Lopes, João et al.

Working Paper

From A to Z: Effects of a 2nd-grade reading intervention program for struggling readers

GLO Discussion Paper, No. 1394

Provided in Cooperation with:
Global Labor Organization (GLO)

Suggested Citation: Lopes, João et al. (2024) : From A to Z: Effects of a 2nd-grade reading intervention program for struggling readers, GLO Discussion Paper, No. 1394, Global Labor Organization (GLO), Essen

This Version is available at:
https://hdl.handle.net/10419/282317

From A to Z:
Effects of a 2nd-grade reading intervention program for struggling readers

João Lopes, *University of Minho*
Pedro S. Martins, *Nova School of Business and Economics & GLO* (Corresponding author)
Célia Oliveira, *Lusófona University - HEI-Lab: Digital Human-Environment Interaction Lab*
João Ferreira, *Nova School of Business and Economics*
João Tiago Oliveira, *CIPsi - Psychology Research Center, School of Psychology, University of Minho*
Nuno Crato, *CEMAPRE, ISEG, University of Lisbon*

Abstract

Many children in primary grades show difficulties with reading fluency, hardly reading text or doing it effortfully and fruitlessly, making intervention programs for struggling readers a priority for researchers and schools. This paper analyzes the results of a reading intervention program for 182 second grade struggling readers from public schools. Students received a multi-component program, including repeated readings, word recognition, morphological analysis, text interpretation, and writing skills. Participants received about fifty 45-minute intervention sessions over the school year. Using a difference-in-differences, quasi-experimental within-group longitudinal design (three-point measurements), we found that the intervention group progressed significantly faster than a classmate control group in all reading outcomes (speed, accuracy, and expressiveness). By the end of the school year, differences between the intervention and control groups in accuracy and expressiveness become small but are still large in reading speed.

Keywords: struggling readers; intervention program; reading speed; accuracy; expressiveness; difference-in-differences

JEL codes: I21, I26

Addressing early reading problems has become an increasing concern worldwide (e.g., Mullis & Martin, 2017; Park, 2019). Despite significant advances in recent decades (Rosling et al., 2018), UNICEF (2022) reports that "only a third of 10-year-olds globally are estimated to be able to read and understand a simple written story. The rest, around two-thirds (64%), cannot achieve this benchmark for minimum proficiency in reading comprehension. This is up from 52 percent pre-pandemic". According to UNICEF (October 2021) and other sources (Domingue et al., 2021; Psacharopoulos et al., 2021), the pandemic created even more severe challenges for students' learning, particularly in reading.

Portugal is no exception in the international scene, despite the country's positive results in international studies, such as the Progress in International Reading Literacy Study 2016 (PIRLS; Mullis & Martin, 2017). However, in the PIRLS ranked results scale, Portugal dropped from 19th to 30th among 61 countries between 2011 and 2016. The country's mean score (i.e., 528 points) was still significantly above the PIRLS scale's midpoint (i.e., 500 points). Fifty percent of Portuguese students achieved average scores equal to or greater than 530 points, and 5% scored 633 points or higher (95th percentile). About 5% of Portuguese students could not reach 500 points (5th percentile) (IAVE, 2017). These struggling students most likely carry their difficulties from their early primary grades (Castles et al., 2018; Fletcher et al., 2018).

Paige et al. (2019), in a study with 1064 US third graders, found that students who could spell and read fluently (i.e., with accuracy, expressiveness, and an adequate reading speed) had a 70% chance of getting good grades in their state reading assessment, against a 20% chance for struggling readers. This study, like many similar studies (e.g., Cadime et al., 2023; Cockerill et al., 2023; Risberg et al., 2023), suggests that failure to develop fluency in elementary students might compromise their school achievement and beyond (Rupley et al., 2020). This fact motivates a strong interest in identifying inexpensive, easily implementable, and scalable remedial interventions. A large literature with such focus has recently emerged, focusing for instance on low-tech interventions in the context of school closures (Angrist et al., 2022). Moreover, tutoring programs saw renewed interest in the context of the pandemic: online provision of individualized tutoring by volunteer college students was particularly promising due to its relatively low cost, although estimated effects on learning were often modest (Kraft et al., 2022). Carlana and La Ferrara (2021), evaluating a similar program in Italy, find more optimistic results across a range of academic, socio-emotional, and psychological outcomes. Indeed,

tutoring programs that do not rely on the hiring of teachers have recently received greater attention: Guryan et al. (2023) is one such example. The paper analyses an intervention employing paraprofessional tutors and finds strong effects on adolescents' academic outcomes. In a cost-benefit analysis, the authors find a ratio that is similar to those obtained in early childhood education programs.

Over the years, governments, schools, and researchers have developed strategies and programs to improve the reading rates of their populations. Some strategies operate at a macro level (i.e., governmental level, e.g., reducing the number of students in the classroom), others at a smaller level, including classroom and individual plans (Ecalle et al., 2019, 2022). Compulsory schooling is one of the most relevant macro-level strategies, especially since the 20th century. However, several authors (e.g., Filmer et al., 2020; Glawe & Wagner, 2022; Kaffenberger & Pritchett, 2017) and international instances (World Bank, 2018a, 2018b) have recently raised concerns about the limits of schooling in improving reading worldwide. For example, Angrist et al. (2021) showed that countries from the Middle East and sub-Saharan Africa increased student enrolment rates from 95% to 99% from 2000 to 2010 but made no progress in PISA or TIMMS. Also, countries with consolidated high schooling (e.g., Gana, Brazil) show persistent low learning levels. Angrist et al. (2021) conclude that while the world is on track to achieve universal primary enrolment by 2030, accomplishing the United Nations Sustainable Development Goals, this will be of little use if learning stagnates.

Mather et al. (2020) provide an interesting picture of what is happening worldwide with services and opportunities for students with reading problems. The authors found that the culture, language orthography, teacher training, and availability of assessments and interventions influence practices intended to tackle reading problems. In turn, Motiejunaite and his colleagues (2014) reviewed the national policies of 32 European countries to improve reading achievement and concluded that providing reading specialists is perhaps the most critical factor distinguishing education systems. The countries that offer targeted continued support seem to achieve higher in reading. However, the intervention of education authorities seems necessary so that reading specialists are actually provided. The authors also found that classroom continuous assessment methods, curriculum guidelines on reading, and national policies are relevant to dealing effectively with reading difficulties.

Unlike governments, researchers and schools typically approach struggling readers' difficulties through small-scale strategies or programs focused on students'

reading weaknesses (e.g., poor decoding, disfluent reading). Although these strategies might assume a myriad of formats, two aspects seem critical for service delivery to struggling readers (the focus of this paper): (1) the organization and structure and (2) the content and didactics (instruction) of the intervention. A review of evidence about those aspects follows.

Organization and Structure of Interventions

Several meta-analyses (MA) and systematic reviews (SR) of literature over the past 20 years have synthesized the organizational and structural conditions under which programs designed for struggling readers are likely to succeed (e.g., Denton et al., 2022; Hall, Dahl-Leonard, Cho, et al., 2022; Maki & Hammerschmidt-Snidarich, 2022). According to See and Gorard (2020), one of these conditions is implementation fidelity, which is crucial for any intervention. The authors suggest that some programs fail because schools attempt to modify the program by altering its intensity and structure (small groups instead of individual support) or applying it to students of different ages.

The tutor-student ratio and the timing and location of the intervention are other factors that might influence the success of educational programs for struggling readers. Nickow et al. (2020) meta-analysis shows a 0.37 effect size of tutoring on learning outcomes, comparable to the 0.36 significant effect found by Dietrichson et al. (2017). Gersten et al. (2020) discovered an average effect size of 0.46 in one-to-one interventions and .31 for small groups of students from grades 1-3. Further analysis revealed that the effect was significant for grades 1-2 but not for grade 3. Gersten et al. (2020) and Al Otaiba et al. (2014) found larger effect sizes when small groups are homogeneous.

According to several MA and SRs (e.g., Nickow et al., 2020), there are various formats for intervention locations, including schools, homes, and educational centers, and the results for differential effectiveness are inconclusive. However, Nickow et al. (2020) found some advantages to interventions during school time. MA and SR also reveal that intervention programs vary considerably in frequency and length and the total number of sessions (e.g., from several weeks to one or two years of school).

The meta-analyses of Wanzek et al. (2016) found no effect of the number of intervention hours in Tier 2 intervention for at-risk students in grades K-3 in reading foundational skills. In the studies reviewed by Nickow et al. (2020), longer interventions produce better reading results. Additionally, more intervention days per week positively impact reading outcomes until it may become exhausting for students. Similarly, more

extended sessions (the authors report programs with 10-15 minutes to one hour, with most sessions ranging from 30-60 minutes) appear to yield larger effect sizes until students' attention cannot be maintained. Dietrichson et al. (2017) found that the duration and frequency of delivery have significant effects, but the intensity did not.

The type of tutor delivering remedial reading tutoring in early elementary grades seems to favor professional teachers over paraprofessionals and nonprofessionals. However, using paraprofessionals appears to be a promising strategy, with lower costs (Jones et al., 2021), but it requires extensive tutor training and specific program scripts (Samson et al., 2015). Wanzek et al. (2016) report studies in which researchers, general education teachers, special education teachers, reading specialists, and paraprofessionals implemented the interventions and found no significant differences. The authors suggest that less intensive Tier 2 interventions might succeed with various implementers.

Contents and Didactics

Programs for early struggling readers vary considerably in content and didactics. However, some meta-analyses (e.g., Gersten et al., 2020) report that most, if not all, programs for primary-grade struggling readers address multiple aspects of foundational reading skills (e.g., phonological awareness, decoding, encoding (spelling), reading fluency, and sometimes writing). Vocabulary and comprehension instruction is rarely approached. Research shows that writing interventions produce larger effect sizes, while those addressing phonological awareness are associated with significantly smaller effects (Gersten et al., 2020; Graham et al., 2013, 2018). The authors suggest that focusing on pre-reading skills when the student can already read might be counterproductive.

Morphological instruction is almost absent from typical classrooms and intervention programs for struggling readers (Moats, 2009). However, morphological knowledge contributes to literacy development through word recognition, comprehension, and motivation (McCutchen et al., 2022; Ng et al., 2022; Savage et al., 2023). Research has also shown that morphological knowledge is involved in word reading accuracy, predicts unique variance in vocabulary knowledge and spelling, and may represent a particular advantage for struggling readers (Simpson et al., 2020).

The didactics of the contents might be as crucial as or even more critical than the contents themselves in intervention programs for struggling readers. Generally, studies (e.g., Dietrichson et al., 2017; Nickow et al., 2020) report larger effect sizes for direct, explicit instruction, immediate feedback, and emphasis on specific reading/writing skills.

Repeated reading (RR) is a vital strategy for intervention programs focusing on reading fluency. Lee and Yoon (2017) demonstrated RR's positive effect on reading fluency and some significant moderators of this relation. For instance, reading the passage four times is more effective than reading three times and listening to the passage beforehand (which might reduce text reading resistance). The authors also found that RR is more effective for elementary grades and when a logical instruction sequence is adopted (e.g., from letter knowledge to fluency and vocabulary knowledge).

Some programs include reading to struggling readers to promote student literacy, following literature recommendations (e.g., Ece Demir-Lira et al., 2019; Niklas et al., 2016). However, some studies found no long-term benefits of such practices (e.g., Klein & Kogan, 2013), and others found no relation between the time first-grade teachers spent reading to their students and students' reading achievement (Meyer et al., 1994). Reading to struggling readers might be a positive strategy to address poor readers' negative feelings towards reading and promote reading foundational skills. As Meyer et al. (1994) put it, "Reading storybooks to children is not a reading program. It is part of a reading program. The direct benefits from storybook exposure can come if children develop print-related skills, such as phoneme awareness and some word recognition" (p. 83).

Lastly, reading interventions on a one-to-one or one-to-two basis favor the customization of learning, which allows for teaching at the right level (Banerjee et al., 2015). Teaching at the right level is one of the most challenging tasks for teachers because of the teacher-student ratio and the variance in students' knowledge. According to Nickow et al. (2020), "individualized reading interventions lead to a massive increase in customization—albeit usually for only a few hours each week—as a supplement to the lower customization classroom setting" (p. 7).

The *A to Z* Program

The literature generally indicates that the specificity of school-based intervention programs for early struggling readers lies in a unique combination of content and delivery format (Gersten et al., 2020). In our study, we tested the efficacy of an intervention program called "AaZ" ("A to Z" in English), designed to develop reading fluency (speed, accuracy, and expressiveness) for second grade struggling readers. The program's particular combination of content and delivery format is presented in the procedure section.

A to Z is an intervention program for first and second graders with reading difficulties. In this article, we will report results for second graders. The program is being conducted in several parts of the country: the nine Azores Islands, the Madeira Island, and three different regions of mainland Portugal. The program has a national coordinator, five regional coordinators, 46 tutor-teachers, and eight associated researchers. Second-grade participants come from 108 classrooms and 73 schools. The schools, the classrooms, and the parents of the students accepted to participate in the program.

At the beginning of the school year, students were selected to join the program according to criteria detailed in another section of this article. Once this process is completed, the intervention begins. The tutor-teachers who work directly with the students receive specific training before joining the program. They also benefited from individual monitoring during the school year, at least once a month, provided by the project's researchers, and daily support from region coordinators. Students included in the program are evaluated every three weeks, and their classmates (control group) are evaluated at the beginning, middle, and end of the school year.

Several significant features distinguish the A to Z Program from other programs conducted in Portugal: (1) struggling readers receive intervention from qualified teachers right from the beginning of the school academic year; (2) the program lasts a whole school year, allowing to control for reading setbacks, often found when intervention programs end (Van Norman et al., 2020, p. 510, for example, stress that "Once students meet exit criteria and Tier 2 reading supports are removed, many do not maintain their current progress and fail to meet future performance benchmarks."); (3) extensive writing, besides decoding training and text interpretation, is emphasized as a strategy to develop reading fluency; (4) teacher readings to students are used as a motivational tool for independent reading.

Most reading intervention studies in Portugal are short-term or small-group studies targeting specific features of reading learning (e.g., phonological awareness) (e.g., Carvalhais et al., 2020; Sucena et al., 2021, 2022). To our knowledge, only one study in the country (Cadime et al., 2022) comes close to the A to Z program. Still, there are significant differences. Cadime et al. (2022) conducted a single-group design study (no control group) with only two measurements in time (pre-test and post-test). Moreover, this program is shorter than the A to Z program (three-month duration), the contents only partially overlap, and the interventionists are educational psychologists (in the A to Z Program, interventionists are qualified teachers).

One important reason for the scarcity of extensive intervention reading programs is the cost of such programs (Slavin et al., 2011). Unlike other programs in the country, the A to Z program could gather the conditions to be extensive, conducted by qualified teachers, and supported by a research team that provided expertise in content and measurement issues. Moreover, unlike other programs, tutor-teachers receive individualized research team support.

Students were assigned to the program (a) if they got a median result ≥ 3.5 in each of the three subscales of the Hong-Kong Learning Difficulties Behaviour Checklist (HKBCL; Artiles & Jiménez, 2008) and (b) if they could read no more than thirty words per minute (wpm) in an oral reading test. We chose the threshold of 30 words per minute (wpm) for reading speed because it corresponds to students between the 10th and the 25th percentile of the norms of the "Oral Reading Fluency" curriculum-based measure (Hasbrouck & Tindal, 2017). We also considered national expected values for reading speed (Buescu et al., 2015). In Portugal, it is expected that an average second grader can read 55 wpm (SD = 10) at the beginning of the school year.

The combination of international (Hasbrouck & Tindal, 2017) and national norms (Buescu et al., 2015) is important because while national norms are adapted to the Portuguese population, Hasbrouck and Tindal norms detail percentiles and provide information for three moments throughout the year (the national norms give only one value for the end of the school year), which match the assessments of the A to Z program.

Possible concerns about using indicators from two languages do not arise since there is evidence of no significant differences in reading rates, independently of the language. For example, Liversedge et al. (2016), exploring the issue of universality in eye movements and reading in Mandarin Chinese, English, and Finnish, found that despite significant differences in the orthographies (e.g., different number of words for the same text, different number of saccades) participants' reading time and reading comprehension were similar.

Verhoeven and Perfetti (2022) investigated the variations between seventeen languages with varying writing systems (e.g., alphabetic, syllabic, morphosyllabic) and concluded the universality of learning to read in these languages. Also, Brysbaert (2019) asserts that the time spent reading information in distinct languages is similar regardless of the number of words used in each language and its length. The special issue of Scientific Studies of Writing (Vol. 26, 2022), titled "Processes of reading and its development across orthographies: state of the science") offers an interesting discussion

about this topic. It is important to stress that random assignment of students was impossible because many teachers, although identifying reading difficulties in some children with results < 30 wpm in the oral reading test, argued that children with more difficulties other than these should be assigned to the program. The difference-in-differences design of our study, however, is a tool that minimizes the effect of the non-randomization of participants (Rose & Bowen, 2019; Steinmann et al., 2023; Strello et al., 2021).

The Present Study

The main goal of our study is to test the effectiveness of the A to Z program using a difference-in-differences quasi-experimental inter- (intervention and control group) and intra-group longitudinal design (three-point measurements). Specifically, we assessed differences over time in *reading speed, accuracy, and expressiveness* between struggling readers involved in the program (the intervention group) (IG) and a non-experimental control group (CG), composed of classmates of the IG. Participants' *reading speed, accuracy, and expressiveness* were evaluated at the beginning (Moment 1 - before the intervention), middle (Moment 2), and end of the school year (Moment 3). According to the selection procedure, in Moment 1, differences between IG and CG in *reading speed, accuracy, and expressiveness* are highly significant.

The hypotheses of the study regarding the effects of the intervention program are as follows:

H1: By Moment 2, differences between IG and CG in *reading speed, accuracy, and expressiveness* are significantly smaller than in Moment 1.

H2: By Moment 3, differences between IG and CG in *reading speed, accuracy, and expressiveness* are significantly smaller than in Moment 2.

H3: The between-group difference reduction is more significant from Moment 1 to Moment 2 than from Moment 2 to Moment 3.

H4: From the beginning to the end of the intervention, the highest performers in the intervention group progress significantly more in reading speed than the lower performers.

Method

Participants

Participants were 1009 second-grade students (boys = 51.3%, aged 7-8, from public schools) from various country regions. One hundred eighty-two students were integrated into the intervention group (IG) (boys = 46.7%) of the program A to Z for struggling readers, and 827 were assigned to the control group (CG) (boys = 52.4%). The participants came from 108 classrooms from 73 schools willing to receive the program. At the beginning of the program, there were 203 students in the intervention group and 1043 in the control group. Attrition is, therefore, 10.34% and 20.71%, respectively. While there are no steadfast rules for attrition, these are acceptable levels. The attrition in the control group might raise concerns because it is around 20%. However, the number of participants is large enough to accommodate this level of attrition. No participants were excluded or eliminated for not obtaining parental consent to participate.

Measures

Hong-Kong Learning Difficulties Behaviour Checklist (HKBCL; Artiles & Jiménez, 2008)

The HKBCL is a Likert format questionnaire filled by the classroom teacher informing about the student's performance in reading, writing, mathematics, general behavior, oral proficiency, and memory. The Spanish version of the scale (Artiles & Jiménez, 2008) has 12 subscales. For our study, only the subscales "*Reading Comprehension*" (ten items), "*Spelling and Dictation Writing*" (five items), and "*Written Expression*" (seven items) were used. The subscales were translated into Portuguese and then retroverted into Spanish. The items are answered on a five-point scale, ranging from 0 to 4, in which 0 means "never" and 4 means "frequently." For each subscale, the maximum score equals the number of items times four. A higher punctuation indicates more learning difficulties. The HKBCL was administered at the beginning of the school year to select students for the intervention group. Details about the questionnaire's psychometric properties can be found elsewhere (i.e., Artilles & Jiménez, 2008, 2010).

A principal component analysis (PCA) was conducted on the 22 items of the questionnaire with oblique rotation (promax) for the sample of participants. The Kaiser–Meyer–Olkin measure verified the sampling adequacy for the analysis, KMO = .96 . Bartlett's test of sphericity χ^2 (231) = 14199.08, $p < .001$, indicated that correlations between items were sufficiently large for PCA. An initial analysis was run to obtain

eigenvalues for each component in the data. Three components had eigenvalues over Kaiser's criterion of 1 and explained 68.40% of the variance in combination. Given the large sample size and the convergence of the scree plot and Kaiser's criterion on three components, this is the number of components retained in the final analysis. The items that cluster on the same components suggest that component 1 (ten items) represents reading comprehension, component 2 (five items) represents spelling and dictation writing, and component 3 (seven items) represents written expression, just as predicted in the original scale. Table 1 presents values for the reliability and construct validity of the instrument in our sample.

Table 1

Reliability and Construct Validity Indicators

Components	*Cronbach's α*	*Composite reliability*	*McDonald's omega (ω)*	*Average variance extracted*
Reading comprehension	.94	.78	.94	.57
Spelling and dictation writing	.91	.84	.91	.68
Written expression	.91	.66	.91	.40

Teste de Avaliação da Fluência [Test of Fluency Evaluation] (TAF; Rodrigues et al., 2022).

The TAF is an oral reading fluency test for primary-grade students (grades 1 through 6). The "TAF test forms from grades 2 to 6 were horizontally equated using equipercentile equating to place test forms of each grade level on the same metric" and "…to ensure the comparability of the scores between grades, the chained equipercentile method with NEAT [non-equivalent groups with anchor test] design was used to vertical link the TAF test forms" (Rodrigues et al., 2022, p. 4). When multiple forms of assessments are used, equating is necessary to allow comparisons between and within students (Kara et al., 2023). For every grade (1 to 6), the reliability values, based on test-retest correlation analysis, are very high, exceeding .90. All coefficients are statistically significant at $p < .001$.

In Grade 2, students have to read aloud three different texts (1 minute for each text) in each evaluation moment: one narrative text with dialogue, one without dialogue, and one with expository text. The texts ranged from 236 to 275 words and differed in the three evaluation moments. The final result for each evaluation moment is the mean performance in the three texts. Three indicators (the outcomes of the study) were collected from the reading measures: *reading speed* (the number of words correctly read per minute), *reading accuracy* (number of words correctly read/number of words read), and *reading expressiveness* (measured in a five-point scale, ranging from "no expressiveness" to "highly expressive").

The tutor-teachers recorded the students' readings and sent the recordings to the research team. Each recording was evaluated independently by two researchers. The Audacity software (Audacity Team, 2000) automatically calculated reading time. The inter-rater agreement was around 100% for reading speed and accuracy and 95% for expressiveness. A third researcher was called upon whenever there was disagreement. We chose an oral test to measure students' reading rate because oral reading aloud is considered a good indicator of reading proficiency in the first school years. Its importance declines in later years (Brysbaert, 2019), although Rasinski et al. (2022) used oral reading measures to develop norms for adult proficient readers.

Most importantly, for adult readers, "There is little evidence that the differences in reading rate lead to better or worse text comprehension … If anything, fast readers tend to be slightly better than slow readers" (Brysbaert, 2019, p. 17). Fumagalli et al. (2019) stress that although oral reading is the most common way to measure reading fluency, it only relates to some reading comprehension processes. Still, Salceda et al. (2020) argue that the reading rate is related to reading comprehension in transparent orthographies such as Spanish (the Portuguese orthography is of Intermediate Depth (Cadime et al., 2023; Seymour et al., 2003)).

Procedure

The ethics council approved the study (CEICSH 083/2020) conducted at the University of Minho. The University follows the principles of the Declaration of Helsinki (World Medical Association, 2013), ensuring that participants and parents were provided with all necessary information about their involvement in the study. Informed consent documents were signed by parents, indicating their children's voluntary participation. The study also adhered to guidelines regarding anonymity, data confidentiality, potential

risks, and participants' access to their data. Although not all students that could benefit from the program could be included, indirect support was provided to their classroom teachers, and a spillover effect (that will be approached explicitly in another article) might have benefited those students and their classmates.

The involvement of students from the intervention group (IG) and the control group (CG) was conducted according to the following process: (1) at the beginning of the school year, teachers from the schools that volunteered for the intervention program indicated students with reading problems (filling the HKBCL instrument); (2) the program staff subsequently assessed the whole class in the TAF reading test (Rodrigues et al., 2022); (3) the students indicated by the teacher whose reading speed was below thirty words per minute, were assigned to the IG. A Cohen's k = .95 was obtained for raters' measurements. Since reading lessons with a tutor-teacher occurred during class time, a maximum of three students per class was selected so the IG would not disturb the lessons too much. Consequently, the CG could also integrate some students reading less than 30 wpm but who could nevertheless read text; (4) the CG was composed of classmates of IG students. Most read more than 30 wpm; (5) students who could not read text were excluded from the analysis since there was no possibility to measure any of the dependent variables (*reading speed, accuracy, and expressiveness*); (6) the IG and CG students were evaluated in three moments: at the beginning (September), middle (January), and final of the school year (May).

It is important to stress that random assignment of students was impossible because many teachers, although identifying reading difficulties in some children with results < 30 wpm in the oral reading test, argued that children with more difficulties other than these should be assigned to the program. The difference-in-differences design of our study, however, is a tool that minimizes the effect of the non-randomization of participants (Rose & Bowen, 2019; Steinmann et al., 2023; Strello et al., 2021).

After the students' selection procedure, 42 full-time and four part-time tutor-teachers initiated the intervention program. To warrant implementation fidelity, the teachers received a two-day training on the program's foundations and main guidelines. Each tutor-teacher had eight students divided into groups of two students. Each group received remedial reading out of the classroom thrice a week, in 45 to 60-minute lessons during class time. To control for implementation fidelity, the teachers filled out a checklist, signaling the phases of the session that have been accomplished. Moreover, each tutor had at least a monthly meeting with the university staff leading the program.

Every lesson had the same structure: (1) the tutor-teacher read aloud a passage of selected children's books (for about five minutes) followed by a short conversation about the text (to show the student that reading can be pleasant, to model reading, to put the student in contact with a new lexicon and to motivate the student for reading); (2) repeated readings of appropriated level text (a different text from 1); (3) fast identification of words in the text; (4) morphology study: production of family words of words found in the text (orally and in writing), the handwriting of synonyms and antonyms; (5) oral and written text interpretation (literal and inferential questions); (6) text summary (written form). The subsequent session always started with a summary of the preceding session, followed by (1). The intervention lasted for a whole school year.

Data Analysis

After performing descriptive statistics for each variable and comparing them between groups using t-student tests, we evaluated short-term program benefits using a difference-in-differences (DID) approach. This methodology originated in economics and econometrics, where it became prevalent. DID is a quasi-experimental design that uses pooled cross-section or longitudinal data from intervention and control groups to estimate a causal effect. It is a valuable tool when the randomization of participants is not feasible or when data is observational. The technique has been used in some large-scale studies in education (e.g., Kodila-Tedika & Otchia, 2022; Lavrijsen & Nicaise, 2016; Rosén & Gustafsson, 2016; Salinas & Solé-Ollé, 2018) but not yet in the evaluation of reading intervention programs.

DID allows us to compare the changes in reading over time of each student enrolled in our program ("A to Z" intervention program) with the same changes of each of their classmates (control group) not enrolled in the program. DID does not require the intervention and control groups to exhibit the same initial characteristics and outcomes. The method estimates the program's impact from the differences in the changes over time of the critical outcomes. Specifically, we compare the changes in reading outcomes of students supported by the program (intervention group) with those not supported by the program (classmates' control group). The latter difference will remove the natural growth in learning outcomes driven by teaching and time (aging). Moreover, we control for differences from the evolving reading tests deployed at subsequent periods. We can thus seek to isolate the specific contribution of the program in a much-improved way compared to simpler before-after or treatment-control comparisons (see Martins, 2017a,

2017b, or a similar illustration based on a different remedial program also implemented in Portugal). Ideally, we would have been able to test for common trends between the treated and control groups before the timing of the intervention. However, such information is not available as there is only one period of data before the intervention is conducted. The main specifications that we take to the data are of the form:

$$y_{it} = \beta_0 + \beta_1 Moment\_2_{it} + \beta_2 Moment\_3_{it} + \beta_3 Moment\_2_{it} * Treatment_i + \beta_4 Moment\_3_{it} * Treatment_i + \alpha_i + u_{it},$$

where y_{it} is the outcome of student i in Evaluation Moment t, $Moment\_t * Treatment_i$ are interaction terms whose associated coefficients (β_3 and β_4) are estimates of the program effects in terms of outcome variable y, at two different periods following the program's start (respectively, moments 2 and 3). α_i depicts student fixed effects, accounting for individual time-invariant (observed or unobserved) heterogeneity. The latter variable captures permanent differences across students (such as sex, race, ethnicity, or socio-economic background), ensuring the estimation is based on relative differences over time in each student's performance. The four different but complementary dependent variables considered in each specification that we take to the data are (1) words per minute (wpm), a measure of reading speed, and (2) its logarithm; (3) reading accuracy; and (4) reading expressiveness. All statistical analyses were performed using Stata (version 18.0; StataCorp, 2023).

Results

Descriptive Statistics

Table 2 presents descriptive statistics of our sample. We analyzed 182 students in the intervention group and 827 in the control group, all of whom could read at least five words per minute right from the beginning of the intervention. The intervention group starts significantly behind in all dimensions analyzed: *reading accuracy, reading speed, and reading expressiveness*. The differences in these three dimensions are large and statistically significant. However, when repeating these comparisons for the second and third evaluation moments, all differences between intervention and control groups become smaller. The analysis of the effect sizes (we used Hedges' g because the sample sizes are significantly different) shows that in moment 1, between-group differences are large in every reading aspect. However, in moment 3, differences become small for

accuracy and *expressiveness* and medium for *reading speed*. These data show a significantly positive effect of the program on struggling readers, but still not enough for them to perform on the average of their classmates, mainly in reading speed.

Table 2

Descriptive Statistics of Reading Features (2nd-grade students)

Variable	Intervention Group		Control Group		Difference		
	Mean	*SD*	*Mean*	*SD*	*Diff.*	*t*	*Hedges' g*
Moment 1 (September)							
Accuracy	80.79	12.87	90.73	11.07	-9.94***	-10.63	-.87
Speed (wpm)	20.37	9.15	47.07	26.32	-26.69***	-13.50	-1.10
Expressiveness	1.63	.74	2.77	1.30	-1.15***	-11.47	-.93
Moment 2 (January)							
Accuracy	91.01	6.69	93.87	5.79	-2.85***	-5.85	-.48
Speed	41.03	12.49	58.70	21.65	-17.67***	-10.62	-.87
Expressiveness	2.52	.73	3.02	.97	-.50***	-6.52	-.54
Moment 3 (May)							
Accuracy	96.06	4.17	96.97	3.22	-.91**	-3.26	-.27
Speed	63.58	17.61	78.42	25.13	-14.83***	-7.56	-.62
Expressiveness	3.26	.97	3.47	.93	-.19**	-2.37	-.19
N	182		827		**Total:** 1009		

Note. The table presents descriptive statistics of the intervention and control groups of students (who could read text by Evaluation Moment 1 *and* for whom there is available information for the three Evaluation Moments), as well as a statistical analysis of the differences in their means, without controlling for regional, school, teacher, or individual fixed effects. Values for Moments 2 and 3 are computed based on the average results obtained by each student regarding the reading of 3 new texts in each moment. *'Reading Accuracy'* is the percentage of words read correctly, and *'Reading Expressiveness*' is an ordinal variable that may take values 1-5 (1 being the lowest level). *Wpm: words per minute*. Significance levels: ** .05; *** .01.

A to Z Efficacy: Difference-in-differences estimates

Table 3 presents our main DID estimates from the equation above's separate ordinary least squares (OLS) regressions. The results indicate that the program had positive and highly statistically significant effects across all measured outcomes from the first to the second evaluation: the intervention is associated with an increase in reading speed of approximately 9.03 words per minute (from a baseline of 20.37 wpm for treated students) and an improvement of around seven percentage points in reading accuracy

(from a baseline of 80.79% in evaluation moment1). The gap between treated and non-treated students concerning reading expressiveness also declined sharply.

We find evidence of increasing effects as we compare the coefficients regarding the second and third periods (β_3 and β_4). These results indicate that the program has cumulative (increasing) effects across all dimensions of interest over time. Computing Glass's delta effect sizes by dividing the impact estimates on the first column by the standard deviation of the control group for both post-intervention moments, the effect sizes are .42 and .47 standard deviations in the cases of moments 2 and 3, respectively. The first effect size is obtained by dividing the coefficient of the moment 2 effect, 9.028, by the standard deviation of the control group during moment 2, 21.65. Similarly, the .47 effect size is obtained by dividing the coefficient of the moment 3 effect, 11.87, by the standard deviation of the control group during moment 3, 25.13. These effect sizes compare favorably to the pooled effect size estimate of .37 obtained by the meta-analysis in Nickow et al. (2020), which considered similar tutoring interventions.

Table 3

Difference-in-Differences Estimates (all evaluation moments)

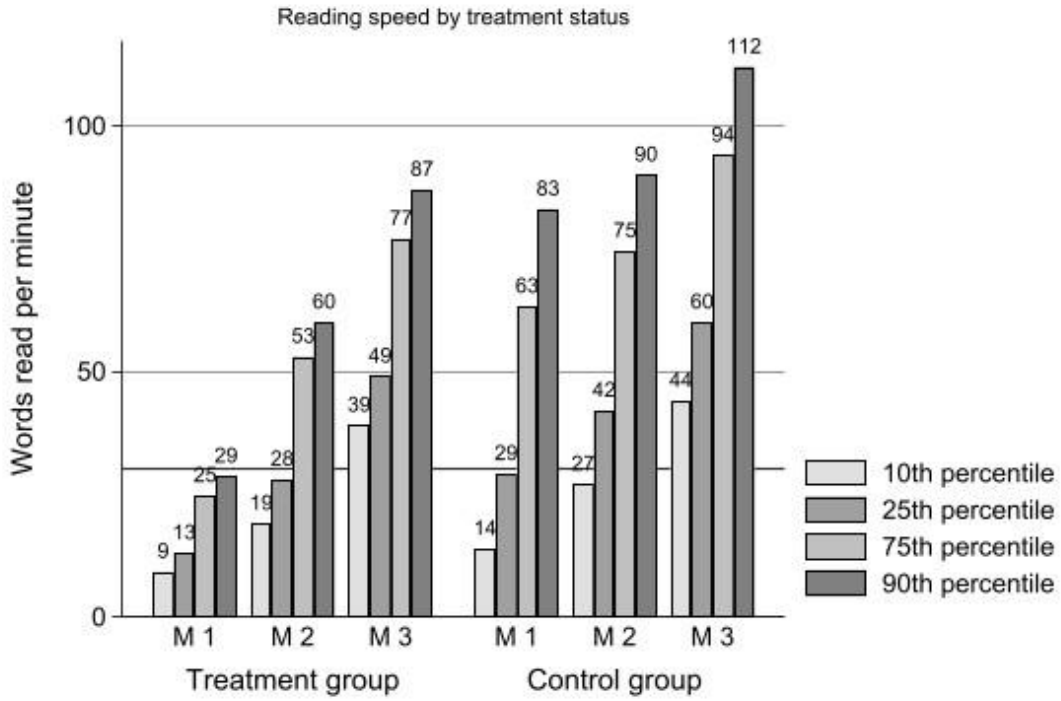
	(1) *Words per minute*	(2) Log (Words per minute)	(3) *Reading Accuracy*	(4) *Reading Expressiveness*
Moment 2	11.62***	0.31***	3.14***	0.25***
	(0.48)	(0.02)	(0.36)	(0.05)
Moment 3	31.34***	0.64***	6.23***	0.68***
	(0.58)	(0.02)	(0.38)	(0.05)
Moment 2 x Treatment	9.03***	0.40***	7.09***	0.65***
	(0.89)	(0.03)	(0.99)	(0.09)
Moment 3 x Treatment	11.87***	0.55***	9.04***	0.96***
	(1.34)	(0.04)	(1.04)	(0.10)
Constant	42.26***	3.52***	88.94***	2.57***
	(0.37)	(0.01)	(0.29)	(0.03)
N	7070	7070	7070	7070
Adjusted R-squared	0.82	0.76	0.43	0.52

Note. Standard errors clustered at the individual level are in parenthesis. Student (individual) fixed effects were included in each regression. Regarding the dependent variables, *'Reading Accuracy'* is the percentage of words read correctly, and *'Reading Expressiveness'* is an ordinal variable that may take values 1-5 (1 being the lowest level). 'Moment 2' and 'Moment 3' are dummies referring to post-intervention evaluations (pre-intervention outcomes were obtained in "Moment 1"). Significance levels: *** .01.

Figures 1, 2, and 3 show different perspectives of the evolution in reading speed throughout the year by treatment status. Some conclusions can be drawn from the figures. First, A to Z students improved significantly faster than their peers in the control group – especially for those who started at the highest reading speed percentiles among the treated students. Second, the patterns of within-group differences are different for intervention and control groups. In the intervention group, the highest reading (wpm) speed percentiles grow faster from moment 1 to moment 2 than from moment 2 to moment 3 (figures 1 and 3). In the control group, the lowest reading (wpm) speed percentiles grow faster from moment 1 to moment 2 but not from moment 2 to moment 3. Third, variance significantly increases in the intervention group and slightly decreases in the control group over the year.

Figure 1

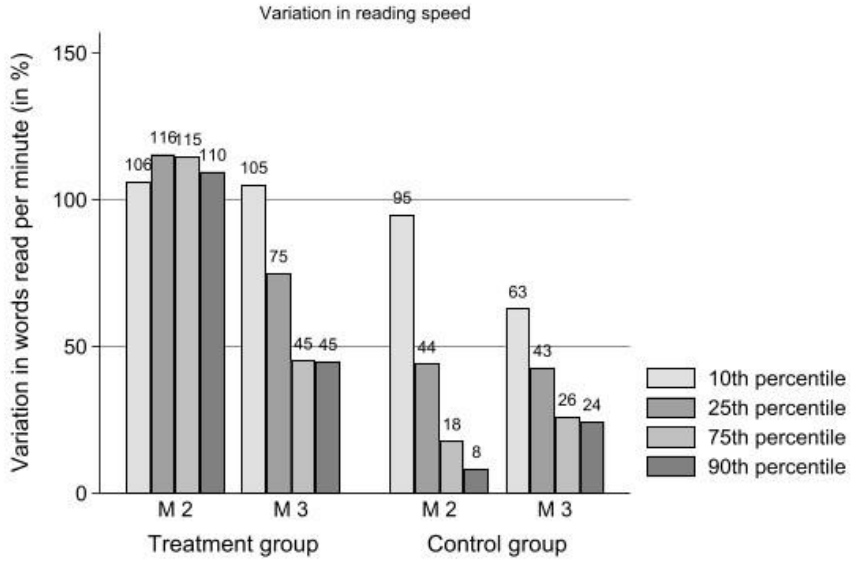
Evolution of Reading Speed Across Evaluation Moments for Selected Percentiles



Note. Only includes students at reading level "Text" by Evaluation Moment 1. The lower horizontal line marks the indicative threshold for program eligibility (30 words per minute). The value in each bar corresponds to a percentile of the reading speed distribution for a specific group (intervention or control) and period (moment 1, 2, or 3). The graph considers the same group of students over time, indicating the level of the 10th, 25th, 75th, and 90th percentiles of each group's reading speed distributions throughout the 2021/2022 school year. Note that the specific student of each percentile will typically not be the same at different periods.

Figure 2

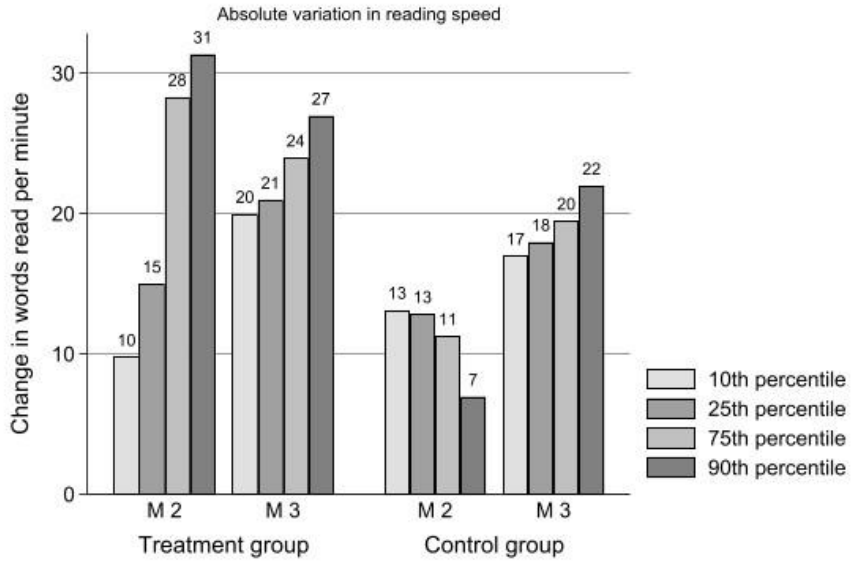
Percentage Variation of Reading Speed Across Evaluation Moments for Selected Percentiles



Note. Only includes students at reading level "Text" by the time of the Moment 1 Evaluation. 'Moment 2' reflects the variation between Moment 1 and Moment 2; Moment 3 considers the variation between Moment 2 and Moment 3.

Figure 3

Absolute Variation in Words per Minute Across Evaluation Moments for Selected Percentiles



Note. Only includes students at reading level "Text" by the time of the Moment 1 Evaluation. Moment 2 reflects the variation between Moment 1 and Moment 2; Moment 3 considers the variation between Moment 2 and Moment 3.

Discussion

The main goal of this study was to estimate the effect of an intervention program for struggling readers on students' *reading speed, accuracy,* and *expressiveness*. In this paper, we do not discuss the impact of the program's specific features (e.g., repeated readings, word recognition, morphological analysis) but the effect of the program as a whole. The results of the study clearly support Hypothesis 1 (H1) ("By Moment 2, differences between IG and CG in reading speed, accuracy, and expressiveness are significantly smaller than in Moment 1") and 2 (H2) ("differences between the intervention and the control group become significantly smaller over time"), but there are some caveats worth mentioning. First, while the differences between the intervention and the control group become smaller in Moment 2 (M2) and moment 3 (M3) in all reading dimensions (see Table 1), these developments are still insufficient for the intervention group to perform at the average of the control group in reading speed (but quite close in *reading accuracy* and *expressiveness*). Second, while the results suggest that the intervention program positively impacts the reading abilities of struggling readers, by M3, differences in *accuracy* and *expressiveness* are highly reduced, but the speed reduction is less expressive.

The general effectiveness of the program is not surprising. Previous studies (e.g., Hall, Dahl-Leonard, & Cannon, 2022; Wheldall et al., 2017) found that most reading interventions with struggling readers in a small group format are effective. Still, compared with intervention programs referred to in several meta-analyses (e.g., Gersten et al., 2020; Wanzek et al., 2016), the results of the A to Z program are encouraging. Wanzek et al. (2018) meta-analysis found weighted mean effect sizes of .39 for reading interventions for struggling readers from kindergarten through third grade, smaller than the reduction in effect sizes of the A to Z program (.58 reduction from M1 to M3 in accuracy; .55 in reading speed; and .75 in expressiveness). Hall, Dahl-Leonard, Cho, et al. (2022) found mean effect sizes of about .33 in 53 studies from 2000 to 2020.

It is important to stress that our training program is explicitly designed to increase *reading speed* (still integrating features such as morphological analysis or writing), with less emphasis on *reading accuracy* (although tutor-teachers provide timely feedback for students' incorrect readings) and *reading expressiveness* (although tutor-teachers model students' reading). Unfortunately, there is insufficient theoretical discussion about transference effects (one or several aspects of reading fluency improve when an isolated reading aspect improves), and our study design does not allow us to be conclusive. A

recent meta-analysis about reading fluency dosage by Maki and Hammerschmidt-Snidarich (2022, p. 149) found that "None of the reviews focused on parsing out the singular effects of targeted reading fluency practice only (i.e., non-multicomponent), which could facilitate our understanding of its causal mechanisms, including the effects of various reading fluency protocols and dosage." Also, Ripoll and Aguado's (2014) systematic review showed that improving decoding ability only does not significantly influence reading comprehension. However, the combination of strategies like enhancing motivation towards reading (a component of our program) and decoding improvement does. In practice, isolating effects in multi-component programs represents a tremendous challenge for reading researchers.

The disparities in the decrease of between-group differences suggest that word recognition (here measured through *reading speed*), perhaps the most relevant aspect of reading fluency (Rasinski et al., 2020), involves more complexity for struggling readers than *accuracy* and *expressiveness* and/or that there is room for improvement in the program. Alternatively, the significant increases in reading speed might bring more significant gains in *accuracy* and *expressiveness*. Moreover, a ceiling effect is much more apparent in these two features than in reading speed. In sum, we generically, but not entirely, confirm H1 and H2. Between-group differences become significantly smaller over time but not at the same rate for *speed*, *accuracy*, and *expressiveness*.

Our Hypothesis 3 (H3) ("the between-group difference reduction is more significant from Moment 1 to Moment 2 than from Moment 2 to Moment 3") stems from the classical learning curve theory, firstly formulated by Ebbinghaus in 1885 and from what is known in economics, as the "law of diminishing marginal returns" (after a certain point, each additional unit of input results in a smaller increase in output) (Blyth et al., 2022; Vazquez-Cognet, 2008). We hypothesized that the gains of our participants could follow a sigmoid-like learning function, with initial significant improvements in reading due to intervention (from M1 to M2), followed by a slowdown in the learning rate (from M2 to M3). According to the theory, we assumed participants would readily improve in reading since they were learning foundational skills and strategies. Still, progress might slow down as texts become more complex, a ceiling effect becomes closer, and motivation and engagement become more difficult (because subjects might perceive they are progressing at a lower rate). Still, while the improvement rate might slow, progress might continue, and the intervention would not become ineffective.

The results do not confirm our H3 but in a positive sense. In truth, the effect sizes of the differences between the intervention and the control group in *accuracy* and *expressiveness* reduce to the same extent between M1-M2 and M2-M3, and the difference reduction in *reading speed* is higher between M2-M3 than between M1-M2. This last result is all the more interesting as the control group also significantly accelerates in reading speed from M2 to M3 (which means that the intervention group accelerates even more).

Overall, we can both assert that participants' progress does not slow down and that, at the very least, the participants maintained the pace of improvement throughout the intervention period. Still, we might hypothesize that diminishing marginal returns could occur if the program is extended. If that happened, turning the program's focus to text comprehension would be defensible, reducing the focus on fluency.

Regarding Hypothesis 4 of our study ("…the highest performers in the intervention group progress significantly more in *reading speed*"), the results suggest within-group differences in response to the intervention. The highest performers readily respond to intervention (from moment 1 to moment 2) and continue to improve at a similar rate. The lowest performers increase significantly more from moment 2 to moment 3. Several studies (e.g., Daniel et al., 2022; Wanzek et al., 2018) found that students who start at lower reading levels tend to perform lower months or years later. Juel (1988) was maybe the first to test this effect, now known as the Matthew effect (Pfost et al., 2014; Protopapas et al., 2016; Stanovich, 1986). Curiously, we found no such effect in the control group.

Moreover, the initial variance in the intervention group was small by a selection effect. Still, the variance significantly increased over the year. This increase can be interpreted at least in two ways: the program does not entirely fit the poorest readers, an effect often found in reading intervention programs (Daniel et al., 2022), or good teaching tends to increase, not decrease, students' variance (Kauffman, 2002). Nevertheless, the results suggest that the program is most effective for struggling students within a specific range of reading speed (e.g., 25-30 wpm) and that providing the lowest-level readers with more intensive or tailored intervention is necessary. We, therefore, confirm our hypothesis. Still, since the lowest performers progress more from Moment 2 to Moment 3, a follow-up of students or the program's extension could clarify to what extent this hypothesis holds over a more extended period. Maybe lower performers take longer to respond but then progress more quickly.

Limitations

Two main limitations of the study are worth mentioning. The first is that the assignment of participants is not random. As we stated, we dealt with this limitation using a difference-in-differences (DID) approach, a valuable impact evaluation methodology when randomization is unavailable. Since most reading intervention studies in school contexts are quasi-experimental, DID is helpful because it removes the effects of groups' characteristics that remain constant over time. Another limitation of our study is that we cannot isolate the impact of specific program components, making it challenging to determine the best combination and dosage of teaching components and accurately assess the program's cost-benefit. Future research must approach this issue as well as the moment from which the improvements are minimal compared to the effort or cost of the intervention, that is, the moment in which the intervention should be discontinued. Also, we still need guidelines for intervention intensity. For example, the A to Z program is more intensive and extensive than similar programs. However, the intervention takes only 135 minutes weekly, apparently insufficient for most struggling readers to catch up with classmates.

Furthermore, since we do not have pre-intervention data for more than one period, we cannot test the validity of the common trends assumption that underlies our DID approach.

Conclusions and implications

In conclusion, our study shows that a systematic, frequent, and focused program like A to Z can significantly improve struggling readers' *reading accuracy*, *expressiveness*, and *speed*. Still, further program refinement might be necessary so the intervention group can come closer to classmates, the most apparent being intervention intensity. Unfortunately, increasing the number of weekly sessions is difficult because many teachers do not welcome students going in and out of the classroom. Moreover, we must be cautious against overloading the students' schedules.

It seems critical to have a permanent structure (or research team) that supports tutor-teachers and involves classroom teachers and schools. Also, the contents of the intervention must be evident, and the intervention must be structured, systematic, and as intense and extended as possible according to the best scientific evidence (e.g., Markovitz et al., 2022; Wanzek et al., 2018). As suggested by Slavin et al. (2011), "The evidence

does not support the idea that a relatively brief tutoring experience in first grade is enough to ensure success throughout elementary school and beyond…" (p. 22).

Although much is said about the importance of initial learning, too many students in Portugal and many other countries are left behind in the early years of school. It is, therefore, fundamental to develop compensatory policies for these students right from the first years of schooling, avoiding an accumulation of deficits progressively more challenging to overcome. These are expensive policies but still less expensive than late and often useless remedial attempts and their subsequent individual and social impacts.

References

Al Otaiba, S., Connor, C. M., Folsom, J. S., Wanzek, J., Greulich, L., Schatschneider, C., & Wagner, R. K. (2014). To wait in Tier 1 or intervene immediately: a randomized experiment examining first-grade response to intervention in reading. *Exceptional Children, 81*(1), 11-27. https://doi.org/10.1177/0014402914532234

Angrist, N., Bergman, P., & Matsheng, M. (2022). Experimental evidence on learning using low-tech when school is out. *Nature Human Behavior*. *6*, 941-950.

Angrist, N., Djankov, S., Goldberg, P. K., & Patrinos, H. A. (2021). Measuring human capital using global learning data. *Nature, 592*(7854), 403-408. https://doi.org/10.1038/s41586-021-03323-7

Artiles, C., & Jiménez, J. E. (2008). *CUDEA. Cuestionario para la detección temprana de las Dificultades de Aprendizaje (DEA) de la lectura, escritura y cálculo*. Gobierno de Canarias. https://drive.google.com/file/d/1LcgHVne8gHCAZsYhQ6bx1Q_3K6f__1pE/view

Audacity Team (2000). *Audacity software for windows. Free Audio Editor and Recorder*. https://audacityteam.org/

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukherji, S., & Walton, M. (2015). *Teaching at the right level: Evidence from randomized evaluations in India*. NBER Working Paper, 22746.

Blyth, M., Pontusson, J., & Baccaro, L. (2022). *Diminishing returns: the new politics of growth and stagnation*. Oxford University Press.

Brysbaert, M. (2019). How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language, 109*, 104047. https://doi.org/10.1016/j.jml.2019.104047

Buescu, H. C., Morais, J., Rocha, M. R., & Magalhães, V. F. (2015). *Programa e Metas Curriculares de Português do Ensino Básico [Portuguese Basic Education Curriculum and Goals*. Lisboa: DGE.

Cadime, I., Ribeiro, I., Cruz, J., Cosme, M. C., Meira, D., Viana, F. L., & Santos, S. (2022). An intervention in reading disabilities using a digital tool during the COVID-19 pandemic. *Frontiers in Psychology, 13:862383.* https://doi.org/10.3389/fpsyg.2022.862383

Cadime, I., Santos, S., Viana, F. L., & Ribeiro, I. (2023). The Relationship of Oral Reading fluency endurance to comprehension in an intermediate depth orthography. *Psicologia Educativa, 29*(2), 143-148. https://doi.org/10.5093/psed2023a6

Carlana, M., & La Ferrara, E. (2021). Apart But Connected: Online Tutoring and Student Outcomes During the Covid-19 Pandemic. *IZA Discussion Paper No. 14094*. http://dx.doi.org/10.2139/ssrn.3785058

Carvalhais, L., Limpo, T., Richardson, U., & Castro, S. L. (2020). Effects of the Portuguese graphogame on reading, spelling, and phonological awareness in second graders struggling to read. *Journal of Writing Research, 12*(1), 9-34. https://doi.org/10.17239/JOWR-2020.12.01.02

Castles, A., Rastle, K., & Nation, K. (2018). Ending the Reading Wars: reading acquisition from novice to expert. *Psychological Science in the Public Interest, 19*(1), 5-51. https://doi.org/10.1177/1529100618772271

Cockerill, M., Thurston, A., & O'Keeffe, J. (2023). Using fluency and comprehension instruction with struggling readers to improve student reading outcomes in

English elementary schools. *International Journal of Educational Research Open, 5*, 100264. https://doi.org/10.1016/j.ijedro.2023.100264
Daniel, J., Vaughn, S., Roberts, G., & Grills, A. (2022). The Importance of Baseline Word Reading Skills in Examining Student Response to a Multicomponent Reading Intervention. *Journal of Learning Disabilities, 55*(4), 259-271. https://doi.org/10.1177/00222194211010349
Denton, C. A., Hall, C., Cho, E., Cannon, G., Scammacca, N., & Wanzek, J. (2022). A meta-analysis of the effects of foundational skills and multi-component reading interventions on reading comprehension for primary-grade students. *Learning and Individual Differences, 93*. https://doi.org/10.1016/j.lindif.2021.102062
Dietrichson, J., Bøg, M., Filges, T., & Klint Jørgensen, A. M. (2017). Academic interventions for elementary and middle school students with low socio-economic status: A systematic review and meta-analysis. *Review of Educational Research, 87*(2), 243-282. https://doi.org/10.3102/0034654316687036
Domingue, B. W., Hough, H. J., Lang, D., & Yeatman, J. (2021). *Changing patterns of growth in oral reading fluency during the COVID-19 pandemic*. PACE.
Ecalle, J., Gomes, C., Auphan, P., Cros, L., & Magnan, A. (2019). Effects of policy and educational interventions intended to reduce difficulties in literacy skills in grade 1. *Studies in Educational Evaluation, 61*, 12-20. https://doi.org/10.1016/j.stueduc.2019.02.001
Ecalle, J., Magnan, A., Auphan, P., Gomes, C., Cros, L., & Suchaut, B. (2022). Effects of targeted interventions and of specific instructional time on reading ability in French children in grade 1. *European Journal of Psychology of Education, 37*(3), 605-625. https://doi.org/10.1007/s10212-021-00566-w
Ece Demir-Lira, Ö., Applebaum, L. R., Goldin-Meadow, S., & Levine, S. C. (2019). Parents' early book reading to children: relation to children's later language and literacy outcomes controlling for other parent language input. *Developmental Science, 22*(3), e12764. https://doi.org/10.1111/desc.12764
Filmer, D., Rogers, H., Angrist, N., & Sabarwal, S. (2020). Learning-adjusted years of schooling (LAYS): Defining a new macro measure of education. *Economics of Education Review, 77*, 101971. https://doi.org/10.1016/j.econedurev.2020.101971
Fletcher, J. M., Lyon, G. R., Fuchs, L. S., & Barnes, M. A. (2018). *Learning disabilities: From identification to intervention*. Guilford.
Fumagalli, J., Barreyro, J. P., & Jaichenco, V. (2019). Reading fluency differences between oral and silent reading comprehension. *Global Journal of Human-Social Science, 19(G9)*, 9-16.
Gersten, R., Haymond, K., Newman-Gonchar, R., Dimino, J., & Jayanthi, M. (2020). Meta-Analysis of the impact of reading interventions for students in the primary grades. *Journal of Research on Educational Effectiveness, 13*(2), 401-427. https://doi.org/10.1080/19345747.2019.1689591
Glawe, L., & Wagner, H. (2022). Is schooling the same as learning? – The impact of the learning-adjusted years of schooling on growth in a dynamic panel data framework. *World Development, 151*, 105773. https://doi.org/10.1016/j.worlddev.2021.105773
Graham, S., Gillespie, A., & McKeown, D. (2013). Writing: importance, development, and instruction. *Reading and Writing, 26*(1), 1-15. https://doi.org/10.1007/s11145-012-9395-2
Graham, S., Liu, X., Aitken, A., Ng, C., Bartlett, B., Harris, K. R., & Holzapfel, J. (2018). Effectiveness of literacy programs balancing reading and writing

instruction: a meta-analysis. *Reading Research Quarterly, 53*(3), 279-304. https://doi.org/10.1002/rrq.194

Guryan, J., Ludwig, J., Bhatt, M. P., Cook, P. J., Davis, J. M. V., Dodge, K., Farkas, G., Fryer Jr., R. G., Mayer, S., Pollack, H., Steinberg, L., & Stoddard, G. (2023). Not Too Late: Improving Academic Outcomes among Adolescents. *American Economic Review, 113*(3), 738-65.

Hall, C., Dahl-Leonard, K., & Cannon, G. (2022). Observing two reading intervention programs for students with dyslexia. *Exceptionality, 30*(2), 109-125. https://doi.org/10.1080/09362835.2021.1938067

Hall, C., Dahl-Leonard, K., Cho, E., Solari, E. J., Capin, P., Conner, C. L., Henry, A. R., Cook, L., Hayes, L., Vargas, I., Richmond, C. L., & Kehoe, K. F. (2022). Forty years of reading intervention research for elementary students with or at risk for dyslexia: a systematic review and meta-analysis. *Reading Research Quarterly.* https://doi.org/10.1002/rrq.477

Hasbrouck, J., & Tindal, G. (2017). *An update to compiled ORF norms (Technical Report No. 1702)*. O. Eugene, Behavioral Research and Teaching.

IAVE. (2017). *Resultados Globais PIRLS 2016 ● ePIRLS 2016 – PORTUGAL. Literacia de leitura & literacia de leitura online*. IAVE.

Jimenez, J. E. (2010). Response to Intervention (RTI) model: a promising alternative for identifying students with learning disabilities? *Psicothema, 22*(4), 932-934. http://www.ncbi.nlm.nih.gov/pubmed/21044534

Jones, B. T., Erchul, W. P., & Geraghty, C. A. (2021). Supplemental reading interventions implemented by paraprofessionals: A meta-analysis. *Psychology in the Schools, 58*(4), 723-741. https://doi.org/10.1002/pits.22427

Juel, C. (1988). Learning to read and write: A longitudinal study of fifty-four children from first through fourth grade. *Journal of Educational Psychology, 80*, 437-447.

Kaffenberger, M., & Pritchett, L. (2017). More school or more learning? Evidence from learning profiles from the financial inclusion insights data. *Research on Improving Systems of Education Working Paper 17/012*

Kara, Y., Kamata, A., Qiao, X., Potgieter, C. J., & Nese, J. F. T. (2023). equating oral reading fluency scores: a model-based approach. *Educational and Psychological Measurement*. https://doi.org/10.1177/00131644221148122

Kauffman, J. M. (2002). *Education deform: Bright people sometimes say stupid things about education*. Scarecrow Press.

Klein, O., & Kogan, I. (2013). Does reading to children enhance their educational success?. *Child Indicators Research, 6*(2), 321-344. https://doi.org/10.1007/s12187-012-9174-2

Kodila-Tedika, O., & Otchia, C. S. (2022). The effects of free primary education in the Democratic Republic of Congo: A difference-in-differences approach. *Review of Development Economics, 26*(4), 2109-2120. https://doi.org/10.1111/rode.12922

Kraft, M. A., List, J. A., Livingston, J. A., & Sadoff, S. (2022). Online Tutoring by College Volunteers: Experimental Evidence from a Pilot Program. *AEA Papers and Proceedings, 112*, 614-18.

Lavrijsen, J., & Nicaise, I. (2016). Educational tracking, inequality and performance: New evidence from a differences-in-differences technique. *Research in Comparative and International Education, 11*(3), 334-349. https://doi.org/10.1177/1745499916664818

Lee, J., & Yoon, S. Y. (2017). The effects of repeated reading on reading fluency for students with reading disabilities: a meta-analysis. *Journal of Learning Disabilities, 50*(2), 213-224. https://doi.org/10.1177/0022219415605194

Liversedge, S. P., Drieghe, D., Li, X., Yan, G., Bai, X., & Hyönä, J. (2016). Universality in eye movements and reading: A trilingual investigation. *Cognition, 147*, 1-20. https://doi.org/https://doi.org/10.1016/j.cognition.2015.10.013

Maki, K. E., & Hammerschmidt-Snidarich, S. (2022). Reading fluency intervention dosage: A novel meta-analysis and research synthesis. *Journal of School Psychology, 92*, 148-165. https://doi.org/10.1016/j.jsp.2022.03.008

Markovitz, C. E., Hernandez, M. W., Hedberg, E. C., & Whitmore, H. W. (2022). Evaluating the effectiveness of a volunteer one-on-one tutoring model for early elementary reading intervention: a randomized controlled trial replication study. *American Educational Research Journal, 59*(4), 788-819. https://doi.org/10.3102/00028312211066848

Martins, P. (2017a). Can non-cognitive skills programs improve achievement? Quasi-experimental evidence from EPIS. *GLO Discussion Paper Series*, 105. https://ideas.repec.org/p/zbw/glodps/105.html

Martins, P. (2017b). (How) Do non-cognitive skills programs improve adolescent school achievement? Experimental Evidence. *IZA Discussion Papers 10950, Institute of Labor Economics (IZA).* https://doi.org/https://ideas.repec.org/p/iza/izadps/dp10950.html

Mather, N., White, J., & Youman, M. (2020). Dyslexia around the world: a snapshot. *Learning Disabilities: A Multidisciplinary Journal, 25*(1), 1-17. https://doi.org/10.18666/ldmj-2020-v25-i1-9552

McCutchen, D., Northey, M., Herrera, B. L., & Clark, T. (2022). What's in a word? Effects of morphologically rich vocabulary instruction on writing outcomes among elementary students. *Reading and Writing, 35*(2), 325-351. https://doi.org/10.1007/s11145-021-10184-z

Meyer, L. A., Wardrop, J. L., Stahl, S. A., & Linn, R. L. (1994). Effects of reading storybooks aloud to children. *Journal of Educational Research, 88*(2), 69-85. https://doi.org/10.1080/00220671.1994.9944821

Moats, L. (2009). Knowledge foundations for teaching reading and spelling. *Reading and Writing, 22*(4), 379-399. https://doi.org/10.1007/s11145-009-9162-1

Motiejunaite, A., Noorani, S., & Monseur, C. (2014). Patterns in national policies for support of low achievers in reading across Europe. *British Educational Research Journal, 40*(6), 970-985. https://doi.org/10.1002/berj.3125

Mullis, I. V. S., & Martin, M. O. (2017). *PIRLS 2016 International Results in Reading*. TIMSS & PIRLS International Study Center, Boston College.

Ng, M. M. R., Bowers, P. N., & Bowers, J. S. (2022). A promising new tool for literacy instruction: The morphological matrix. *PLoS ONE, 17*(1 January), e0262260 https://doi.org/10.1371/journal.pone.0262260

Nickow, A., Oreopoulos, P., & Quan, V. (2020). *The impressive effects of tutoring on prek-12 learning: a systematic review and meta-analysis of the experimental evidence*. NBER Working Paper, 27476. http://www.nber.org/papers/w27476

Niklas, F., Cohrssen, C., & Tayler, C. (2016). The Sooner, the Better: Early Reading to Children. *SAGE Open, 6*(4) https://doi.org/10.1177/2158244016672715

Paige, D. D., Smith, G. S., Rasinski, T. V., Rupley, W. H., Magpuri-Lavell, T., & Nichols, W. D. (2019). A path analytic model linking foundational skills to

Grade 3 state reading achievement. *The Journal of Educational Research, 112*(1), 110-120. https://doi.org/10.1080/00220671.2018.1445609

Park, Y. (2019). Effectiveness of repeated reading intervention packages for improving the reading skills of students with disabilities. *Asia Life Sciences, Suppl18*(2), 567-578. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85063692621&partnerID=40&md5=629fa7d04f2dd8858e7d22a625cccac7

Pfost, M., Hattie, J., Dörfler, T., & Artelt, C. (2014). individual differences in reading development: a review of 25 years of empirical research on Matthew Effects in reading. *Review of Educational Research, 84*(2), 203-244. https://doi.org/10.3102/0034654313509492

Protopapas, A., Parrila, R., & Simos, P. G. (2016). In search of Matthew Effects in reading. *Journal of Learning Disabilities, 49*(5), 499-514. https://doi.org/10.1177/0022219414559974

Psacharopoulos, G., Collis, V., Patrinos, H. A., & Vegas, E. (2021). The COVID-19 cost of school closures in earnings and income across the world. *Comparative Education Review, 65*(2), 271-287. https://doi.org/10.1086/713540

Rasinski, T., Galeza, A., Vogel, L., Viton, B., Rundo, H., Royan, E., Nemer Shaheen, R., Bartholomew, M., Kaewkaemket, C., Stokes, F., Young, C., & Paige, D. (2022). Oral reading fluency of college graduates: toward a deeper understanding of college ready fluency. *Journal of Adolescent & Adult Literacy, 66*(1), 23-30. https://doi.org/10.1002/jaal.1248

Rasinski, T., Rupley, W. H., Paige, D., & Young, C. (2020). *Reading fluency*. MDPI.

Salceda, J. C. R., Montesinos, M. M. T., & Alonso, G. A. (2020). Reading rate in Spanish-speaking students: A meta-analysis. *Revista de Psicodidactica, 25*, 158-165.

Ripoll, J. C., & Aguado, G. (2014). Reading comprehension improvement for Spanish students: A meta-analysis. *Revista de Psicodidáctica, 19*(1), 27-44. https://doi.org/10.1387/RevPsicodidact.9001

Risberg, A. K., Widlund, A., Hellstrand, H., Vataja, P., & Salmi, P. (2023). Profiles of reading and fluency and spelling skills: stability and change across the early school years. *Scandinavian Journal of Educational Research* https://doi.org/10.1080/00313831.2023.2228822

Rodrigues, B., Cadime, I., Freitas, T., Choupina, C., Baptista, A., Viana, F. L., & Ribeiro, I. (2022). Assessing oral reading fluency within and across grade levels: Development of equated test forms. *Behavior Research Methods, 54*(6), 3043-3054. https://doi.org/10.3758/s13428-022-01806-7

Rose, R. A., & Bowen, N. K. (2019). Difference-in-differences as an alternative to pretest-posttest regression for social work intervention evaluation and research. *Social Work Research, 43*(4), 247-258. https://doi.org/10.1093/swr/svz017

Rosén, M., & Gustafsson, J. E. (2016). Is computer availability at home causally related to reading achievement in grade 4? A longitudinal difference in differences approach to IEA data from 1991 to 2006. *Large-Scale Assessments in Education, 4*(1), 5. https://doi.org/10.1186/s40536-016-0020-8

Rosling, H., Rönnlund, A., R., & Rosling, O. (2018). *Factfulness: ten reasons we're wrong about the world--and why things are better than you think.* Flatiron Books.

Rupley, W. H., Nichols, W. D., Rasinski, T. V., & Paige, D. (2020). Fluency: deep roots in reading instruction. *Education Sciences, 10*(6), 1-11, 155. https://doi.org/10.3390/educsci10060155

Salinas, P., & Solé-Ollé, A. (2018). Partial fiscal decentralization reforms and educational outcomes: A difference-in-differences analysis for Spain. *Journal of Urban Economics, 107*, 31-46. https://doi.org/10.1016/j.jue.2018.08.003

Samson, J. F., Hines, S. J., & Li, K. (2015). Effective use of paraprofessionals as early intervention reading tutors in grades K-3. *Mentoring and Tutoring: Partnership in Learning*, *23*(2), 164-177. https://doi.org/10.1080/13611267.2015.1049014

Savage, R., Maiorino, K., Gavin, K., Horne-Robinson, H., Georgiou, P. G., & Deacon, H. (2023). Contrasting Direct Instruction in Morphological Decoding and Morphological Inquiry-Analysis Interventions in Grade 3 Children With Poor Morphological Awareness. *Journal of Learning Disabilities*. https://doi.org/10.1177/00222194231161117

See, B. H., & Gorard, S. (2020). Effective classroom instructions for primary literacy: A critical review of the causal evidence. *International Journal of Educational Research, 102*, 101577. https://doi.org/https://doi.org/10.1016/j.ijer.2020.101577

Seymour, P. H. K., Aro, M., Erskine, J. M., Wimmer, H., Leybaert, J., Elbro, C., Lyytinen, H., Gombert, J. E., Le Normand, M. T., Schneider, W., Porpodas, C., Ragnarsdottir, H., Tressoldi, P., Vio, C., De Groot, A., Licht, R., Iønnessen, F. E., Castro, S. L., Cary, L., Defior, S., Martos, F., & Olofsson, Å. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology, 94*(2), 143-174. https://doi.org/10.1348/000712603321661859

Simpson, I. C., Moreno-Pérez, F. J., Rodríguez-Ortiz, I. R., Valdés-Coronel, M., & Saldaña, D. (2020). The effects of morphological and syntactic knowledge on reading comprehension in Spanish speaking children. *Reading and Writing, 33*(2), 329-348. https://doi.org/10.1007/s11145-019-09964-5

Slavin, R. E., Lake, C., Davis, S., & Madden, N. A. (2011). Effective programs for struggling readers: A best-evidence synthesis. *Educational Research Review, 6*(1), 1-26. https://doi.org/10.1016/j.edurev.2010.07.002

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*, 360-407.

StataCorp (2023). *Stata Statistical Software: Release 18*. College Station, TX: StataCorp LLC.

Steinmann, I., Strello, A., & Strietholt, R. (2023). The effects of early between-school tracking on gender segregation and gender gaps in achievement: a differences-in-differences study. *School Effectiveness and School Improvement, 34*(2), 189-208. https://doi.org/10.1080/09243453.2023.2165510

Strello, A., Strietholt, R., Steinmann, I., & Siepmann, C. (2021). Early tracking and different types of inequalities in achievement: difference-in-differences evidence from 20 years of large-scale assessments. *Educational Assessment, Evaluation and Accountability, 33*(1), 139-167 https://doi.org/10.1007/s11092-020-09346-4

Sucena, A., Silva, A. F., & Marques, C. (2021). Promoting foundation reading skills with at-risk students. *Frontiers in Psychology, 12*, 671733 https://doi.org/10.3389/fpsyg.2021.671733

Sucena, A., Silva, A. F., & Marques, C. (2022). Reading skills intervention during the Covid-19 pandemic. *Humanities and Social Sciences Communications, 9*(1), 45. https://doi.org/10.1057/s41599-022-01059-x

UNICEF. (2022, 19 September). *Only a third of 10-year-olds globally are estimated to be able to read and understand a simple written story.*

https://www.unicef.org/bulgaria/en/press-releases/unicef-only-third-10-year-olds-globally-are-estimated-be-able-read-and-understand
UNICEF. (October 2021). *The State of the World's Children 2021: On my mind – Promoting, protecting and caring for children's mental health*. UNICEF.
Van Norman, E. R., Nelson, P. M., & Klingbeil, D. A. (2020). Typical Rates of Regression After Exiting Supplemental Interventions in Reading. *School Psychology Review, 49*(4), 510-515. https://doi.org/10.1080/2372966X.2020.1720801
Vazquez-Cognet, J. J. (2008). The production of mathematical problems: a diminishing marginal returns experiment. *International Review of Economics Education, 7*(1), 103-116. https://doi.org/10.1016/s1477-3880(15)30093-1
Verhoeven, L., & Perfetti, C. (2022). Universals in learning to read across languages and writing systems. *Scientific Studies of Reading, 26*(2), 150-164. https://doi.org/10.1080/10888438.2021.1938575
Wanzek, J., Stevens, E. A., Williams, K. J., Scammacca, N., Vaughn, S., & Sargent, K. (2018). Current evidence on the effects of intensive early reading interventions. *Journal of Learning Disabilities, 51*(6), 612-624. https://doi.org/10.1177/0022219418775110
Wanzek, J., Vaughn, S., Scammacca, N., Gatlin, B., Walker, M. A., & Capin, P. (2016). Meta-Analyses of the effects of tier 2 type reading interventions in grades K-3. *Educational Psychology Review, 28*(3), 551-576. https://doi.org/10.1007/s10648-015-9321-7
Wheldall, K., Wheldall, R., Madelaine, A., Reynolds, M., & Arakelian, S. (2017). Further evidence for the efficacy of an evidence-based, small group, literacy intervention program for young struggling readers. *Australian Journal of Learning Difficulties, 22*(1), 3-13. https://doi.org/10.1080/19404158.2017.1287102
World Bank. (2018a). *World development indicators*, Washington, DC: World Bank.
World Bank. (2018b). *World development report 2018: Learning to realize education's promise*, Washington DC
World Medical Association. (2013). World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA, 310(20)*, 2191–2194. doi:10.1001/jama.2013.281053