



## UvA-DARE (Digital Academic Repository)

### Implementation takes time: Reduction of literacy problems in schools implementing an early-literacy intervention

van der Weijden, Fae A.; van den Boer, Madelon; Zijlstra, Bonne J.H.; de Jong, Peter F.

**DOI**

[10.1080/19345747.2024.2384365](https://doi.org/10.1080/19345747.2024.2384365)

**Publication date**

2025

**Document Version**

Final published version

**Published in**

Journal of Research on Educational Effectiveness

**License**

CC BY

[Link to publication](#)

**Citation for published version (APA):**

van der Weijden, F. A., van den Boer, M., Zijlstra, B. J. H., & de Jong, P. F. (2025). Implementation takes time: Reduction of literacy problems in schools implementing an early-literacy intervention. *Journal of Research on Educational Effectiveness*, 18(4), 918-950. <https://doi.org/10.1080/19345747.2024.2384365>

**General rights**




It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

**Disclaimer/Complaints regulations**

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

*UvA-DARE is a service provided by the library of the University of Amsterdam (<https://dare.uva.nl>)*

# Implementation Takes Time: Reduction of Literacy Problems in Schools Implementing an Early-Literacy Intervention

Fae A. van der Weijden , Madelon van den Boer, Bonne J. H. Zijlstra , and Peter F. de Jong 

Department of Child Development and Education, University of Amsterdam, Amsterdam, the Netherlands

## ABSTRACT

Early-literacy interventions might prevent reading problems in the long term, but effects are rarely examined at scale. In this study, we examined whether the large-scale implementation of the Dutch early-literacy intervention *Build!* reduced the percentage of readers with difficulties and improved mean reading skills at the school level. Transfer effects to spelling and reading comprehension were also examined. Over the course of 6 years, schools not implementing *Build!* (61–126 schools, depending on the outcome measure) were compared to 72 to 145 schools that introduced *Build!* during the project. Per year, intervention schools were modeled as using or not using the intervention. Using difference-in-difference models, we examined changes in literacy skills from the moment the intervention was introduced. Findings indicated that there was no immediate effect of the intervention. However, after the intervention had been used for 2 years, the percentage of children with difficulties in reading, spelling, and reading comprehension started to decrease and the mean reading and spelling ability increased. Results suggest that large-scale evaluations of interventions should be continued for several years, as effects might show several years after the implementation of the intervention.

## ARTICLE HISTORY

Received 17 May 2023  
Revised 24 June 2024  
Accepted 2 July 2024


## KEYWORDS

Prevention; literacy difficulties; reading; spelling; large-scale

In primary school, 3% to 10% of children experience severe reading difficulties (Fluss et al., 2009; Snowling, 2013). Reading difficulties are associated with a negative academic self-concept (Bear et al., 2002; Zeleke, 2004), lower school achievement (Ferrer et al., 2015; Mol & Bus, 2011), and school dropout in adolescence (Daniel et al., 2006). In turn, school careers tend to affect children's later employability (Annie E. Casey Foundation, 2010). To prevent these negative outcomes for children and society at large, there is a need for reading interventions that can improve reading skills and reduce the number of children with reading difficulties.

Reading problems have been proven difficult to overcome once they have arisen, especially when it concerns problems in reading fluency (Ferrer et al., 2015; Torgesen

**CONTACT** Fae A. van der Weijden  [f.a.vanderweijden@uva.nl](mailto:f.a.vanderweijden@uva.nl)  University of Amsterdam, Postbus 15667, Code 1115, t.a.v. Madelon van den Boer, 1001 ND Amsterdam, the Netherlands.

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/19345747.2024.2384365>.

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

et al., 2001). Therefore, early-literacy interventions have been developed with a focus on preventing reading difficulties. These interventions, often starting in kindergarten or first grade, have generally been shown to be effective in reducing later reading problems (Ehri et al., 2001a; Lovett et al., 2017; Wanzek & Vaughn, 2007). However, most of the evidence for the effects of early-literacy interventions comes from relatively small-scale studies. Moreover, in most studies, interventions were implemented under the strict guidance of researchers. Researchers were involved in, for example, selecting qualified teachers, providing training and support, and frequently visiting schools to monitor and stimulate the implementation of the intervention (e.g., Mathes et al., 2005; Zijlstra et al., 2021). Little is known about the effectiveness of these interventions when implemented on a large scale by schools, without researchers involved. In the current study, we examined whether the large-scale school-based implementation of the early-literacy intervention *Build!* leads to an improvement in reading skills within schools as well as a decrease in the number of children with reading difficulties.

*Build!* is a Dutch computer-assisted early-literacy intervention for children at risk for reading difficulties. The intervention starts in kindergarten and continues for two years. Children practice pre-literacy skills, including letter-sound correspondences and phoneme blending, as well as decoding of monosyllabic words. As in many pre- and early-literacy interventions (e.g., Suggate, 2016), *Build!* supports the acquisition of (pre) literacy skills (Regtvoort & van der Leij, 2007). To avoid fade-out effects, the intervention is continued in grades 1 and 2, with the start of formal reading instruction (Bailey et al., 2017; Zijlstra et al., 2021). From grade 1 onward, the focus of the program shifts from letters to letter clusters, from monosyllabic to bisyllabic words, from consistent to inconsistent words, and from reading accuracy to reading fluency. The intervention is meant to be provided in three to four sessions of 10 to 15 minutes per week. The child is assisted by a tutor who reads aloud instructions from the screen and stimulates the child to stay on task.

Two randomized controlled trials (RCTs) with children at risk for reading difficulties have been conducted to evaluate the effects of this early-literacy intervention (Regtvoort et al., 2013; Zijlstra et al., 2021). Regtvoort et al. (2013) examined the effect of the second part of the intervention, the period from the middle of grade 1 to the middle of grade 2. Unfortunately, treatment integrity in part of the intervention group was low. Therefore, the intervention group was split in groups that did and did not complete the intervention. These intervention groups did not differ in reading ability at the start of the intervention. Regtvoort et al. (2013) found that the intervention group displayed better word-reading and reading-comprehension abilities than the no-intervention group at posttest and 1 year after the conclusion of the intervention. The effect of the *Build!* intervention, from the second year of kindergarten through the middle of grade 2, was tested in an RCT by Zijlstra et al. (2021, see also Zijlstra, 2015). The intervention and control groups were followed until the end of second grade, half a year after the intervention had finished. Results showed that the intervention was only effective in a subgroup of children. This subgroup comprised children whose parents had provided information about the prevalence of dyslexia in the family, termed the family risk information (FRinfo) intervention group. The subsample in which the intervention was not effective consisted mostly of children from immigrant, non-Dutch-speaking families with a low socioeconomic status. Children in the FRinfo

intervention group were more fluent in word, pseudoword, and text reading than children in the control group. Follow-up of the FRinfo subsample showed that these effects were sustained until sixth grade—4 years after the intervention had finished. In addition, the percentage of children with reading difficulties (defined as the lowest-scoring 25% based on national norms) in grade 6 was substantially lower than in the control group.

Because of these promising results, school districts and school boards in the Netherlands have started to stimulate and facilitate the implementation of the intervention, aiming for an overall decrease in the number of children with reading difficulties in their schools. As a result, since 2014, the number of schools that have implemented the intervention has gradually increased to about 80% of the primary schools in the Netherlands (about 5,000 schools). However, unlike the researcher-guided RCTs in which the implementation of the intervention is closely monitored by the researchers, in these schools, the implementation of the intervention is the schools' own responsibility. Effects of an intervention can be lower when used in natural school settings (e.g., Sirinides et al., 2018). Therefore, it seems apt to examine whether this policy has led to the desired outcomes. Accordingly, in the current study, the short- and long-term effects of the early-literacy intervention *Build!* on literacy outcomes were examined at the school level for schools who have autonomously implemented the intervention.

### ***Effects of Early-Literacy Interventions***

Current evidence suggests that interventions for children with or at risk for reading difficulties can be effective (Galuschka et al., 2014; Lovett et al., 2017; Suggate, 2016; Wanzek & Vaughn, 2007). However, few studies have examined the long-term effects of early-literacy interventions in preschool and kindergarten, that is, around 1 year after the intervention concluded (Suggate, 2010). Such studies show that effects fade out (Suggate, 2016). There is only little evidence that early-literacy interventions can produce effects that are still visible after second grade (see Lovett et al. (2017) and Zijlstra et al. (2021) as exceptions). Clearly, longer-lasting effects are needed, as reading difficulties might continue or even emerge after second grade (Simmons et al., 2008; Torppa et al., 2015).

It may take schools a couple of years to implement an intervention properly and to reach intervention effects. Harn et al. (2013) argued that interventions should be adapted to local circumstances. For example, in the case of the current intervention, schools have to decide which tutors they prefer (adult volunteers, older peers, parents) and how they will instruct and motivate them. Moreover, schools have to choose and implement a procedure for the selection of children at risk for reading difficulties, which often involves regular testing of preliteracy skills. There are also issues concerning the planning and routine of the intervention sessions as well as integrating these sessions into the ongoing processes and policies at the school (Prenger et al., 2022). Thus, for schools, it is no small feat to implement such an intervention. It requires time, resources, and leadership (Durlak & DuPre, 2008), and it might take more than 1 year to reach full treatment integrity, that is, to implement the intervention as intended (Gresham et al., 2000).

To our knowledge, there are very few studies that have considered the effects of a literacy intervention over time, that is, in subsequent cohorts. One exception, a study by Torgesen (2009), showed that in a large-scale implementation of the *response to intervention instructional model*, intervention effects increased during 3 years of implementation. In the current large-scale study, we followed schools for multiple years to evaluate the effects of the early-literacy intervention *Build!*.

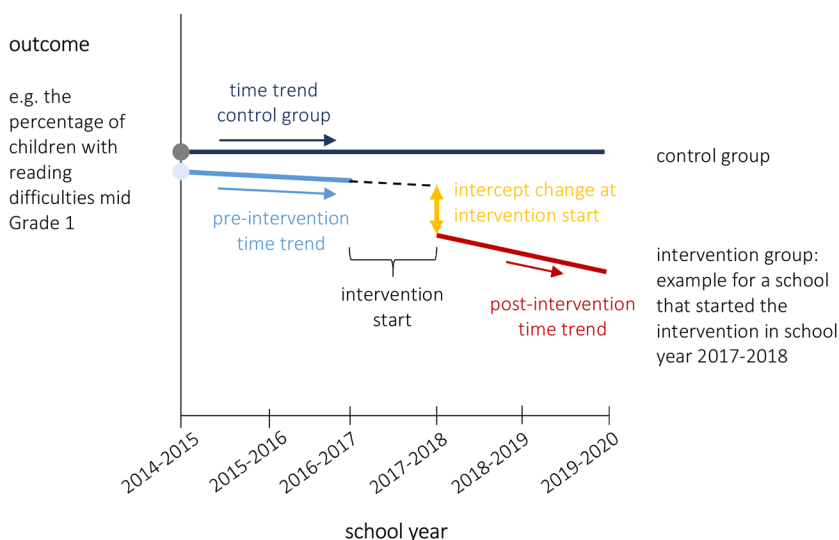
### **Difference-in-Difference (DiD) Design**

Effectiveness studies with random assignment of participants to an intervention and a control group are generally considered to have the best qualifications for assessing the effect of an intervention (Thompson & Panacek, 2006). However, large-scale RCTs in educational settings are often not feasible. In addition to the costs of large-scale RCTs as a factor, schools are inclined to implement a promising intervention and do not want students in a control group until long-term effects have been established. In the current study, we used the next-best solution, an extended DiD design (Mascha & Sessler, 2019; Wing et al., 2018). Using this design, we analyzed the effects of *Build!* at the school level, rather than at the individual level, answering the question of whether schools found an improvement in the mean reading ability and/or a decrease in the number of students with reading difficulties from the moment *Build!* was implemented. The DiD design is often used in applied econometrics and public health research to examine the effects of large-scale interventions and policy decisions, when an RCT is not feasible (Wing et al., 2018). It has also been used to evaluate interventions in education (Sims et al., 2022), but—to our knowledge—not yet in the field of reading research.

The DiD design in its simplest form is a pretest-posttest quasi-experimental design with an intervention and a control group (Fredriksson & de Oliveira, 2019; Wing et al., 2018). It is used to evaluate whether the gains are larger in the intervention group than in the control group. These gains can be measured at the individual level and also at the school level. The DiD is defined as the difference between pretest and posttest in the intervention group minus the difference between pretest and posttest in the control group. It is literally a “difference of differences” (Fredriksson & de Oliveira, 2019). The design assumes that the gains in the control group and intervention group are similar in the absence of the intervention.

The DiD design can be extended by the addition of multiple pretests and posttests, resulting in a comparative interrupted time series (CITS) design (Jacob et al., 2016). This CITS design controls for the potential threat that larger gains in the intervention group are due to differences in pre-intervention growth between the groups. Therefore, the main assumption in the CITS design is that the pre-intervention time trends in schools that implemented the intervention are similar to the concurrent trends in schools that did not implement the intervention (Jacob et al., 2016; Mascha & Sessler, 2019). An example of this design is displayed in Figure 1.

A potential threat to the CITS design is that the control and intervention group are not sufficiently comparable, for example, when co-occurring events during the period of the study, unrelated to the intervention, affect the intervention and control group differently (Jacob et al., 2016). Some have advocated that schools in both groups should be located in the same geographical region so that they are more or less



**Figure 1.** Difference-in-difference design with multiple pretests and posttests. *Note.* A difference-in-difference model with multiple pretests and posttests controls for differences between the intervention and control group at the intervention start (grey vs. light blue) and in pre-intervention time trend (dark blue vs. blue). It determines a change immediately after the intervention is introduced (yellow) and after the intervention is used for multiple years (red).

susceptible to the same regime (Cook et al., 2008; but see Jacob et al., 2016). It might also be useful when clusters of schools start to implement the intervention in different years (e.g., Mascha & Sessler, 2019; van de Werfhorst, 2019). When the start of the intervention differs across schools, various clusters of schools are expected to show changes in time trends (i.e., trends over cohorts) at different points in time. Thus, schools are modeled as using or not using the intervention at a particular point in time (Mascha & Sessler, 2019). This might provide some extra control for co-occurring events that influence the intervention and control group differently.

### Current Study

The main question in this study was whether the early-literacy intervention *Build!* led to an improvement in reading skills and a reduction of children with reading difficulties following implementation. We examined short-term effects in first and second grade, that is, during the intervention (the middle of grade 1 and the end of grade 1) and at the planned end of the intervention (the middle of grade 2). In addition, we investigated follow-up effects half a year and 1 year after the planned end of the intervention (the end of grade 2 and the middle of grade 3, respectively). Furthermore, the study investigated whether the effects were visible after schools had been using the program for 1 year and whether intervention effects increased with the number of years that schools had been using the program. More experience with the program could result in larger intervention effects (Harn et al., 2013; Torgesen, 2009).

Two additional research questions were addressed. The first was whether the effects of the intervention would transfer to spelling, as the trained skills also contribute to

spelling (Ehri et al., 2001a; Suggate, 2016), and to reading comprehension, as reading fluency is an important prerequisite for reading comprehension (Florit & Cain, 2011). Effects of *Build!* on spelling were previously found in Zijlstra et al. (2021), and effects on reading comprehension were reported in Regtvoort et al. (2013). The second additional research questions pertained to effects on an unrelated skill: mathematics. If no effects on mathematics were found, this would increase the likelihood that the effects on literacy skills were due to the introduction of *Build!* and not to other concurrent events that led to an overall improvement in school achievement.

## Method

### Design

The study had a CITS design with a no-intervention control group and an intervention group in which the intervention was implemented in different school years. The units of analysis were the cohorts within schools during the time period from 2014–2015 to 2019–2020, with six cohorts per school. School achievement was assessed in first and second grade, in the middle and at the end of each school year, and in third grade in the middle of the school year. For each of these measurement occasions, a separate CITS model was tested.

### Participants

Three hundred eighteen schools, clustered in 26 school boards and 4 geographical locations, were asked to participate in the study. Schools were located in 5 (out of 12) provinces in the west and middle parts of the Netherlands. Half of the schools were located in villages, and half of the schools were in cities (mostly large cities, e.g., Amsterdam, Rotterdam, Utrecht).

Of the 318 schools, 55 schools from two school boards participated in a larger research project on the intervention *Build!* (van der Weijden et al., 2024). The 318 schools had a student population that was representative of the national population according to the school weight of the schools, a composite measure of the socioeconomic status and ethnic composition of the children of a school determined by the educational level, ethnicity, and financial means of the parents. A higher school weight implies a more complex student population. School weights range from 20 to 40. Schools with a higher school weight are allotted extra funds by the government. The average school weight of the 318 schools was 29.48 ( $SD=4.92$ ), which was similar to that of the national population ( $M=29.84$ ,  $SD=3.91$ ),  $t(6568) = 1.51$ ,  $p = .131$ , and the 55 schools that participated in the larger research project ( $M=28.76$ ,  $SD=2.72$ ),  $t(337) = 1.02$ ,  $p = .301$ . The average school size of the 318 schools ( $M=249.52$  students,  $SD=139.28$ ) was somewhat larger than the average school size of the 55 schools ( $M=198.23$  students,  $SD=104.19$ ),  $t(359) = 2.56$ ,  $p = .011$ .

Permission was obtained from 233 schools to retrieve the test scores of reading fluency, spelling, and reading comprehension from ParnasSys, a student information system in which Dutch schools can register scores on the various measures of school achievement. There were 199 schools that also gave permission to retrieve test scores

of mathematics. Data came from the cohorts of children who were in grades 1 to 3 from 2014–2015 to 2019–2020. Thus, per school and grade, the data of six successive cohorts were retrieved. Data were obtained and treated in accordance with the guidelines of the Ethics Review Board of the Faculty Social and Behavioral Sciences of the University of Amsterdam (approval obtained with project number 2021-CDE-12989).

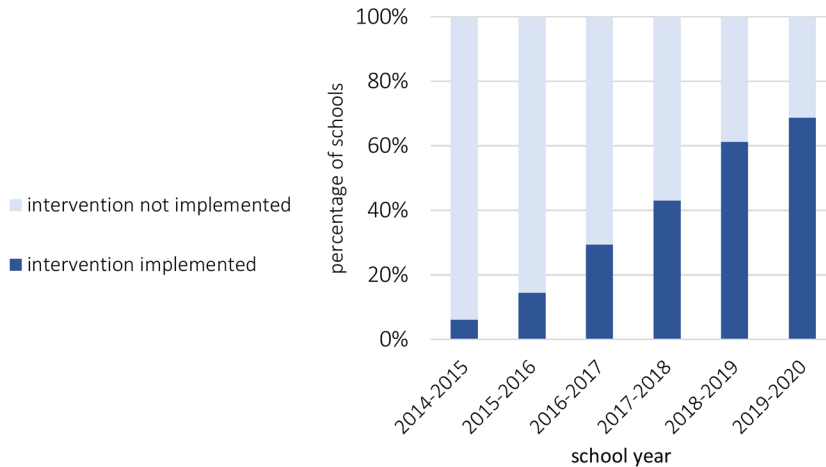
Schools were asked to indicate whether they had implemented the intervention *Build!* and, if so, when they had started, to determine the cohorts for which *Build!* had or had not been implemented. Twelve schools did not provide this information. Working with *Build!* was coded at the cohort level. We do not know which specific children did or did not work with the program. A number of schools had not registered all school achievement measures in the ParnasSys system. As a result, the number of schools per outcome measure varied slightly, that is, 213 schools were included for reading, 214 for spelling, 215 for reading comprehension, and 184 for mathematics. The average school weights of the final samples varied from 29.70 to 29.80 (*SDs* varied from 4.65 to 4.67), which were similar to the average school weight of the 318 approached schools ( $M=29.48$ ,  $SD=4.92$ ). The average school sizes of the final samples, varying from 253.60 to 259.59 (*SDs* varied from 142.62 to 145.71), were also similar to the average school size of the 318 approached schools ( $M=249.52$ ,  $SD=139.28$ ).

For unknown reasons, the data of the first 2 to 3 years (2014–2015 to 2016–2017) could no longer be retrieved for a number of schools. As a result, the number of schools also varied across school years: between 142 and 207 for reading fluency, between 150 and 203 for spelling, and between 115 to 206 for reading comprehension. The low number of 115 referred to reading comprehension at the end of grade 1: Around half of the schools started to test reading comprehension from grade 2 onward. On all other occasions, the number of schools for reading comprehension varied between 143 and 206. Data on mathematics were almost complete. The number of schools varied between 169 and 177 across school years.

Schools started to implement the intervention at different points in time from the school year 2014–2015 until school year 2019–2020. The percentage of schools that had implemented the intervention over these years is displayed in [Figure 2](#).

For each measurement occasion and for each outcome measure, groups of intervention and control schools were determined. The intervention schools had at least one cohort that had been involved in the intervention, whereas the control schools did not have any cohort that had worked with the intervention. As a result, the number of schools that were included in the intervention and control group differed across measurement occasions. For example, a school that started to use the intervention in 2019 would only have cohorts of the middle and end of grade 1 who had worked with the intervention and would thus be included in the control group for the measurement occasions in second and third grade. Therefore, at later measurement occasions, control groups were somewhat larger and intervention groups somewhat smaller compared to the earlier measurement occasions.

Depending on the outcome and measurement occasion, the intervention group consisted of 72 to 145 schools (reading fluency: 72–144; spelling: 89–145; reading comprehension: 83–113; mathematics: 75–123) and the control group of 61 to 126 schools (reading fluency: 62–118; spelling: 63–120; reading comprehension: 85–126;



**Figure 2.** Percentage of schools implementing *Build!* over time.

mathematics: 61–107). In the control group, there were always 48 to 58 schools that did not implement the intervention at all. All schools of this part of the control group and the majority of schools in the intervention group came from the same three geographical regions. This similarity in geographical location across groups increases the comparability of control and intervention groups (Cook et al., 2008), being subject to the same district policies, demographic shifts, and contextual factors. Comparability of the control and intervention group supports the likelihood that in the absence of the intervention, the treatment group would have made the same average gains (or losses) as the comparison group (parallel trends assumption; Jacob et al., 2016).

In principle, data from six cohorts of children should have been available for each school on each measurement occasion. However, this was not the case. Assessments at the end of school year 2019–2020 could not be used, because they were conducted right after the first COVID-19 pandemic school closure. Thus, there were six cohorts for assessments in the middle of each grade and five cohorts for assessments at the end of each grade.

### **Intervention**

The intervention program *Build!* consists of 526 digital lessons, divided in 12 program parts. In parts 1 to 5, children learn the sounds of 14 letters and digraphs in the Dutch language (e.g., /o/ in *sok* [sock] and /oo/ in *boot* [boat]). Children also learn to decode regular one-syllable words with these letters. In parts 6 to 9, children are introduced to words including regular consonant clusters (e.g., *glas* [glass] and *warm* [warm]) and common irregular consonant clusters (e.g., /sch/ in *schoen* [shoe] and /ng/ in *zing* [sing]). Part 10 continues with compound words (e.g., *maandag* [Monday] and *zeezout* [sea salt]). In parts 11 and 12, children learn to read two-syllable words with open syllables (e.g., *letter* [letter] and *rozen* [roses]) and closed syllables (e.g., *winter* [winter]). From part 2 onward, there are reading exercises with and without time limits.

The instruction is characterized by (a) direct instruction (Stockard et al., 2018), (b) direct feedback (Wisniewski et al., 2019), and (c) the minimal pairing technique (McCandliss et al., 2003). This technique requires children to start with a word, change one letter at a time, and read the resulting words. Intervention sessions were guided by a tutor. The tutor could be a professional (e.g., a teacher) or a nonprofessional (e.g., a parent, volunteer, or older student). Previous research has shown that the majority of professional and nonprofessional tutors provided sufficient support (Zijlstra et al., 2014). More information on the program is provided by Regtvoort et al. (2007, 2013) and Zijlstra et al. (2014, 2021).

In this study, schools implemented the intervention. If they wanted, they could be supported by a 2-day training provided by the publisher of *Build!*. In this training, schools were informed how *Build!* is intended to be used, how to find tutors, how to organize intervention sessions at school, and how to monitor children's practice. It was recommended to start the intervention in the middle of the second kindergarten year (in the Netherlands children go to school when they are 4 years old and follow 2 years of kindergarten before entering grade 1) and to finish the intervention in the middle of second grade. It was also recommended to provide children with three to four intervention sessions a week lasting 10 to 15 minutes each.

## **Measures**

Within schools, we assessed per cohort whether *Build!* was used and, if so, for how long. We also determined the performance of each cohort on a range of outcomes (word reading fluency, spelling, reading comprehension, and mathematics).

### **Information on the Use of *Build!***

We asked schools whether they (had) used the intervention and, if so, to fill out a table to indicate in which school years (between 2014–2015 and 2019–2020) and in which grades (from kindergarten through grade 3). Based on this information, we created one variable at the school level and two variables at the cohort level. At the school level, we coded whether the school had ever used *Build!* in this period in a particular grade. At the cohort level, we distinguished (a) whether the intervention was used in the particular cohort and (b) the number of years the school had already been using the intervention.

**Implementation of the Intervention.** Cohorts in which *Build!* was implemented from kindergarten or grade 1 onward were coded “1.” All other cohorts were coded “0.” Note that we had no information on which children did or did not participate in the intervention but only on whether the school offered *Build!* to this cohort or not.

**Number of Years That *Build!* Is Used.** Within schools, we determined per cohort how many previous cohorts at the school had worked with *Build!*. Cohorts in which *Build!* was implemented by the school for the first time were coded as “0.” The score increased if the school had used *Build!* in additional previous years.

### **Outcome Measures**

Within schools, we determined per cohort, outcome, and measurement occasion (a) the percentage of children with difficulties (i.e., children scoring below the 25th percentile based on national norms) and (b) the mean ability. Both the percentage of children with difficulties and the mean ability were based on the individual test scores of the children in a cohort.

In the Netherlands, primary schools are obliged to monitor children's school performance with (national) standardized tests. As most schools in the Netherlands (85%) use the tests of Central Institute for Test Development (Cito), those tests were used in this study. Cito had designed the tests in such a way that all school staff members could administer the test by following the instructions. The tests are specifically designed to be able to measure children's growth. Items on different tests within a certain domain (e.g., the items of the spelling tests in grade 1, grade 2, and grade 3) are all positioned on the same underlying scale so that raw scores could be converted to an ability score, a measure of the child's ability across grades. Tests were evaluated by the Dutch Committee on Tests and Testing; Reliability and validity were sufficient for all tests (Egberink et al., 2009–2023). Schools administered the tests twice a year: between mid-January and mid-February and between mid-May and the end of June. Reading comprehension was measured from the end of grade 1 onward. Thus, we obtained test scores of word reading fluency, spelling, and mathematics from school years 2014–2015 to 2019–2020 from children in grades 1 through 3 at the middle and end of the school year.

**Word Reading Fluency.** This ability was measured with the *Three-Minute Test* (Cito, 2017; Krom et al., 2010). The test consisted of three cards of 150 words each. Word difficulty increased per card, from one-syllable words with a CVC structure (card 1; e.g., *kat* [cat]) to one-syllable words with consonant clusters (Card 2; e.g., *melk* [milk]), and words with two to four syllables (Card 3; e.g., *mentaliteit* [mentality]). The test was administered individually. Per card, children were asked to read aloud as many words as possible within 1 minute, without making errors.

In the middle of grade 1, only cards 1 and 2 were administered. At the end of grade 1 and in grade 2, all three cards were used. From grade 3 onward, card 1 is skipped for high-performing children. The score per card was the number of words read correctly. The scores on the administered cards were summed and converted into an ability score (Cito, 2017; Krom et al., 2010). All ability scores are on the same scale. Depending on the administered cards and grade, Cronbach's alpha varied between .92 and .97 (Krom et al., 2010) and test-retest reliability between .90 and .97 (van Til et al., 2018). National norms are available for the ability scores per grade and measurement occasion. The levels vary from A (>75th percentile), B (51–75th percentile), C (26–50th percentile), D (11–25th percentile), to E (≤10th percentile). We recoded levels D and E as “difficulties in reading fluency” and levels A, B, and C as “no difficulties in reading fluency.”

Between 2015 and 2021, two different versions of this test were in use: version 2010 (Krom et al., 2010) and version 2017 (Cito, 2017). The words presented on the cards differed between versions. A dummy was created to account for test version.

**Spelling.** Spelling was measured with *Cito Spelling* (Cito, 2014b; de Wijs, 2010). This consisted of two parts, administered at the same day, with a break or activity in between. In each task, children had to spell 20 to 30 words. Target words differed across age groups, corresponding to the national learning goals, varying from consistent one-syllable words (e.g., *man* [man]) to inconsistent four-syllable words (e.g., *informatie* [information]). The test was administered by the teacher in a classroom setting. Teachers read aloud a sentence containing the target word and asked children to spell the target word on an answer sheet. There was no time limit. The teacher continued to the next sentence when all children had finished spelling the word. The score was the number of words that were spelled correctly. Depending on the grade, test-retest reliability varied between .86 and .93 (Tomesen et al., 2015a, 2015b, 2016b). Similarly to word reading fluency, the raw scores in all grades were converted to an ability score. Based on national norms, these scores were recoded into “difficulties in spelling” and “no difficulties in spelling.”

Between 2015 and 2021, two different versions of this test were in use: version LOVS (de Wijs, 2010) and version 3.0 (Cito, 2014b). In the 3.0 version, the second part was similar for all children, whereas in the LOVS version, children receive an easier or more difficult part 2 depending on the child’s performance on the first part. In grades 2 and 3, the more difficult version of the second part did not involve spelling words, but included multiple-choice items in which children had to select the one word (out of four) that was spelled incorrectly. A dummy was created to account for test version.

**Reading Comprehension.** Reading comprehension was measured with *Cito Begrijpend Lezen* (Cito, 2014a; Feenstra et al., 2010). The test consisted of two parts, administered at the same day, with a break or activity in between. Each part consisted of a booklet containing around eight texts and 20 to 25 multiple-choice questions, displayed right after the corresponding text. Texts were stories, news items, articles, reviews, advertisements, announcements, poems, requests, game manuals, recommendations, recipes, songs, reports, instructions, or letters. In the higher grades, texts were longer and more formal than in the lower grades. The questions tested children’s understanding and interpretation of the texts, that is, whether they could process information that was explicitly mentioned in the texts (e.g., questions on a number, fact, or opinion in the text; questions on the relations between sentences) and whether they could make connections between the texts and their own knowledge (e.g., questions on the meaning of a word, questions on the main idea of the text). The test was administered by the teacher in a classroom setting. Children received their own booklet and had to read the texts and answer the corresponding questions by circling their answers in the booklet. The score was the number of questions answered correctly. Depending on the grade, test-retest reliability varied between .86 and .93 (Jolink et al., 2015; Tomesen et al., 2016a). Similarly to word reading fluency, the raw scores in all grades were converted to an ability score. Based on national norms, these scores were recoded into “difficulties in reading comprehension” and “no difficulties in reading comprehension.”

Between 2015 and 2021, two different versions of this test were in use: version LOVS (Feenstra et al., 2010) and version 3.0 (Cito, 2014a). In the LOVS version, the difficulty of the second part depended on the child’s performance on the first part,

whereas in the 3.0 version, the second part was similar for all children. A dummy was created to account for test version.

**Mathematics.** Mathematics was measured with *Cito Rekenen-Wiskunde* (Cito, 2013; Janssen et al., 2010). The test consisted of two or three parts, administered at the same day, with a break or activity in between. Each task consisted of 26 to 32 arithmetical questions, including both equations and word problems. In grades 1 to 3, the questions covered number estimation, arithmetic, measurement, geometry, time, and money. Difficulty increased per grade and matched the national learning goals. There were both multiple-choice and short-answer questions. The test was administered in the classroom by the teacher. In grades 1 and 2, teachers read aloud the questions (once repeated), children received a booklet with the corresponding equations and pictures, and they wrote down their answer in the booklet. In grade 3, children worked independently: They read the questions themselves and wrote down the answers on a separate answer sheet. There was no time limit: In Grades 1 and 2, the teacher only continued when all children had finished the question. The score was the number of questions answered correctly. Depending on the grade, test–retest reliability varied between .92 and .95 (Hop et al., 2016; Janssen et al., 2015a, 2015b). Similarly to word reading fluency, the raw scores in all grades were converted to an ability score. Based on national norms, these scores were recoded into “difficulties in mathematics” and “no difficulties in mathematics.”

Between 2015 and 2021, two different versions of this test were in use: version LOVS (Janssen et al., 2010) and version 3.0 (Cito, 2013). In the 3.0 version, children were allowed to write down their calculations during the test, whereas in the LOVS version they were not. A dummy was created to account for test version.

## Analyses

### Data Processing

From all participating schools, we had scores from successive school years (2014–2015 to 2019–2020) from cohorts in grades 1 to 3, with two measurement occasions per grade (middle and end). For each outcome, we analyzed the data per grade and measurement occasion. Data were not analyzed for the end of grade 3 because there were fewer cohorts available at the end of each grade due to the COVID-19 pandemic (the second measurement occasion in school year 2019–2020 was excluded; see Participants section). In grade 3, that was particularly problematic because in this grade few cohorts had received *Build!* from kindergarten or grade 1 onward, while the schools had been using *Build!* for 2 years or longer (needed to estimate the post-intervention time trend; see DiD Models section).

### DiD Models

Our main interest concerned time trends, which are changes in reading fluency, spelling, reading comprehension, and mathematics over time. In particular, we expected a change in time trend in literacy outcomes (reading fluency, spelling, and reading comprehension) from the moment the intervention was implemented onward. For mathematics, we did not foresee such a change in time trend.

We used a DiD model to estimate changes in outcomes over time (Mascha & Sessler, 2019). The model is presented in Figure 1. Note that schools did not start the intervention at the same time; therefore, the length of the pre- and post-intervention period differs across schools that have implemented *Build!*. We modeled a time trend in the control group and, separately, a pre-intervention time trend in the intervention group. For the intervention group, we also modeled a change in the intercept at the intervention start and a post-intervention time trend. The latter parameter estimated the effect of the years since implementation.

Based on the eighth formula from Mascha and Sessler (2019), we used the following model to estimate all the parameters of interest:

$$\begin{aligned}
 y_t = & \beta_0 + \beta_1 \text{test version} + \beta_2 \text{intervention group} \\
 & + \beta_3 \text{school year} + \beta_4 \text{intervention group} * \text{school year} \\
 & + \beta_5 \text{implementation of the intervention} \\
 & + \beta_6 \text{number of years that Build! is used},
 \end{aligned}$$

in which  $\beta_0$  is the intercept for schools in the control group,  $\beta_2$  is the difference in intercepts for schools in the intervention group compared to those in the control group,  $\beta_3$  is the general time trend in the control group,  $\beta_4$  is the difference in pre-intervention time trend in the intervention group compared to the control group,  $\beta_5$  is the intercept change in the first intervention year compared to the predicted level based on the intercept and pre-intervention time trend, and  $\beta_6$  is the difference in the post-intervention time trend in the intervention group from the second intervention year onward (measuring the effect of the number of years that *Build!* is used) compared to the pre-intervention time trend.

Unfortunately, during our period of interest new tests were introduced for all outcome variables. To control for differences between test versions, we included a dummy in the model,  $\beta_1$ . However, the newer test versions were introduced when more schools had implemented the intervention. Even by including test version, it might be possible that the model could not make a clear distinction between the introduction of the intervention and introduction of the new test version, leading to overestimation or underestimation of intervention effects. Therefore, we additionally ran all models on the data of the separate test versions. The new test version was used for spelling, reading comprehension, and mathematics because this version was most frequently used (i.e., in 65% to 90% of the cohorts). Both test versions were analyzed for reading fluency because the oldest version was used as often as the newest version (i.e., in 42% to 74% of the cohorts).

The models were built using multilevel modeling, in which cohorts were nested within schools. This way, we accounted for dependencies between cohorts within the same school (Snijders & Bosker, 2012). We included random intercepts for schools, as schools can vary in general ability level. Schools can also differ in general time trend. We decided per outcome whether it was necessary to include random slopes for time. If random slopes models fit the data better for one of the measurement occasions (e.g., the middle of grade 1), we added random slopes for all five occasions. Using the full maximum likelihood estimator, all schools were included in the analysis

also when some cohorts were missing. Analyses were carried out with *R* (R Core Team, 2022), using package: nlme (Pinheiro et al., 2019).

**Model Evaluation.** The first evaluation looked at whether there was a change in test scores in the first year *Build!* was introduced, that is, whether  $\beta_5$  was significant. Second, we evaluated whether cohorts benefited more from the intervention if schools had been using the intervention for a longer time, that is, whether  $\beta_6$  was significant. These effects were evaluated with *t* tests, using a significance level of 5%. To estimate its effect size, we calculated the difference in  $R^2$  (explained variance at level 1 and 2 combined; Snijders & Bosker, 2012) between the model with and without parameter  $\beta_6$ . Cohen (2013) suggested that an  $R^2$  of .01 can be considered small, an  $R^2$  of .09 can be considered medium, and an  $R^2$  of .25 can be considered large.

**Assumptions.** We checked assumptions for all models. The assumptions of linearity and homogeneity of residuals were checked by visually inspecting the plots of level-1 and level-2 residuals (y-axis) and predicted outcomes (x-axis). Normality of residuals was checked by visually inspecting the Q-Q plots and histograms of residuals at levels 1 and 2. Assumptions were met, although homogeneity was not perfect due to the two different test versions in use. Therefore, we not only included test version as a predictor but also allowed the models to estimate the variance for each test version separately. In addition, we ran the models for one test version, using only the data of the test version that was used most often during the research project (either the old or new version, depending on the outcome).

## Results

### Data Processing

Prior to analyses, we checked for outliers on all outcome variables (values more than 3 standard deviations from the mean). The number of outliers varied from 0.9% to 3.9% (reading fluency: 1.0%–3.9%; spelling: 1.1%–3.7%; reading comprehension: 0.9%–2.8%; mathematics: 1.5%–3.2%). Outliers were coded as missing values. Those observations were not included in the analyses.

There was a specific problem with reading fluency. The test needed to be administered individually. It appeared that some schools therefore tested only the students with reading difficulties from grade 3 onward. As a result, some cohorts had more than 80% readers with difficulties, which is unrealistic. Based on the spelling test, which was administered in a classroom setting, we determined the size of each cohort. When the number of children tested on spelling was similar to the number of children tested on reading fluency (by a 20% margin), we included the cohort in the analyses of reading fluency. As a consequence, the analyses in the middle of grade 3 included fewer schools. Across school years, the number of schools varied between 136 and 195 at the middle of each grade (except for reading fluency in the middle of grade 3: 41–160 schools) and between 147 and 195 at the end of each grade.

### Descriptive Statistics

Means and standard deviations of the percentage of children with difficulties at the end of the study (school year 2019–2020 at the middle of each grade and school year

2018–2019 at the end of each grade) are shown in Table 1 per outcome and measurement occasion and split into schools that did not use *Build!*, used *Build!* for 1 or 2 years, and used *Build!* for 3 or more years. Effect sizes of group differences are presented in the last two columns. The results for the mean ability on the newest test version are shown in Table 2.

Regarding reading fluency, schools that had used *Build!* for 1 or 2 years had a similar percentage of children with difficulties and a similar cohort mean ability as schools without *Build!* at the end of the period. In contrast, schools that had used *Build!* for 3 or more years showed fewer children with difficulties and a higher mean ability than schools that did not use *Build!* (small differences). Regarding spelling, reading comprehension, and mathematics, schools without *Build!* mostly showed more children with difficulties and a lower mean ability than schools that used *Build!* for 1 or 2 years (small difference) and schools that used *Build!* for 3 or more years (medium difference). These descriptive statistics give the impression that *Build!* may also have affected mathematics, but Tables 1 and 2 do not show whether the differences between groups already existed at the beginning of the study (before the intervention had started). Therefore, Tables 1 and 2 do not show whether any differences

**Table 1.** Descriptive statistics for the percentage of children with difficulties at the end of the study.

Outcome	Schools without <i>Build!</i>			Schools with <i>Build!</i>						Cohen's <i>d</i>	
	<i>n</i>	<i>M</i>	<i>SD</i>	1–2 years of <i>Build!</i>			≥3 years of <i>Build!</i>			Without ↔ 1–2 years <sup>a</sup>	Without ↔ ≥3 years <sup>b</sup>
Reading fluency											
Mid-grade 1 <sup>c</sup>	43	37.87	13.19	43	34.30	12.31	69	29.83	14.20	–0.28	–0.58
End of grade 1 <sup>c</sup>	54	21.18	13.48	52	20.20	13.14	49	15.99	10.13	–0.07	–0.43
Mid-grade 2 <sup>c</sup>	62	29.06	13.49	52	28.62	11.38	53	25.74	10.60	–0.03	–0.27
End of grade 2 <sup>c</sup>	70	21.25	10.92	42	19.23	13.65	21	16.86	8.49	–0.17	–0.42
Mid-grade 3 <sup>c</sup>	79	24.03	13.43	40	21.69	12.11	19	22.02	12.98	–0.18	–0.15
Spelling											
Mid-grade 1 <sup>c</sup>	54	32.00	16.30	50	29.64	17.39	82	23.39	16.30	–0.14	–0.53
End of grade 1 <sup>c</sup>	76	25.67	15.34	59	24.46	17.10	59	14.76	11.83	–0.08	–0.71
Mid-grade 2 <sup>c</sup>	77	26.83	14.83	64	22.86	15.84	58	18.28	14.44	–0.26	–0.58
End of grade 2 <sup>c</sup>	113	24.67	15.65	55	18.54	12.91	29	18.56	16.87	–0.41	–0.38
Mid-grade 3 <sup>c</sup>	112	32.16	16.44	56	25.72	16.18	29	22.09	13.99	–0.39	–0.63
Reading comprehension											
End of grade 1 <sup>c</sup>	58	31.83	23.11	43	23.78	16.95	30	13.77	11.83	–0.39	–0.90
Mid-grade 2 <sup>c</sup>	69	34.18	20.61	53	28.88	17.72	48	24.54	18.00	–0.27	–0.49
End of grade 2 <sup>c</sup>	110	30.00	18.75	51	20.65	13.64	29	18.58	12.69	–0.54	–0.65
Mid-grade 3 <sup>c</sup>	115	32.27	18.85	57	29.01	16.49	28	22.22	17.75	–0.18	–0.54
Mathematics <sup>a</sup>											
Mid-grade 1 <sup>c</sup>	56	34.06	19.71	46	30.37	18.66	71	24.45	16.02	–0.19	–0.54
End of grade 1 <sup>c</sup>	75	29.43	18.13	48	22.60	14.01	51	15.72	10.86	–0.41	–0.88
Mid-grade 2 <sup>c</sup>	73	32.34	17.59	48	27.63	15.98	50	21.81	13.02	–0.28	–0.66
End of grade 2 <sup>c</sup>	97	27.32	16.29	46	21.65	16.41	28	17.29	13.25	–0.35	–0.64
Mid-grade 3 <sup>c</sup>	100	31.58	15.80	43	25.77	16.73	27	16.58	12.93	–0.36	–0.98

<sup>a</sup>Cohen's *d*: difference between schools without *Build!* and schools that used *Build!* for 1 or 2 years, divided by the pooled *SD*.

<sup>b</sup>Difference between schools without *Build!* and schools that used *Build!* for 3 or more years, divided by the pooled *SD*. A negative Cohen's *d* means that the percentage of children with difficulties was lower in schools with *Build!* than in schools without *Build!*

<sup>c</sup>This table presents descriptives for the end of the study on the new test version. That is school year 2019–2020 for the middle of each grade and 2018–2019 for the end each grade (as the COVID-19 pandemic affected scores at the end of 2019–2020).

**Table 2.** Descriptives for the mean ability at the end of the study.

Outcome	Schools without <i>Build!</i>			Schools with <i>Build!</i>						Cohen's <i>d</i>	
	<i>n</i>	<i>M</i>	<i>SD</i>	1–2 years of <i>Build!</i>			≥3 years of <i>Build!</i>			Without ↔ 1–2 years <sup>a</sup>	Without ↔ ≥ 3 years <sup>b</sup>
				<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>		
<b>Reading fluency</b>											
Mid-grade 1 <sup>c</sup>	47	14.52	3.52	45	15.26	3.06	74	15.72	3.95	0.22	0.32
End of grade 1 <sup>c</sup>	55	29.09	6.52	53	28.95	5.63	47	30.62	5.40	–0.02	0.25
Mid-grade 2 <sup>c</sup>	62	44.20	7.04	52	43.73	5.00	52	45.38	5.77	–0.08	0.18
End of grade 2 <sup>c</sup>	70	53.21	5.89	41	54.06	5.73	21	56.08	5.35	0.15	0.50
Mid-grade 3 <sup>c</sup>	78	63.14	7.50	40	64.07	5.55	19	66.19	47.85	0.13	0.14
<b>Spelling</b>											
Mid-grade 1 <sup>c</sup>	55	144.06	26.45	55	146.79	28.82	83	158.13	25.25	0.10	0.55
End of grade 1 <sup>c</sup>	78	200.48	20.82	78	205.07	21.56	65	218.10	19.06	0.22	0.88
Mid-grade 2 <sup>c</sup>	76	237.11	19.31	64	240.63	18.21	57	250.58	20.71	0.19	0.68
End of grade 2 <sup>c</sup>	112	269.54	18.03	55	277.25	17.81	29	279.46	19.48	0.43	0.54
Mid-grade 3 <sup>c</sup>	112	290.20	16.78	55	298.06	17.74	29	301.60	14.25	0.46	0.70
<b>Reading comprehension</b>											
End of grade 1 <sup>c</sup>	56	113.50	56.00	43	119.76	43.00	30	124.60	30.00	0.12	0.23
Mid-grade 2 <sup>c</sup>	69	128.94	14.78	52	132.31	12.71	49	135.42	13.21	0.24	0.46
End of grade 2 <sup>c</sup>	110	136.94	13.05	51	141.28	11.40	29	145.77	10.86	0.35	0.70
Mid-grade 3 <sup>c</sup>	113	150.98	11.79	56	153.26	11.00	27	159.01	9.20	0.20	0.71
<b>Mathematics<sup>a</sup></b>											
Mid-grade 1 <sup>c</sup>	54	109.70	16.86	45	114.25	16.04	73	117.52	15.08	0.28	0.49
End of grade 1 <sup>c</sup>	73	137.32	13.10	48	142.73	11.53	53	146.97	8.70	0.43	0.84
Mid-grade 2 <sup>c</sup>	73	160.10	13.16	48	161.24	11.11	52	166.46	11.82	0.09	0.50
End of grade 2 <sup>c</sup>	97	182.30	13.44	45	186.75	10.23	28	191.13	10.84	0.36	0.68
Mid-grade 3 <sup>c</sup>	101	198.66	12.14	44	203.65	12.53	27	211.32	9.50	0.41	1.09

<sup>a</sup>Cohen's *d*: difference between schools without *Build!* and schools that used *Build!* for 1 or 2 years, divided by the pooled *SD*.

<sup>b</sup>Difference between schools without *Build!* and schools that used *Build!* for 3 or more years, divided by the pooled *SD*. A positive Cohen's *d* means that the mean ability was higher in schools with *Build!* than in schools without *Build!*.

<sup>c</sup>This table presents descriptives for the end of the study on the new test version. That is school year 2019–2020 for the middle of each grade and 2018–2019 for the end each grade (as the COVID-19 pandemic affected scores at the end of 2019–2020).

between groups emerged during the study and might be related to the introduction of *Build!*. We need the DiD models (controlling for preexisting differences and test version and modeling time trends within groups) to draw any further conclusions on the significance of these differences and their relation with the use of *Build!*.

### DiD Models

We ran separate DiD models for reading fluency, spelling, reading comprehension, and mathematics for each measurement occasion, for both test versions together as well as for the two test versions separately. Before the intervention started, two effects are relevant: (1) the mean difference between the intervention and control group intercepts ( $\beta_2$ ) and (2) the difference between the pre-intervention time trend of the control group and the intervention group ( $\beta_4$ ). Particularly, the similarity of the trends in the control and intervention group is important for the ability to interpret the effects of the intervention in a CITS design (parallel-trends assumption). The pre-intervention time trend in the control group was always based on five (end of a grade) or six (middle of a grade) cohorts. The number of pre-intervention cohorts in the intervention group varied between one and four at the end of a grade and between



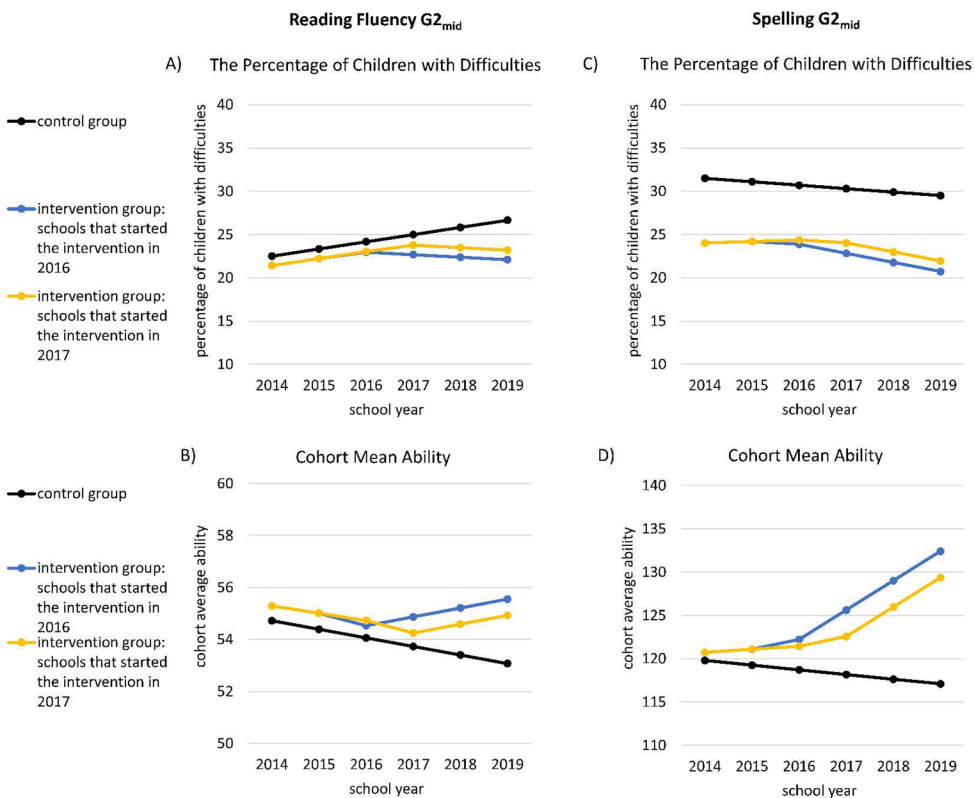
**Table 3.** Post-intervention time trends: changes in test scores after 2 years using *Build!*.

	Percentage of children with difficulties												Mean ability												
	G1 <sub>mid</sub>			G1 <sub>end</sub>			G2 <sub>mid</sub>			G2 <sub>end</sub>			G3 <sub>mid</sub>			G1 <sub>mid</sub>			G2 <sub>mid</sub>			G3 <sub>mid</sub>			
	$\beta$	<i>p</i>	$\beta$	<i>p</i>	$\beta$	<i>p</i>	$\beta$	<i>p</i>	$\beta$	<i>p</i>	$\beta$	<i>p</i>	$\beta$	<i>p</i>	$\beta$	<i>p</i>	$\beta$	<i>p</i>	$\beta$	<i>p</i>	$\beta$	<i>p</i>	$\beta$	<i>p</i>	
Reading fluency																									
TVs together	-1.39	.006	-0.84	.162	-1.10	.040	-1.27	.101	-0.38	.705	0.30	.052	0.70	.011	0.62	.018	1.19	.001	0.83	.093					
Old TV	-2.05	.052	-2.51	.018	-2.44	.032	-2.72	.022	-4.74	.068	0.90	.014	1.68	<.001	1.84	<.001	1.94	<.001	2.04	.093					
New TV	-2.07	.001	-0.75	.324	-1.36	.048	-1.46	.190	-0.51	.656	0.39	.018	0.83	.021	0.99	.008	1.54	.007	1.14	.090					
Spelling																									
TVs together	-1.44	.022	-1.76	.023	-1.22	.091	-1.82	.079	-2.07	.043	1.84	.026	1.75	.056	3.05	<.001	4.26	<.001	3.34	<.001					
New TV	-1.65	.006	-2.06	.010	-1.50	.033	-1.61	.105	-2.27	.021	3.30	<.001	2.95	.005	2.84	.003	3.11	.016	3.19	.005					
Reading																									
comprehension																									
TVs together	-		-2.26	.045	-1.17	.184	-2.17	.040	-2.08	.042	-		1.48	.062	1.11	.071	1.61	.027	0.52	.416					
New TV	-		-3.15	.007	-1.58	.101	-2.38	.037	-2.67	.019	-		2.23	.009	1.30	.054	1.72	.028	1.33	.076					
Mathematics																									
TVs together	-0.08	.907	-1.50	.061	-0.08	.907	0.08	.938	-1.50	.084	-0.04	.942	0.96	.110	0.14	.811	0.97	.228	0.82	.243					
New TV	-0.57	.409	-1.66	.041	-0.06	.931	0.18	.869	-1.67	.100	0.06	.919	0.84	.168	0.17	.765	0.40	.623	1.06	.180					

Note. TV: Test Version. Significant parameters are presented in bold.

one and five for occasions in the middle of a grade. At least four baseline cohorts are needed for a reliable estimation of a baseline trend (Jacob et al., 2016). Across analyses, the percentage of schools in the intervention group with at least four baseline measures varied between approximately 25% and 50% (22–54 schools).

The intervention effects were evaluated based on two post-intervention parameters: (1) the direct effect of *Build!*, that is, after the first cohort had used *Build!* ( $\beta_5$ ), and (2) the change in time trend when *Build!* had been used for 2 or more years ( $\beta_6$ ). Our main results concern the second parameter, that is, the post-intervention time trend. For reasons of clarity, we present the parameter estimates for this parameter in a single table (Table 3). We present them for each test version separately and the two test versions together. In the [Supplemental Materials](#), all parameter estimates of the DiD models are provided (see Tables A1–A4, model 1). We also ran analyses with one of the test versions. The results are also listed in the [Supplementary Material](#) (Tables A5



**Figure 3.** Model-based estimates of difference-in-difference models for reading fluency and spelling. *Note.* 2014 refers to school year 2014–2015, 2015 to school year 2015–2016, etc. Graphs illustrate time trends in the percentage of children with difficulties in reading fluency (A) and spelling (C) and in the cohort mean ability on reading fluency (B) and spelling (D) in the middle of grade 2, the moment that the intervention should be ended. Separate lines are shown for schools that did not use *Build!*, schools that used *Build!* from 2016 onward, and schools that used the program from 2017 onward. Graphs are based on the model parameters in [Supplemental Materials Table A1](#) and [A2](#) (two test versions) and the formula in the Methods section (see Analyses, Difference-in-Difference Models).

and A6). In what follows, we describe patterns per outcome, rather than significant differences per measurement occasion, and summarize findings across test versions.

### **Word Reading Fluency**

**Pre-Intervention Differences.** Except for a few incidental significant differences, the intervention and control group showed a similar percentage of children with difficulties and a similar mean ability on reading fluency before the intervention had started (see [Supplemental Materials Table A1](#), model 1, intercept<sub>int</sub>,  $\beta_2$ ). The pre-intervention time trend in the intervention and control group was similar (see [Supplemental Materials Table A1](#), model 1, time trend pre<sub>int</sub>,  $\beta_4$ ). In both groups, there was a slight increase in children with reading difficulties over time, but this was less clear for the two test versions separately (see [Supplemental Materials Table A1](#), model 1, and [Table A5](#), time trend<sub>control</sub>,  $\beta_3$ ). The mean ability was highly stable over time.

**Intervention Effects.** There was no significant change in the percentage of children with difficulties and the mean ability on reading fluency after the school had been using *Build!* for 1 year (see [Supplemental Materials Table A1](#), model 1, intervention start<sub>int</sub>,  $\beta_5$ ). However, overall there was a post-intervention time trend: The percentage of children with difficulties decreased and the mean ability increased from 2 years of *Build!* onward. Depending on test version, this effect was significant from the middle or end of grade 1 to the middle or end of grade 2 ( $\beta_6$ ; see [Table 3](#)). From the second year, for schools that had been using *Build!*, the percentage of children with difficulties was stable or dropped by 1% per school year, while it increased by 1% per school year before *Build!* had been introduced. Put differently, the percentage of readers with difficulties after 2 years of intervention was 1% to 2% lower than expected based on the pre-intervention slope. The mean ability increased by around 1 point per school year, while it was stable before *Build!* had been implemented. The post-intervention time trend explained 1% to 4% of the variance in reading fluency, which might be qualified as a small effect.

In [Figure 3A and 3B](#), we provide an illustration of time trends in the percentage of children with difficulties and the mean ability in the middle of grade 2, when the intervention should have been finished. Based on the model parameters in [Supplemental Materials Table A1](#) (model 1) and the formula in the Methods section (see Analyses, DiD Models), the graphs illustrate time trends for schools that did not use *Build!*, schools that used *Build!* from 2016 onward, and schools that used the program from 2017 onward. The graphs illustrate that before the intervention had started, the differences between the groups were small and there was no immediate change when *Build!* was used for 1 year. However, after *Build!* was used for 2 years, time trends positively changed in the intervention groups, while the control group continued to have a negative time trend.

### **Spelling**

**Pre-Intervention Differences.** The intervention group had fewer children with difficulties and a higher mean ability on spelling before the intervention had started (see [Supplemental Materials Table A2](#), model 1, intercept<sub>int</sub>,  $\beta_2$ ). This effect is most pronounced

for the new test version (see [Supplemental Materials Table A6](#), intercept<sub>int</sub>,  $\beta_2$ ). Regarding the percentage of children with spelling difficulties, the pre-intervention time trend in the control group and intervention group were similar (see [Supplemental Materials Table A2](#), model 1, time trend pre<sub>int</sub>,  $\beta_4$ ). In both groups, the percentage was stable over time (see [Supplemental Materials Table A2](#), model 1, time trend<sub>control</sub>,  $\beta_3$ ). With respect to the mean spelling ability, pre-intervention trends in the control group and intervention group were not always similar, in particular the trends in second grade on the new version (see [Supplemental Materials Table A6](#), time trend pre<sub>int</sub>,  $\beta_4$ ). Here, the mean ability slightly increased over time in the control group (see [Supplemental Materials Table A6](#), time trend<sub>control</sub>,  $\beta_3$ ), whereas it remained stable in the intervention group (adding time trend<sub>control</sub>,  $\beta_3$ , and time trend pre<sub>int</sub>,  $\beta_4$ ).

**Intervention Effects.** There was no change in the percentage of children with difficulties and the mean ability on spelling when the school had been using *Build!* for 1 year (see [Supplemental Materials Table A2](#), model 1, intervention start<sub>int</sub>,  $\beta_5$ ). After working with *Build!* for 2 years, the time trend changed: The percentage of children with difficulties decreased and the mean ability increased. This was significant for most measurement occasions, mid-grade 1 to mid-grade 3 ( $\beta_6$ , see [Table 3](#)). From the second year *Build!* was used, the percentage of children with difficulties dropped by 1% to 3% per school year and the mean ability increased by around 3 points per school year. The post-intervention time trend explained 1% to 4% of the variance in spelling (a small effect).

See [Figure 3C and 3D](#) for an illustration of the effects at the planned end of the intervention, that is, mid-grade 2. Graphs were created in the same way as in [Figure 3A and 3B](#) but were based on the parameters in the [Supplemental Materials Table A2](#) (model 1). The graphs show that before the intervention had started, there were differences between schools that did not use *Build!* and schools that used *Build!* from 2016 or 2017 onward. Nonetheless, there was a clear change in time trends, not immediately after *Build!* was introduced but after 2 years. The intervention group showed more favorable outcomes at the end of the research period.

### Reading Comprehension

**Pre-Intervention Differences.** The intervention group had fewer children with difficulties and a higher mean ability on reading comprehension before the intervention had started (see [Supplemental Materials Table A3](#), model 1, intercept<sub>int</sub>,  $\beta_2$ ). The pre-intervention time trend in the control group and intervention group was similar (see [Supplemental Materials Table A3](#), model 1, time trend pre<sub>int</sub>,  $\beta_4$ ). In both groups, there was a slight increase in children with difficulties (1% to 2% per school year) and a slight decrease in the mean ability (around 1 point per school year) on reading comprehension over time, but this was less clear for the new test version (see [Supplemental Materials Table A3](#), model 1, and [Table A6](#), time trend<sub>control</sub>,  $\beta_3$ ). In this version, the percentage of children with difficulties and the mean ability were mostly stable.

**Intervention Effects.** There was mostly no change in the percentage of children with difficulties and the mean ability on reading comprehension, when the school had been using *Build!* for 1 year (see [Supplemental Materials Table A3](#), model 1, intervention

start<sub>int</sub>,  $\beta_5$ ). However, after working with *Build!* for 2 years, the post-intervention trend for the percentage of children with difficulties changed. While it increased by 1% or 2% before *Build!* was introduced, it kept stable or decreased by 1% after working with *Build!* for 2 years (post-intervention slope was around 2% lower than the pre-intervention slope). This effect was significant for most measurement occasions, including at the end of grade 1 to the middle of grade 3 ( $\beta_6$ , see Table 3). The post-intervention time trend explained 1% to 2% of the variance in the percentage of children with difficulties in reading comprehension (a small effect). The mean ability mostly did not change after working with *Build!* for 2 years.

### Mathematics

We also checked effects of *Build!* on mathematics. If there is no effect on this unrelated skill, the likelihood increases that the effects on trained skills are due to the intervention and not to another event.

**Pre-Intervention Differences.** The intervention group had fewer children with difficulties and a higher mean ability on mathematics before the intervention had started (see Supplemental Materials Table A4, model 1, intercept<sub>int</sub>,  $\beta_2$ ). The pre-intervention time trend in the control group and intervention group was mostly similar (see Supplemental Materials Table A4, model 1, time trend pre<sub>int</sub>,  $\beta_4$ ). In both groups, the percentage of children with difficulties and the mean ability were highly stable over time (see Supplemental Materials Table A4, time trend<sub>control</sub>,  $\beta_3$ ).

**Intervention Effects.** As expected, there was no change in mathematics scores when the school had been using *Build!* for 1 year (see Supplemental Materials Table A4, model 1, intervention start<sub>int</sub>,  $\beta_5$ ) or after working with *Build!* for 2 years ( $\beta_6$ , see Table 3).

### Additional Analyses on Early and Late Adopters

The overall results thus far show that the effects of the intervention did not occur immediately after schools implemented the intervention, but in the following years of working with the intervention. The implementation of the intervention was associated with a downward trend in literacy problems and an upward trend in mean level of literacy performance. This change in trend was found, irrespective of the version of the test that was used to assess literacy skills (see Tables A5 and A6). However, it should be acknowledged that pre-intervention trends were more affected by schools that started the intervention later (i.e., late adopters) and that post-intervention trends were more affected by schools that started the intervention earlier (i.e., early adopters) because late adopters inevitably had more pre-intervention cohorts and early adopters had more post-intervention cohorts than late adopters. This raises the possibility that the pre- and post-intervention trends could be ascribed to differences between early- and late-adopting schools. For example, children in the early-adopting schools could profit more from the intervention than the children in the late-adopting schools because schools that started early might be more capable and motivated to implement the intervention and thereby reach a higher treatment fidelity, resulting in larger effects soon after implementing the intervention. In the case of such systematic differences

between early and late adopters, we would expect a significant interaction between time of the implementation of the intervention (early or late) and the pre-intervention time trend, the intervention effect after 1 year, and/or the intervention effect after 2 years.

With respect to the pre-intervention trend, it should first be noted that we did not observe a difference between the schools in the intervention and the control group (i.e., schools that did not start with the intervention in the research period). It would be quite surprising if we found significant differences between pre-intervention trends of early and late adopters, whereas at the same time the pre-intervention trends in the intervention and control group hardly ever differed significantly. Nevertheless, we tested for differences between the pre-intervention time trends of early and late adopters within the intervention group. We distinguished three groups of intervention schools: schools with two, three, and four or more pre-intervention cohorts, where schools with more pre-intervention cohorts were later adopters. Depending on the outcome and measurement occasion, there were 5 to 31 schools ( $M=21$ ) with two pre-intervention cohorts, 8 to 28 schools ( $M=22$ ) with three pre-intervention cohorts, and 8 to 54 schools ( $M=34$ ) with four or more pre-intervention cohorts. We specified a model on only the pre-intervention measures of these groups of schools with three parameters of the original model—intercept ( $\beta_0$ ), test version ( $\beta_1$ ), and school year ( $\beta_3$ )—whereby school year represents the pre-intervention time trend. Next, we created two dummies to distinguish schools with two, three, or four or more pre-intervention cohorts, and we added an interaction between school year ( $\beta_3$ ) and the two dummies. Hardly any of the interaction effects appeared to be significant, indicating—as expected—that early- and late-adopting schools mostly had similar pre-intervention time trends.

It is more likely that post-intervention trends, which differ from the trend in the control group, could differ between early and late adopters. Therefore, we more thoroughly checked whether the intervention effects differed across early and late adopters. The number of years that a school had implemented the intervention ranged from 1 to 6, but the post-intervention trend started from the second year of implementation. Thus, in principle, we could split the intervention group into five groups based on the number of years a school worked with *Build!*. However, the number of schools that had implemented the intervention for 4, 5, or 6 years was low. Therefore, to increase the power of our analyses, we distinguished between schools that had implemented the intervention for 2 years (late adopters) and those that had used the program for 3 or more years (early adopters). If the post-intervention trends differed between these groups, we should find a difference between the groups after 1 and/or after 2 years of implementation. We took several steps to test these effects. We started to extend our original model (model 1 in [Tables A1–A5](#)) by splitting the post-intervention trend ( $\beta_5$ ) into two dummy variables: 2 years and 3 years or more after implementation. Note that the effect after 1 year was already in the model. The parameter estimates for part of this model are given in the [Supplementary Materials](#) (see model 2 in the bottom part of [Tables A1–A5](#)). Overall, the results show that the effect after 2 years is smaller than after 3 or more years of intervention, which nicely aligns with the post-intervention trend observed in our earlier model. Note that the model with the dummy variables has one extra parameter as compared to our previous CITS model. Per outcome measure and measurement occasion, we tested the difference between this model and the simpler model with one trend estimate with a chi-square difference test. Hardly any

of the results were significant (see [Supplementary Materials, Table A7](#)), indicating that this model mostly did not fit the data better than the simpler model did.

Next, we made a dichotomous variable distinguishing between early and late adopters. The schools that had worked with the intervention for 3 years or more were considered early adopters, schools that used the intervention for less than 3 years were late adopters. Depending on the outcome and time point, 6% to 29% of the schools (11 to 57 schools) were early adopters. Finally, we inserted two interaction terms in the model: the interaction of adopter group (early or late) with the immediate effect (after 1 year) and the interaction of adopter group with the effect after 2 years. A significant interaction of adopter group with the immediate effect would imply a difference between early and late adopters after 1 year of implementation. More importantly, a significant interaction of adopter group with the effect after 2 years would mean that, on top of the immediate effect, there would be a difference between early and late adopters in the effect after schools worked with *Build!* for 2 years, a clear indication of differences in post-intervention trend between early and late adopters. Note that we did not specify an interaction of adopter group with the dummy denoting 3 or more years of intervention, because the late adopters had only 2 post-intervention measures. In these analyses, we assumed similar intercepts and pre-intervention time trends for early and late adopters. The results showed that the model with the interaction effects hardly ever differed significantly from the model without these effects (see [Supplementary Materials, Table A7](#)), indicating that early- and late-adopting schools mostly showed the same post-intervention effects. Except for the percentage of children with difficulties in reading comprehension in the middle of grade 3, significant interaction effects were always in the opposite direction than expected: Early adopters showed less favorable outcomes than late adopters (i.e., lower mean abilities). Results indicate that it is not likely that the post-intervention time trends in the original models can be ascribed to the fact that there were more post-intervention cohorts in early-adopting schools than in late-adopting schools and, thus, the larger effects as the implementation proceeded were not confounded by adopter group.

## Discussion

In this study, we examined whether the large-scale implementation of an early-literacy intervention (*Build!*) led to an improvement in reading performance and a reduced number of students with reading difficulties at the school level. We investigated whether intervention effects possibly transferred to spelling and reading comprehension and also included mathematics, a skill unrelated to the intervention on which we did not expect an effect. We assessed whether the effects increased when schools had been using the intervention for a longer time. These questions were answered with a DiD model, specifically a CITS model. In a sample of 207 schools and over 6 school years (2014–2015 to 2019–2020), we investigated per outcome variable whether there was a change in the percentage of children with difficulties and the mean ability at the school level, after schools had used the intervention for 1 year and for 2 or more years.

The models showed that levels of reading fluency, spelling, and reading comprehension did not change immediately after the intervention had been implemented but that time trends changed, resulting in differences between schools using and not using

the intervention after 2 or more years. More specifically, after schools used the intervention for at least 2 years, there were decreases in the percentage of children with difficulties in reading fluency (middle/end of grade 1 to middle/end of grade 2), spelling (middle/end of grade 1 to middle/end of grade 3), and reading comprehension (end of grade 1 to middle of grade 3). Moreover, the mean ability on reading fluency (middle of grade 1 to end of grade 2) and spelling (middle of grade 1 to middle of grade 3) started to increase. A contrary trend was observed in schools that did not use the intervention. As expected, no effects were found on mathematics scores.

As this was a quasi-experimental study, it is important to note that we controlled for preexisting differences between schools that did and did not implement *Build!*. Before the intervention was introduced, schools that used *Build!* had a lower percentage of children with difficulties and a higher cohort average performance on spelling, reading comprehension, and mathematics—but not reading fluency—than schools that did not use the intervention. This indicates that schools with *Build!* were initially “better-performing” schools. However, we found hardly any systematic differences between the pre-intervention time trend in the intervention group and the time trend in the control group, suggesting that the percentage of children with difficulties and the mean ability of the children developed similarly across schools before *Build!* was introduced. Using multilevel modeling, we examined changes in trends *within* schools. Thus, it seems unlikely that our findings could be the result of preexisting differences across schools.

### **Effects of *Build!***

Before drawing any conclusions on the effects of *Build!*, we first discuss the size of the effects and alternative explanations for the findings.

#### **Effect Size**

The number of years *Build!* was used explained 1% to 4% of the variance in the percentage of children with difficulties and the mean ability on reading fluency, spelling, and reading comprehension. At first glance, these effects are small. However, larger effects can hardly be expected. First, the target group of *Build!* is small. The mean ability was exhibited by many more children than only the small group of children who received the *Build!* intervention. The intervention is meant for children who are at risk for reading problems, that is, the 25% lowest-scoring children in kindergarten or grade 1, but the effects were examined at the school level, including the performance of all children. Second, the selection of children is never perfect, and thus not all of the readers with the 25% lowest scores might have received the intervention. Especially in kindergarten (before formal reading instruction has started), the selection of children is challenging, as precursors of reading can only partly predict later reading ability (e.g., van Viersen et al., 2018). Early selection inevitably results in underidentification and overidentification of children with reading problems (Fletcher et al., 2021).

Third, *Build!* is an evidence-based intervention, but it might not prevent all children from experiencing reading problems. Some children have *severe* reading problems (Zijlstra et al., 2021). Within the response-to-instruction (RTI) model (Fuchs & Fuchs,

2006), *Build!* could be seen as a tier 2 intervention (individual reading support from a nonprofessional tutor at school), while some children need tier 3 instruction (individual support from a specialist). Suppose that only half of the lowest-scoring readers in the 25% identified were provided with *Build!* and that it was effective for 25% of these children. Then, the percentage of readers with difficulties would drop from 25% to 22%, which is similar to the drop in readers with difficulties in the middle of grade 1 in our study after *Build!* had been used for 4 years—from 25.51% to 22.45%. The cohort average ability on reading fluency would increase, for example, in the middle of grade 1, by around half a point, similar to the increase in our study after *Build!* had been used for 3 years—from 22.83 to 23.38 points. Please note that with more than 1 million children in primary schools in the Netherlands, a reduction of 3% in readers experiencing difficulties would equate to thousands of children for whom severe literacy problems have been prevented.

Fourth, it is unclear the extent to which schools implemented the intervention as intended. Another study on the school-based implementation of *Build!* (van der Weijden et al., 2024) showed that many schools start the intervention in grade 1 instead of in kindergarten. Although there might be good reasons to do so, the intervention can no longer be considered a prevention program in that case. Effects for remediation programs tend to be smaller (Wanzek et al., 2013). Thus, treatment fidelity, that is, the extent to which the intervention was implemented as intended (Gresham et al., 2000), may have affected intervention outcomes.

### ***Alternative Explanations for the Effects of Build!***

It is probable that the small effects of *Build!* on literacy skills were not caused only by the implementation of the intervention. In that case, we would have found smaller effects on the mean ability than on the percentages of children with difficulties, because the intervention is only used for children at risk for reading difficulties. The finding of similar effects on these two outcomes (reading fluency and reading comprehension) or even larger effects on the mean ability than on the percentage of children with difficulties on spelling suggests that the implementation of *Build!* might be associated with a development within schools that affected *all* children. It is possible that the implementation of *Build!* created more attention for reading and spelling within schools or better monitoring of children's reading and spelling skills. When selecting children for the intervention, schools might become aware of how many children have poor pre-literacy skills. This may have acted as an incentive to provide all students with better instruction and to provide extra help to students who do not reach target levels. Similarly, Torgesen (2009) suggested that the implementation of the RTI instructional model in that research might have caused a change in behavior within the schools. Torgesen speculated that teachers and schools may have become more confident in their ability to meet the needs of children as a result of the implementation of the RTI instruction model, leading to lower percentages of readers with difficulties. In all, it seems likely that the implementation of *Build!* by a school has broader effects than increasing the literacy abilities of the children who are provided with the intervention.

It is not likely that the outcomes were due to large educational improvements, such as changes in general policy. First, there were no improvements in reading fluency in schools that did not use the intervention. Second, the absence of effects on

mathematics suggests that the improvements were not caused by something that was not related to literacy, for example, more funds from the government for education in general. Third, the schools in this study started with *Build!* at different time points, making it less likely that the increase in test scores was due to a particular event at a particular time. Fourth, such a large movement can be expected to start before the school had been using *Build!*, while we did not find that reading or spelling improved before *Build!* was used or after it was used for 1 year. Taken together, findings suggest that the observed improvements in literacy skills could (partly) be a direct effect of the intervention by training at-risk children's literacy skills as well as an indirect effect by causing a change within the school that affected literacy instruction for all children.

### **Long-Term Effects**

Few studies have examined long-term effects of early-literacy interventions, that is, around 1 year after the intervention had finished (Suggate, 2010). Our findings show that there was no long-term effect on word reading fluency. Effects disappeared after the end of grade 2, that is, half a year after the intervention should be finished. In contrast, there were long-term effects on spelling and reading comprehension, that is, the middle of grade 3, 1 year after the intervention had finished. The absence of long-term effects on reading fluency could be explained by the sample size in the higher grades: In these grades, fewer cohorts had worked with *Build!* for 2 or more years. This was even more pressing for reading fluency, as we had to filter out schools that tested only the readers with difficulties. Thus, these findings may just have been a matter of power.

Alternatively, a long-term decrease in reading difficulties may not be established by an early-literacy intervention only. Children need continuous support during all the phases of reading development to establish long-term effects. The current intervention lasted 2 years (covering the pre-reading and reading phase in kindergarten and grade 1 and 2), and earlier research showed that the effects of *Build!* were still visible in sixth grade (Zijlstra et al., 2021). However, it is possible that schools in the current study did not continue the intervention for 2 years. A previous study (van der Weijden et al., 2024) showed that many children stopped the intervention in grade 1 instead of in grade 2.

### **Transfer Effects**

The transfer effects to spelling and reading comprehension are in line with earlier studies on *Build!* (Regtvoort et al., 2013; Zijlstra et al., 2021). Regarding spelling, the intervention includes extensive training on letter–sound correspondences, which has been shown to improve spelling in the short and long term (Ehri et al., 2001a; Suggate, 2016). To a lesser extent, the intervention includes phonological training, that is, phoneme blending and reading with the minimal pairing technique. From around the middle of grade 1, children read words with time limits. Both phonological training and reading fluency training have been shown to improve spelling skills (Ehri et al., 2001b; Suggate, 2016). Reading comprehension also could have been improved by reading fluency training (Álvarez-Cañizo et al., 2015; Kim et al., 2010). The simple view of reading suggests that reading fluency is one of the two components that determine reading comprehension (Florit & Cain, 2011).

### **Years Since Implementation**

The most remarkable finding of this study is that effects only emerged after *Build!* had been used for 2 years or longer. To the best of our knowledge, there was no existing (quasi)experimental study to support the idea that interventions become more effective when schools use them for longer periods. Note that many large-scale studies on K–12 interventions do not show any effects (Lortie-Forgues & Inglis, 2019), perhaps because they included only effects immediately after implementation. Our findings suggest that it takes time and effort to implement an intervention properly.

There are generally two ways to interpret a “proper implementation.” It is widely accepted that treatment integrity is key (for a review study, see Durlak & DuPre, 2008). Only when the intervention is sufficiently implemented as intended can the intervention be effective. Harn et al. (2013) have a different view. They suggest that when it comes to studies in the field, higher levels of treatment integrity are not necessarily better. In their view, interventions become more effective when schools make adjustments to the intervention, leading to a better fit with the school context and better fulfillment of the needs of school staff and children. The implementation of an intervention leaves room for many choices. It may take a couple of years of experience with the intervention to figure out which choices fit the local situation best.

### **Limitations and Suggestions for Future Research**

In this study, we used a DiD approach, more specifically a CITS model, to study the effects of an early-literacy intervention. This design was extended with multiple pre- and post-intervention measurement occasions. Moreover, schools did not start the intervention at the same time. Both extensions strengthen the causal interpretation of the observed effects. Nonetheless, we need to point out two limitations. First, the number of measurement occasions after the intervention had been implemented was limited. As a consequence, estimation of the post-intervention time trend was less precise, and we cannot draw conclusions about a further decrease in the percentage of readers with difficulties after 3 years. Second, schools decided themselves when they started the intervention. They were not randomly assigned to those moments. Our additional analyses indicated that schools who started the intervention earlier did not show more favorable intervention outcomes than those that started later. Schools that started early were thus not more capable or motivated to implement the intervention and did not have greater intervention outcomes in the first 2 years after the implementation. That early adopters had more post-intervention cohorts is thus not an alternative explanation for the finding that the intervention became effective after a number of years. This alternative explanation is also less likely, as we controlled for differences between schools at the beginning of the study. We measured changes within schools, and at each time point we studied whether the time trends changed.

Our findings suggest that future large-scale studies on evidence-based (reading) interventions should not stop after 1 year of implementation. Instead, effects might only appear after 2 or 3 years of implementing the intervention. Although large-scale

studies are time-consuming and costly, it would be worthwhile to include multiple cohorts or continue the study for several years. Moreover, it could be investigated how schools' experience with the intervention affects the implementation of the intervention, in terms of treatment integrity, fit within the school context, and fulfillment of the needs of school staff and children.

### **Practical Implications**

Our findings suggest that *Build!* is slightly more effective than business-as-usual education in decreasing literacy problems. Findings indicate that schools that (want to) use an evidence-based (reading) intervention cannot expect an immediate effect after implementing the intervention for 1 year. Instead, schools have to keep on going for multiple years to observe any effects. Effects in the field might not be as large as found in RCTs guided by researchers and conducted on a small scale. Even with an evidence-based intervention, it takes time and effort to mitigate reading and spelling problems in schools.

### **Conclusion**

Our findings indicated that after 2 years of implementation, *Build!* had a small effect on the percentage of children with difficulties in word reading fluency, spelling, and reading comprehension and the mean ability scores in word reading fluency and spelling. Findings suggest that it takes time to reach or increase effects, which is of special interest in light of large-scale studies on reading interventions. It has been suggested that it is very hard to find effects in large-scale implementation studies (Lortie-Forgues & Inglis, 2019; Thomas et al., 2018). Our findings suggest that effects might not be found immediately after implementation but instead after schools have had several years of experience with the intervention.

### **Acknowledgements**

We are grateful to the school boards and schools that participated in the study. We also like to express thanks to Ellen van der Steene, Judith Kuipers, Rosalin van der Hoeven, Cindy Deijle, and Niels de Ruig, who helped to recruit the schools for this study.

### **Open Research Statements**

#### **Study and Analysis Plan Registration**

There is no registration associated with this study.

#### **Data, Code, and Materials Transparency**

The materials, data, and code associated with this study are not publicly available.

#### **Design and Analysis Reporting Guidelines**

Not applicable.

## Transparency Declaration

The lead author (the manuscript's guarantor) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

## Replication Statement

This manuscript reports an original study.

## Disclosure Statement

The authors report there are no competing interests to declare.

## Funding

This work was supported by The Netherlands Initiative for Education Research (NRO) under Grant 40.5.18540.065.

## ORCID

Fae A. van der Weijden  <http://orcid.org/0000-0001-8117-7112>

Bonne J. H. Zijlstra  <http://orcid.org/0000-0001-9924-4387>

Peter F. de Jong  <http://orcid.org/0000-0002-8806-0563>

## References

- Álvarez-Cañizo, M., Suárez-Coalla, P., & Cuetos, F. (2015). The role of reading fluency in children's text comprehension. *Frontiers in Psychology*, 6, 1810. <https://doi.org/10.3389/fpsyg.2015.01810>
- Annie E. Casey Foundation. (2010). *Early warning! Why reading by the end of third grade matters*. <https://files.eric.ed.gov/fulltext/ED509795.pdf>
- Bailey, D., Duncan, G. J., Odgers, C. L., & Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of Research on Educational Effectiveness*, 10(1), 7–39. <https://doi.org/10.1080/19345747.2016.1232459>
- Bear, G. G., Minke, K. M., & Manning, M. A. (2002). Self-concept of students with learning disabilities a meta-analysis. *School Psychology Review*, 31(3), 405–427. <https://doi.org/10.1080/02796015.2002.12086165>
- Cito. (2013). *Handleiding. Cito Volgsysteem. Primair en Speciaal Onderwijs. Rekenen-Wiskunde 3.0. Groep 3 tot en met 8* [Manual. Cito. Primary and Special Education. Arithmetic-Math, version 3.0. Grade 1 to 6]. Cito.
- Cito. (2014a). *Handleiding. Cito Volgsysteem. Primair en Speciaal Onderwijs. Begrijpend Lezen 3.0. Groep 4 tot en met 8* [Manual. Cito. Primary and Special Education. Reading Comprehension, version 3.0. Grade 2 to 6]. Cito.
- Cito. (2014b). *Handleiding. Cito Volgsysteem. Primair en Speciaal Onderwijs. Spelling 3.0. Groep 3 tot en met 8* [Manual. Cito. Primary and Special Education. Spelling, version 3.0. Grade 1 to 6]. Cito.
- Cito. (2017). *Handleiding. Cito Volgsysteem. Primair en Speciaal Onderwijs. DMT (Drie-Minuten-Toets). Groep 3 tot en met 8* [Manual. Cito. Primary and Special Education. Three-Minute-Test. Grade 1 to 6]. Cito.

- Cohen, J. (2013). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. <https://doi.org/10.4324/9780203771587>
- Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, 27(4), 724–750. <https://doi.org/10.1002/pam.20375>
- Daniel, S. S., Walsh, A. K., Goldston, D. B., Arnold, E. M., Reboussin, B. A., & Wood, F. B. (2006). Suicidality, school dropout, and reading problems among adolescents. *Journal of Learning Disabilities*, 39(6), 507–514. <https://doi.org/10.1177/00222194060390060301>
- de Wijs, A., Kamphuis, F., Kleintjes, F., & Tomesen, M. (2010). *Leerling- en onderwijsvolgsysteem. Spelling Groep 3 t/m 5* [Student information system. Spelling Grade 1 to 3]. Cito.
- Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, 41(3-4), 327–350. <https://doi.org/10.1007/s10464-008-9165-0>
- Egberink, I. J. L., Leng, W. E. D., & Vermeulen, C. S. M. (2009–2023). *COTAN Documentatie*. Boom Uitgevers Amsterdam. [www.cotandocumentatie.nl](http://www.cotandocumentatie.nl)
- Ehri, L. C., Nunes, S. R., Stahl, S. A., & Willows, D. M. (2001a). Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis. *Review of Educational Research*, 71(3), 393–447. <https://doi.org/10.3102/00346543071003393>
- Ehri, L. C., Nunes, S. R., Willows, D. M., Schuster, B. V., Yaghoub-Zadeh, Z., & Shanahan, T. (2001b). Phonemic awareness instruction helps children learn to read: Evidence from the National Reading Panel's meta-analysis. *Reading Research Quarterly*, 36(3), 250–287. <https://doi.org/10.1598/RRQ.36.3.2>
- Feenstra, H., Kleintjes, F., Kamphuis, F., & Krom, R. (2010). *Leerling- en onderwijsvolgsysteem. Begrijpend Lezen. Groep 3 t/m 5* [Student information system. Reading Comprehension. Grade 1 to 3]. Cito.
- Ferrer, E., Shaywitz, B. A., Holahan, J. M., Marchione, K. E., Michaels, R., & Shaywitz, S. E. (2015). Achievement gap in reading is present as early as first grade and persists through adolescence. *The Journal of Pediatrics*, 167(5), 1121–1125.e2. <https://doi.org/10.1016/j.jpeds.2015.07.045>
- Fletcher, J. M., Francis, D. J., Foorman, B. R., & Schatschneider, C. (2021). Early detection of dyslexia risk: Development of brief, teacher-administered screens. *Learning Disability Quarterly: journal of the Division for Children with Learning Disabilities*, 44(3), 145–157. <https://doi.org/10.1177/0731948720931870>
- Florit, E., & Cain, K. (2011). The simple view of reading: Is it valid for different types of alphabetic orthographies? *Educational Psychology Review*, 23(4), 553–576. <https://doi.org/10.1007/s10648-011-9175-6>
- Fluss, J., Ziegler, J. C., Warszawski, J., Ducot, B., Richard, G., & Billard, C. (2009). Poor reading in French elementary school: The interplay of cognitive, behavioral, and socioeconomic factors. *Journal of Developmental and Behavioral Pediatrics: JDBP*, 30(3), 206–216. <https://doi.org/10.1097/DBP.0b013e3181a7ed6c>
- Fredriksson, A., & de Oliveira, G. M. (2019). Impact evaluation using difference-in-differences. *RAUSP Management Journal*, 54(4), 519–532. <https://doi.org/10.1108/RAUSP-05-2019-0112>
- Fuchs, D., & Fuchs, L. S. (2006). Introduction to response to intervention: What, why, and how valid is it? *Reading Research Quarterly*, 41(1), 93–99. <https://doi.org/10.1598/RRQ.41.1.4>
- Galuschka, K., Ise, E., Krick, K., & Schulte-Körne, G. (2014). Effectiveness of treatment approaches for children and adolescents with reading disabilities: A meta-analysis of randomized controlled trials. *PloS One*, 9(2), e89900. <https://doi.org/10.1371/journal.pone.0089900>
- Gresham, F. M., MacMillan, D. L., Beebe-Frankenberger, M. E., & Bocian, K. M. (2000). Treatment integrity in learning disabilities intervention research: Do we really know how treatments are implemented? *Learning Disabilities Research and Practice*, 15(4), 198–205. [https://doi.org/10.1207/SLDRP1504\\_4](https://doi.org/10.1207/SLDRP1504_4)
- Harn, B., Parisi, D., & Stoolmiller, M. (2013). Balancing fidelity with flexibility and fit: What do we really know about fidelity of implementation in schools? *Exceptional Children*, 79(2), 181–193. <https://doi.org/10.1177/001440291307900204>

- Hop, M., Janssen, J., & Engelen, R. (2016). *Wetenschappelijke verantwoording Rekenen-Wiskunde 3.0 voor groep 5* [Scientific justification arithmetic-mathematics 3.0 for Grade 3]. Cito.
- Jacob, R., Somers, M. A., Zhu, P., & Bloom, H. (2016). The validity of the comparative interrupted time series design for evaluating the effect of school-level interventions. *Evaluation Review*, 40(3), 167–198. <https://doi.org/10.1177/0193841X16663414>
- Janssen, J., Hop, M., & Wouda, J. (2015a). *Wetenschappelijke verantwoording Rekenen Wiskunde 3.0 voor groep 4* [Scientific justification of arithmetic-mathematics 3.0 for Grade 2]. Cito.
- Janssen, J., Hop, M., Wouda, J., & Hollenberg, J. (2015b). *Wetenschappelijke verantwoording Rekenen-Wiskunde 3.0 voor groep 3* [Scientific justification arithmetic-mathematics 3.0 for Grade 1]. Cito.
- Janssen, J., Verhelst, N., Engelen, R., & Scheltens, F. (2010). *Wetenschappelijke verantwoording van de toetsen LOVS Rekenen-Wiskunde voor groep 3 tot en met 8* [Scientific justification of LOVS tests arithmetic-mathematics for Grade 1 to 6]. Cito.
- Jolink, A., Tomesen, M., Hilte, M., Weekers, A., & Engelen, R. (2015). *Wetenschappelijke verantwoording. Begrijpend lezen 3.0 voor groep 4* [Scientific justification. Reading comprehension 3.0 for Grade 2]. Cito.
- Kim, Y. S., Petscher, Y., Schatschneider, C., & Foorman, B. (2010). Does growth rate in oral reading fluency matter in predicting reading comprehension achievement? *Journal of Educational Psychology*, 102(3), 652–667. <https://doi.org/10.1037/a0019643>
- Krom, R., Jongen, I., Verhelst, N., Kamphuis, F., & Kleintjes, F. (2010). *DMT en AVI. Groep 3 tot en met 8* [Three-Minute-Test and Text Reading. Grade 1 to 6]. Cito.
- Lortie-Forgues, H., & Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educational Researcher*, 48(3), 158–166. <https://doi.org/10.3102/0013189X19832850>
- Lovett, M. W., Frijters, J. C., Wolf, M., Steinbach, K. A., Sevcik, R. A., & Morris, R. D. (2017). Early intervention for children at risk for reading disabilities: The impact of grade at intervention and individual differences on intervention outcomes. *Journal of Educational Psychology*, 109(7), 889–914. <https://doi.org/10.1037/edu0000181>
- Mascha, E. J., & Sessler, D. I. (2019). Segmented regression and difference-in-difference methods: Assessing the impact of systemic changes in health care. *Anesthesia and Analgesia*, 129(2), 618–633. <https://doi.org/10.1213/ANE.0000000000004153>
- Mathes, P. G., Denton, C. A., Fletcher, J. M., Anthony, J. L., Francis, D. J., & Schatschneider, C. (2005). The effects of theoretically different instruction and student characteristics on the skills of struggling readers. *Reading Research Quarterly*, 40(2), 148–182. <https://doi.org/10.1598/RRQ.40.2.2>
- McCandliss, B., Beck, I. L., Sandak, R., & Perfetti, C. (2003). Focusing attention on decoding for children with poor reading skills: Design and preliminary tests of the word building intervention. *Scientific Studies of Reading*, 7(1), 75–104. [https://doi.org/10.1207/S1532799XSSR0701\\_05](https://doi.org/10.1207/S1532799XSSR0701_05)
- Mol, S. E., & Bus, A. G. (2011). To read or not to read: A meta-analysis of print exposure from infancy to early adulthood. *Psychological Bulletin*, 137(2), 267–296. <https://doi.org/10.1037/a0021890>
- Pinheiro, J., Bates, D., DebRoy, S., & Sarkar, D. (2019). *nlme: Linear and nonlinear mixed effects models*. R-Project. <https://cran.r-project.org/package=nlme>
- Prenger, R., Tappel, A. P. M., Poortman, C. L., & Schildkamp, K. (2022). How can educational innovations become sustainable? A review of the empirical literature. *Frontiers in Education*, 7, 970715. <https://doi.org/10.3389/educ.2022.970715>
- R Core Team. (2022). *R: A language and environment for statistical computing*. R foundation for statistical computing. <https://www.R-project.org/>
- Regtvoort, A. G. F. M., & van der Leij, A. (2007). Early intervention with children of dyslexic parents: Effects of computer-based reading instruction at home on literacy acquisition. *Learning and Individual Differences*, 17(1), 35–53. <https://doi.org/10.1016/j.lindif.2007.01.005>
- Regtvoort, A., Zijlstra, H., & van der Leij, A. (2013). The effectiveness of a 2-year supplementary tutor-assisted computerized intervention on the reading development of beginning readers at risk for reading difficulties: A randomized controlled trial. *Dyslexia (Chichester, England)*, 19(4), 256–280. <https://doi.org/10.1002/dys.1465>

- Simmons, D. C., Coyne, M. D., Kwok, O-M., McDonagh, S., Ham, B. A., & Kame'enui, E. J. (2008). Indexing response to intervention: A longitudinal study of reading risk from kindergarten through third grade. *Journal of Learning Disabilities, 41*(2), 158–173. <https://doi.org/10.1177/0022219407313587>
- Sims, S., Anders, J., & Zieger, L. (2022). The internal validity of the school-level comparative interrupted time series design: Evidence from four new within-study comparisons. *Journal of Research on Educational Effectiveness, 15*(4), 876–897. <https://doi.org/10.1080/19345747.2022.2051652>
- Sirinides, P., Gray, A., & May, H. (2018). The impacts of reading recovery at scale: Results from the 4-year i3 external evaluation. *Educational Evaluation and Policy Analysis, 40*(3), 316–335. <https://doi.org/10.3102/0162373718764828>
- Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Sage Publications Ltd.
- Snowling, M. J. (2013). Early identification and interventions for dyslexia: A contemporary view. *Journal of Research in Special Educational Needs: JORSEN, 13*(1), 7–14. <https://doi.org/10.1111/j.1471-3802.2012.01262.x>
- Stockard, J., Wood, T. W., Coughlin, C., & Rasplica Khoury, C. (2018). The effectiveness of direct instruction curricula: A meta-analysis of a half century of research. *Review of Educational Research, 88*(4), 479–507. <https://doi.org/10.3102/0034654317751919>
- Suggate, S. P. (2010). Why what we teach depends on when: Grade and reading intervention modality moderate effect size. *Developmental Psychology, 46*(6), 1556–1579. <https://doi.org/10.1037/a0020612>
- Suggate, S. P. (2016). A meta-analysis of the long-term effects of phonemic awareness, phonics, fluency, and reading comprehension interventions. *Journal of Learning Disabilities, 49*(1), 77–96. <https://doi.org/10.1177/0022219414528540>
- Thomas, J., Cook, T. D., Klein, A., Starkey, P., & DeFlorio, L. (2018). The sequential scale-up of an evidence-based intervention: A case study. *Evaluation Review, 42*(3), 318–357. <https://doi.org/10.1177/0193841X18786818>
- Thompson, C. B., & Panacek, E. A. (2006). Research study designs: Experimental and quasi-experimental. *Air Medical Journal, 25*(6), 242–246. <https://doi.org/10.1016/j.amj.2006.09.001>
- Tomesen, M., Weekers, A., Hilde, M., Jolink, A., & Engelen, R. (2016a). *Wetenschappelijke verantwoording. Begrijpend lezen 3.0 voor groep 5* [Scientific justification. Reading comprehension 3.0 for Grade 3]. Cito.
- Tomesen, M., Wouda, J., & Horsels, L. (2016b). *Wetenschappelijke verantwoording van de LVS-toetsen: Spelling 3.0 groep 5* [Scientific justification of the LVS tests: Spelling 3.0 Grade 3]. Cito.
- Tomesen, M., Wouda, J., Mols, A., & Horsels, L. (2015a). *Wetenschappelijke verantwoording van de LVS-toetsen: Spelling 3.0 groep 3* [Scientific justification of the LVS tests: Spelling 3.0 Grade 1]. Cito.
- Tomesen, M., Wouda, J., Mols, A., & Horsels, L. (2015b). *Wetenschappelijke verantwoording van de LVS-toetsen: Spelling 3.0 groep 4* [Scientific justification of the LVS tests: Spelling 3.0 Grade 2]. Cito.
- Torgesen, J. K. (2009). The response to intervention instructional model: Some outcomes from a large-scale implementation in reading first schools. *Child Development Perspectives, 3*(1), 38–40. <https://doi.org/10.1111/j.1750-8606.2009.00073.x>
- Torgesen, J. K., Alexander, A. W., Wagner, R. K., Rashotte, C. A., Voeller, K. K., & Conway, T. (2001). Intensive remedial instruction for children with severe reading disabilities: Immediate and long-term outcomes from two instructional approaches. *Journal of Learning Disabilities, 34*(1), 33–58, 78. <https://doi.org/10.1177/002221940103400104>
- Torppa, M., Eklund, K., van Bergen, E., & Lyytinen, H. (2015). Late-emerging and resolving dyslexia: A follow-up study from age 3 to 14. *Journal of Abnormal Child Psychology, 43*(7), 1389–1401. <https://doi.org/10.1007/s10802-015-0003-1>
- van der Weijden, F., van den Boer, M., Zijlstra, A. H., & de Jong, P. F. (2024). *A school-based implementation of an early-literacy intervention: Relations among dosage, familial risk, parental education, and reading acquisition*. Department of Child Development and Education, University of Amsterdam.
- van de Werfhorst, H. G. (2019). Early tracking and social inequality in educational attainment: Educational reforms in 21 European countries. *American Journal of Education, 126*(1), 65–99. <https://doi.org/10.1086/705500>

- van Til, A., Kamphuis, F., Keuning, J., Gijssel, M., Vloedraven, J., & de Wijs, A. (2018). *Wetenschappelijke verantwoording LVS-toetsen DMT* [Scientific justification LVS Three-Minute-Tests]. Cito.
- van Viersen, S., de Bree, E. H., Zee, M., Maassen, B., van der Leij, A., & de Jong, P. F. (2018). Pathways into literacy: The role of early oral language abilities and family risk for dyslexia. *Psychological Science*, 29(3), 418–428. <https://doi.org/10.1177/0956797617736886>
- Wanzek, J., & Vaughn, S. (2007). Research-based implications from extensive early reading interventions. *School Psychology Review*, 36(4), 541–561. <https://doi.org/10.1080/02796015.2007.12087917>
- Wanzek, J., Vaughn, S., Scammacca, N. K., Metz, K., Murray, C. S., Roberts, G., & Danielson, L. (2013). Extensive reading interventions for students with reading difficulties after Grade 3. *Review of Educational Research*, 83(2), 163–195. <https://doi.org/10.3102/0034654313477212>
- Wing, C., Simon, K., & Bello-Gomez, R. A. (2018). Designing difference in difference studies: Best practices for public health policy research. *Annual Review of Public Health*, 39(1), 453–469. <https://doi.org/10.1146/annurev-publhealth-040617-013507>
- Wisniewski, B., Zierer, K., & Hattie, J. (2019). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, 10, 3087. <https://doi.org/10.3389/fpsyg.2019.03087>
- Zelege, S. (2004). Self-concepts of students with learning disabilities and their normally achieving peers: A review. *European Journal of Special Needs Education*, 19(2), 145–170. <https://doi.org/10.1080/08856250410001678469>
- Zijlstra, A. H. (2015). *Early grade learning: The role of teacher-child interaction and tutor-assisted intervention* [Doctoral dissertation, University of Amsterdam]. Digital Academic Repository. [https://pure.uva.nl/ws/files/2615928/166609\\_DEF\\_met\\_correctie\\_Proefschrift\\_Early\\_grade\\_learning\\_H\\_Zijlstra\\_compleet.pdf](https://pure.uva.nl/ws/files/2615928/166609_DEF_met_correctie_Proefschrift_Early_grade_learning_H_Zijlstra_compleet.pdf)
- Zijlstra, A. H., Koomen, H. M. Y., Regtvoort, A. G. F. M., & van der Leij, D. A. V. (2014). Effects of quantitative and qualitative treatment fidelity of an individualized computer-supported early reading intervention delivered by non-professional tutors. *Learning and Individual Differences*, 33, 55–62. <https://doi.org/10.1016/j.lindif.2014.04.004>
- Zijlstra, H., van Bergen, E., Regtvoort, A., de Jong, P. F., & van der Leij, A. (2021). Prevention of reading difficulties in children with and without familial risk: Short- and long-term effects of an early intervention. *Journal of Educational Psychology*, 113(2), 248–267. <https://doi.org/10.1037/edu0000489>