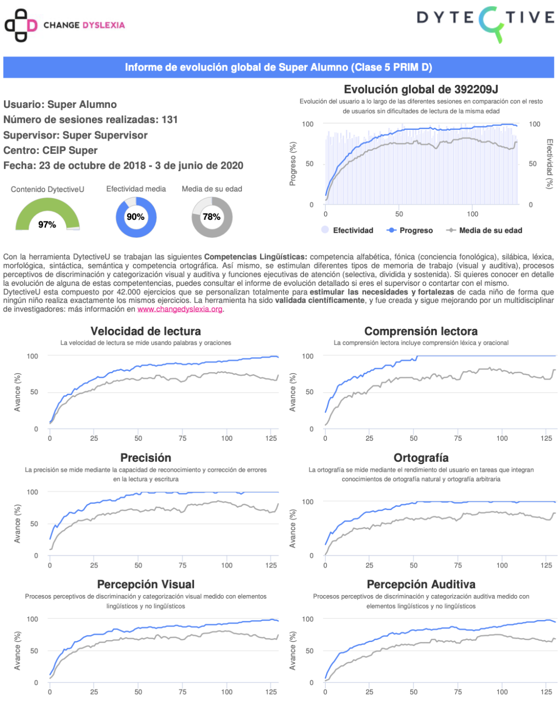

*Notes: Sample text message shared with parents in WhatsApp groups. The English translation is: "When shopping, ask your child to be the one that picks the products from the supermarket or store, to compare prices and ask him to check which items of the shopping list have already been added to the cart."*

Figure A.2: Contextual details

(a) Game interface


(b) Exercise interface


(c) Performance card

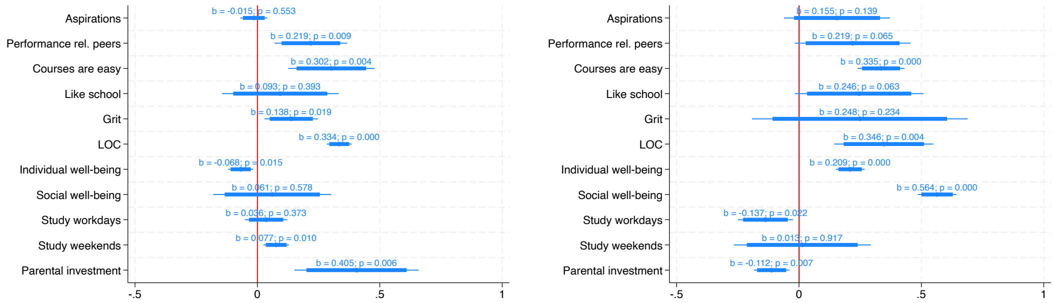



(d) In-class session

*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 psychopedagogues 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.*

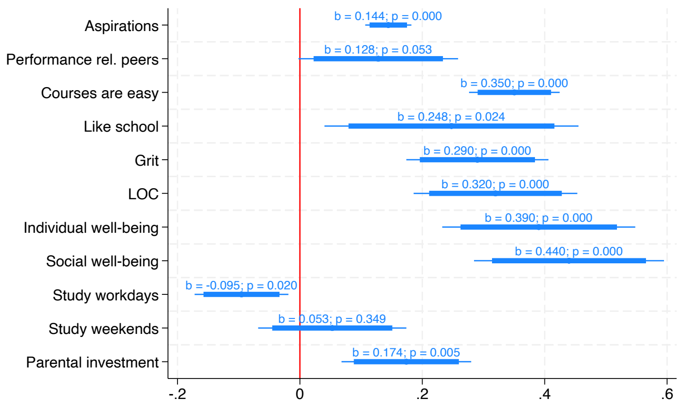
Figure A.3: Treatment effects by gender

(a) Males (b) Females



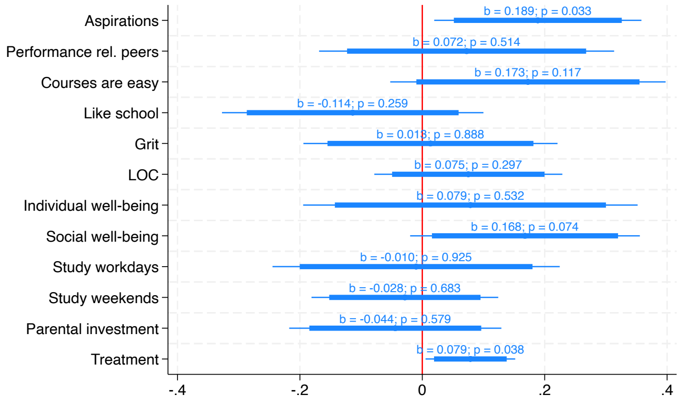

*Notes: The figure reports results from running the same regressions as in Figure 1 separately by gender.*

Figure A.4: Treatment effects on the sample available on the main survey day



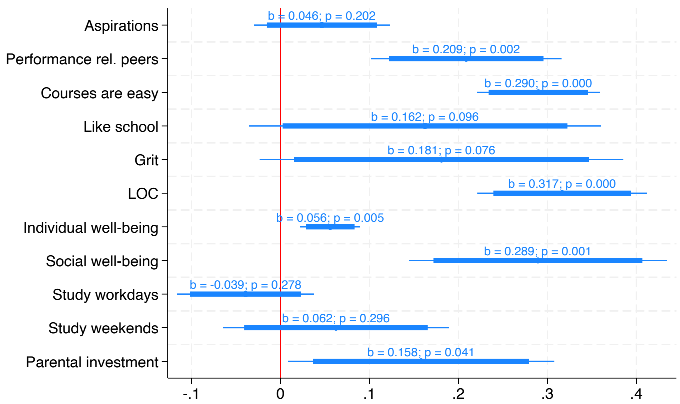

*Notes: Replication of Figure 1 where the estimating sample is restricted to those individuals who completed the survey on the main data collection day for each school. This aims at further homogenizing the estimating samples for treatment and control schools. Sample size is 377.*

Figure A.5: Non-selective attrition




*Notes: Replication of the comparison of means in Table 1 where the indicator for treatment status is replaced by an indicator taking the value 1 if the respondent is observed also at endline, and 0 otherwise. The graph therefore shows whether those individuals that attrite differ at baseline from those that do not attrite along the twelve dimensions explored (the last one being an indicator for whether the respondent belongs to a treated school). Both 90% and 95% confidence intervals are reported.*

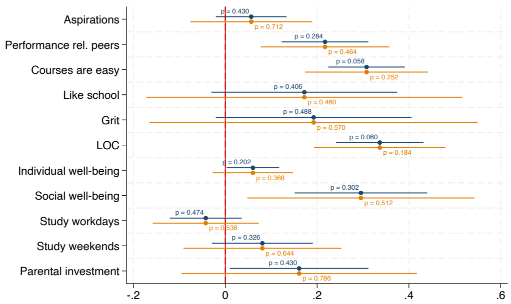
Figure A.6: Robustness – Exclusion of vector of individual characteristics



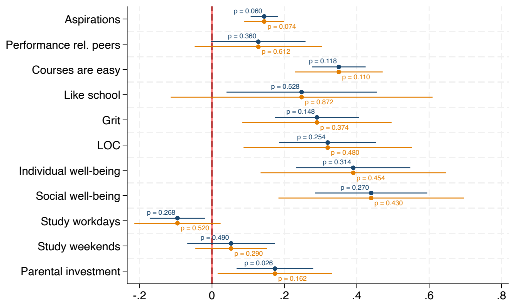

*Notes: Replication of Figure 1 where the individual characteristics in vector X have been removed.*

Figure A.7: Robustness to alternative clusterings

(a) Main sample




(b) Sample from main interview date




*Notes: Robustness check of the results in Figures 1 and A.4 to alternative choices of clustering of standard errors. 95% confidence intervals when clustering at the school level are reported in blue. 95% confidence intervals when clustering at the strata level are reported in orange. The reported p-values are the ones obtained after implementing wild-bootstrapping. Panel (a) uses the same sample as in Figure 1 while panel (b) uses the same sample as in Figure A.4.*

# B Online appendix: Tables

Table B.1: Summary statistics

| Variable | Mean | Std. Dev. | Min. | Max. | N |
|---|---|---|---|---|---|
| Male | 0.449 | 0.498 | 0 | 1 | 595 |
| Repeater | 0.087 | 0.283 | 0 | 1 | 595 |
| Screening score | 0.197 | 0.072 | 0.067 | 0.494 | 531 |
| **Index: aspirations** | 0.067 | 0.964 | -1.607 | 0.621 | 595 |
| Aspires to university (dummy) | 0.751 | 0.433 | 0 | 1 | 595 |
| **Index: perceived performance relative to peers** | 0.018 | 1.004 | -2.931 | 1.424 | 595 |
| Math relative to peers | 3.612 | 1.21 | 1 | 5 | 595 |
| Spanish relative to peers | 3.716 | 1.188 | 1 | 5 | 595 |
| Reading relative to peers | 3.824 | 1.246 | 1 | 5 | 595 |
| **Index: finds courses easy** | 0.103 | 0.924 | -2.843 | 1.256 | 595 |
| Math is easy | 3.563 | 1.33 | 1 | 5 | 595 |
| Spanish is easy | 3.997 | 1.183 | 1 | 5 | 595 |
| Reading is easy | 4.145 | 1.25 | 1 | 5 | 595 |
| **Index: likes school courses** | 0.052 | 0.957 | -3.487 | 1.19 | 595 |
| Likes school | 4.126 | 1.156 | 1 | 5 | 595 |
| Likes math | 4.032 | 1.276 | 1 | 5 | 595 |
| Likes Spanish | 3.918 | 1.181 | 1 | 5 | 595 |
| Likes reading | 4.04 | 1.2 | 1 | 5 | 595 |
| **Index: grit** | 0.077 | 1.017 | -2.816 | 2.045 | 595 |
| Likes hard tasks | 3.271 | 1.434 | 1 | 5 | 595 |
| Easily gives up (reversed) | 3.18 | 1.52 | 1 | 5 | 595 |
| Gives up if losing (reversed) | 3.756 | 1.5 | 1 | 5 | 595 |
| Wasted effort if not known (reversed) | 3.329 | 1.609 | 1 | 5 | 595 |
| **Index: locus-of-control** | 0.051 | 0.982 | -2.662 | 1.613 | 595 |
| Can improve if try | 4.052 | 1.228 | 1 | 5 | 595 |
| Luck matters in exams (reversed) | 3.378 | 1.432 | 1 | 5 | 595 |
| Can reach goals | 3.945 | 1.293 | 1 | 5 | 595 |
| Makes plans | 4.027 | 1.324 | 1 | 5 | 595 |
| Thinks of future | 3.886 | 1.323 | 1 | 5 | 595 |
| **Index: individual well-being** | 0.038 | 1.018 | -3.322 | 2.086 | 595 |
| Feeling happy | 4.244 | 1.091 | 1 | 5 | 595 |
| Many things worry me (reversed) | 3.037 | 1.431 | 1 | 5 | 595 |
| Feeling sad (reversed) | 3.872 | 1.337 | 1 | 5 | 595 |
| Easy to get mad (reversed) | 3.301 | 1.487 | 1 | 5 | 595 |
| Feel like doing nothing (reversed) | 3.079 | 1.429 | 1 | 5 | 595 |
| I do badly (reversed) | 2.879 | 1.398 | 1 | 5 | 595 |
| Hard to focus (reversed) | 3.007 | 1.421 | 1 | 5 | 595 |
| **Index: social well-being** | 0.108 | 0.964 | -2.76 | 1.445 | 595 |
| Feeling alone (reversed) | 3.603 | 1.478 | 1 | 5 | 595 |
| Classmates respect me | 3.56 | 1.321 | 1 | 5 | 595 |
| Feel safe at school | 4.062 | 1.31 | 1 | 5 | 595 |
| **Index: study workdays** | 0.067 | 0.995 | -1.124 | 1.628 | 595 |
| Hours study weekday | 2.731 | 1.445 | 1 | 5 | 595 |
| **Index: study weekends** | 0.026 | 0.982 | -1.067 | 1.747 | 595 |
| Hours study weekend | 2.553 | 1.396 | 1 | 5 | 595 |
| **Index: parental investment** | -0.048 | 0.972 | -3.403 | 0.997 | 595 |
| Parents help with homework | 3.782 | 1.213 | 1 | 5 | 595 |
| Parents care about school | 4.442 | 0.869 | 1 | 5 | 595 |
| **Reading test score** | 56.998 | 20.52 | 0 | 100 | 419 |
| Locate information | 65.298 | 25.651 | 0 | 100 | 419 |
| Interpret information | 61.246 | 21.09 | 0 | 100 | 419 |
| Reflect | 44.451 | 33.182 | 0 | 100 | 419 |

*Notes: Summary statistics at baseline of main predetermined variables, indices, and their individual elements. The sample includes those individuals used in our main estimations (i.e., those that are present both at baseline and at endline). Indices were constructed to have a mean of 0 and a standard deviation of 1 for the control group. 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 | 0.461 | 0.499 | 0 | 1 | 723 |
| Repeater | 0.094 | 0.292 | 0 | 1 | 723 |
| Screening score | 0.199 | 0.075 | 0.067 | 0.525 | 642 |
| **Index: aspirations** | 0.035 | 0.981 | -1.607 | 0.621 | 723 |
| Aspires to university (dummy) | 0.737 | 0.44 | 0 | 1 | 723 |
| **Index: perceived performance relative to peers** | 0.004 | 1.015 | -2.931 | 1.424 | 723 |
| Math relative to peers | 3.609 | 1.207 | 1 | 5 | 723 |
| Spanish relative to peers | 3.701 | 1.18 | 1 | 5 | 723 |
| Reading relative to peers | 3.801 | 1.263 | 1 | 5 | 723 |
| **Index: finds courses easy** | 0.07 | 0.961 | -2.843 | 1.256 | 723 |
| Math is easy | 3.548 | 1.344 | 1 | 5 | 723 |
| Spanish is easy | 3.957 | 1.223 | 1 | 5 | 723 |
| Reading is easy | 4.097 | 1.278 | 1 | 5 | 723 |
| **Index: like school courses** | 0.069 | 0.958 | -3.487 | 1.19 | 723 |
| Likes school | 4.144 | 1.149 | 1 | 5 | 723 |
| Likes math | 4.053 | 1.268 | 1 | 5 | 723 |
| Likes Spanish | 3.927 | 1.188 | 1 | 5 | 723 |
| Likes reading | 4.053 | 1.186 | 1 | 5 | 723 |
| **Index: grit** | 0.073 | 1.032 | -2.816 | 2.045 | 723 |
| Likes hard tasks | 3.266 | 1.435 | 1 | 5 | 723 |
| Easily gives up (reversed) | 3.198 | 1.522 | 1 | 5 | 723 |
| Gives up if losing (reversed) | 3.733 | 1.507 | 1 | 5 | 723 |
| Wasted effort if not known (reversed) | 3.331 | 1.615 | 1 | 5 | 723 |
| **Index: locus-of-control** | 0.035 | 1.007 | -3.625 | 1.613 | 723 |
| Can improve if try | 4.036 | 1.247 | 1 | 5 | 723 |
| Luck matters in exams (reversed) | 3.371 | 1.436 | 1 | 5 | 723 |
| Can reach goals | 3.928 | 1.298 | 1 | 5 | 723 |
| Makes plans | 3.996 | 1.341 | 1 | 5 | 723 |
| Thinks of future | 3.896 | 1.308 | 1 | 5 | 723 |
| **Index: individual well-being** | 0.022 | 1.017 | -3.322 | 2.086 | 723 |
| Feeling happy | 4.241 | 1.095 | 1 | 5 | 723 |
| Many things worry me (reversed) | 3.006 | 1.424 | 1 | 5 | 723 |
| Feeling sad (reversed) | 3.842 | 1.343 | 1 | 5 | 723 |
| Easy to get mad (reversed) | 3.31 | 1.49 | 1 | 5 | 723 |
| File like doing nothing (reversed) | 3.079 | 1.444 | 1 | 5 | 723 |
| I do badly (reversed) | 2.871 | 1.398 | 1 | 5 | 723 |
| Hard to focus (reversed) | 2.981 | 1.426 | 1 | 5 | 723 |
| **Index: social well-being** | 0.076 | 0.978 | -2.76 | 1.445 | 723 |
| Feeling alone (reversed) | 3.577 | 1.484 | 1 | 5 | 723 |
| Classmates respect me | 3.537 | 1.334 | 1 | 5 | 723 |
| Feel safe at school | 4.019 | 1.333 | 1 | 5 | 723 |
| **Index: study workdays** | 0.064 | 0.99 | -1.124 | 1.628 | 723 |
| Hours study weekday | 2.728 | 1.438 | 1 | 5 | 723 |
| **Index: study weekends** | 0.029 | 0.985 | -1.067 | 1.747 | 723 |
| Hours study weekend | 2.557 | 1.401 | 1 | 5 | 723 |
| **Index: parental investment** | -0.042 | 1.006 | -3.942 | 0.997 | 723 |
| Parents help with homework | 3.802 | 1.206 | 1 | 5 | 723 |
| Parents care about school | 4.436 | 0.902 | 1 | 5 | 723 |
| **Reading test score** | 55.6 | 20.93 | 0 | 100 | 557 |
| Locate information | 63.986 | 25.711 | 0 | 100 | 557 |
| Interpret information | 59.86 | 21.698 | 0 | 100 | 557 |
| Reflect | 42.953 | 32.965 | 0 | 100 | 557 |

*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

| Variable | (1) **Control** N | Mean/(SE) | (2) **Treatment** N | Mean/(SE) | (1)-(2) **Pairwise t-test** N | Beta/(SE) |
|---|---|---|---|---|---|---|
| Male | 366 | 0.481<br>(0.102) | 357 | 0.440<br>(0.102) | 723 | 0.016<br>(0.016) |
| Repeater | 366 | 0.096<br>(0.017) | 357 | 0.092<br>(0.019) | 723 | 0.008<br>(0.024) |
| Screening score | 328 | 0.193<br>(0.007) | 314 | 0.204<br>(0.004) | 642 | -0.009<br>(0.006) |
| Index: aspirations | 366 | -0.000<br>(0.103) | 357 | 0.072<br>(0.061) | 723 | -0.137*<br>(0.063) |
| Index: perceived performance relative to peers | 366 | -0.000<br>(0.098) | 357 | 0.009<br>(0.070) | 723 | 0.058<br>(0.069) |
| Index: finds courses easy | 366 | -0.000<br>(0.103) | 357 | 0.141<br>(0.038) | 723 | -0.086<br>(0.063) |
| Index: like school courses | 366 | 0.000<br>(0.150) | 357 | 0.140<br>(0.085) | 723 | -0.091<br>(0.168) |
| Index: grit | 366 | 0.000<br>(0.144) | 357 | 0.148<br>(0.070) | 723 | -0.189<br>(0.118) |
| Index: locus of control | 366 | 0.000<br>(0.118) | 357 | 0.071<br>(0.075) | 723 | -0.074<br>(0.068) |
| Index: individual well-being | 366 | 0.000<br>(0.026) | 357 | 0.046<br>(0.103) | 723 | -0.053<br>(0.056) |
| Index: social well-being | 366 | 0.000<br>(0.122) | 357 | 0.154<br>(0.107) | 723 | -0.265***<br>(0.075) |
| Index: study workdays | 366 | -0.000<br>(0.068) | 357 | 0.130<br>(0.066) | 723 | 0.012<br>(0.093) |
| Index: study weekends | 366 | -0.000<br>(0.069) | 357 | 0.058<br>(0.067) | 723 | 0.040<br>(0.104) |
| Index: parental investment | 366 | 0.000<br>(0.051) | 357 | -0.084<br>(0.074) | 723 | 0.101<br>(0.065) |
| Reading test score | 242 | 58.492<br>(2.965) | 315 | 53.378<br>(3.784) | 557 | 2.252<br>(2.831) |

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

Table B.4: Attrition by treatment status

| | (1) Attrited | (2) Attrited based on main interview date |
|---|---|---|
| Treated | -0.074*** | -0.003 |
| | (0.018) | (0.025) |
| Observations | 723 | 526 |
| R-squared | 0.013 | 0.047 |

*Notes: Regression of an indicator taking the value of 1 if a respondent observed at baseline is not observed at endline (and 0 otherwise) on the treatment indicator and month of interview and strata fixed effects. The second column defines the relevant sample as those individuals that responded the baseline survey on the main data collection date for each school, and the outcome variable takes the value of 1 if the endline was not completed on the main data collection date for each school (or was not completed at all). Out of the 723 individuals in column (1), 128 attrite. Out of the 526 individuals in column (2), 149 attrite. Standard errors clustered at the school level in parentheses. p<0.01, ** p<0.05, * p<0.1*

Table B.5: Indices' weights for PCA

| Index | Weight |
|---|---|
| Aspirations | 0.164 |
| Performance rel. peers | 0.352 |
| Like school | 0.420 |
| Courses are easy | 0.329 |
| Grit | 0.360 |
| LOC | 0.381 |
| Individual well-being | 0.355 |
| Social well-being | 0.324 |
| Parental investment | 0.214 |
| Study workdays | 0.058 |
| Study weekends | 0.091 |

*Notes: The table reports the weights generated for the first principal component when combining all main indices. The resulting variable is then employed as a robustness check to multiple hypothesis testing.*

Table B.6: Robustness to multiple hypothesis testing

| | (1) Baseline | (2) Endline |
|---|---|---|
| Treated | 0.119 | 0.306*** |
| | (0.118) | (0.049) |
| Observations | 595 | 595 |
| R-squared | 0.055 | 0.382 |

*Notes: The outcome in column (1) is the z-score of the variable created based on the first principal component out of all the indices used in the main analysis (see Table B.5 for the statistical weights employed). The outcome in column (2) is its endline counterpart, where the weights of each variable from the baseline have also been applied to yield the endline measure. The regression in column (1) replicates the ones in Table 1 while the regression in column (2) replicates those in Figure 1. Standard errors clustered at the school level in parentheses. p<0.01, ** p<0.05, * p<0.1*

# 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 in order 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.