Weeks of practice and years of experience

Factors related to the outcomes of an early-literacy intervention in schools

van der Weijden, F.A.

Publication date
2025

Document Version
Final published version

Weeks of practice and years of experience

Factors related to the outcomes of an early-literacy intervention in schools

FAE VAN DER WEIJDEN

Weeks of practice and years of experience
Factors related to the outcomes of an early-literacy intervention in schools

Fae van der Weijden

ISBN: 978 94 6506 745 2

Cover: Nicole Siers
Lay-out: André van Delft
Printing: Ridderprint | www.ridderprint.nl

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

This dissertation was sponsored by Stichting Kohnstamm Fonds.

Weeks of practice and years of experience
Factors related to the outcomes of an early-literacy intervention in schools

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College
voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Aula der Universiteit
op woensdag 22 januari 2025, te 14.00 uur

door Fae Aimée van der Weijden
geboren te Woerden

Promotiecommissie

Promotor:	prof. dr. P.F. de Jong	Universiteit van Amsterdam
Copromotores:	dr. M. van den Boer	Universiteit van Amsterdam
	dr. A.H. Zijlstra	Universiteit van Amsterdam
Overige leden:	prof. dr. L.J.F. Cornelissen	Universiteit van Amsterdam
	prof. dr. E.H. de Bree	Universiteit van Amsterdam
	prof. dr. P. Ghesquière	KU Leuven
	prof. dr. J.E. Rispens	Universiteit van Amsterdam
	prof. dr. P.C.J. Segers	Radboud Universiteit

Faculteit der Maatschappij- en Gedragswetenschappen

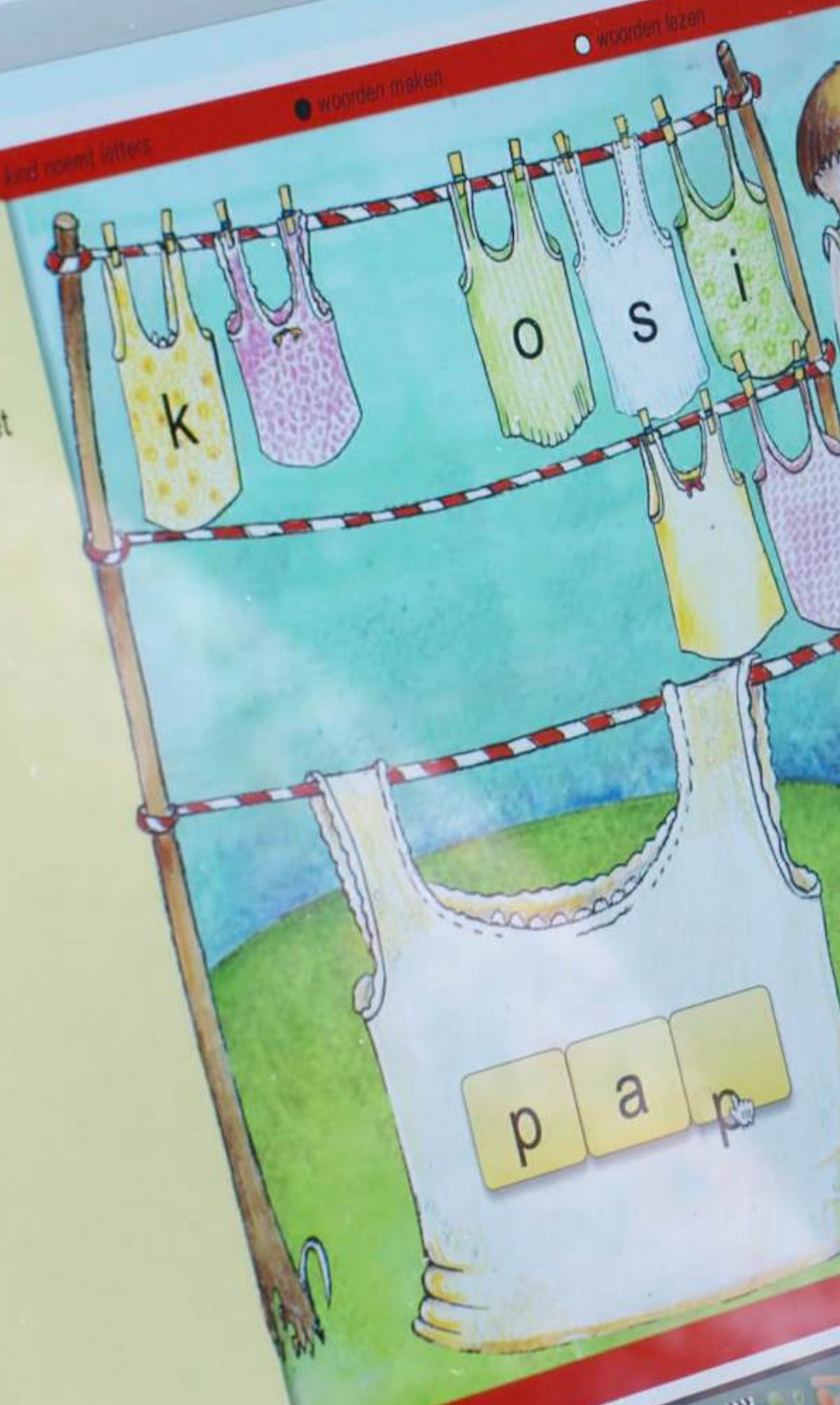
Table of Contents

yer.php

kind noemt letters
woorden maken
woorden lezen

et woord **pap**
maken.

p de p naar het
grote hemd)



CHAPTER 1

General Introduction

Reading is an essential skill that children need to develop to succeed in school and beyond. It paves the way for all other learning and communication in our literate society. Therefore, reading is considered to be one of the most important learning goals in primary education. In the early school years, large parts of the curriculum are devoted to reading or reading-related activities. Special efforts are made to raise attention for the quality of reading education, such as *Leesoffensief* in the Netherlands (de Leescoalitie, 2020) and *The Reading First Program* in the USA (NCEE, 2008).

Most children learn to read fairly well. However, 3% to 12% of children in primary school struggles with reading (Fluss et al., 2009; Snowling, 2013). The Dutch *Inspectie van het Onderwijs* (2019) reported that 7.5% of the Dutch sixth graders had a diagnosis of dyslexia (i.e. severe reading and/or spelling problems), of which 92% struggled with reading fluency. Reading problems can lead to academic failure and lowered academic self-esteem (Bear et al., 2002; Luyten & Bruggencate, 2011; Mol & Bus, 2011; Poskiparta et al., 2003). Moreover, reading difficulties can have long-term consequences for children's school career, future employability, and functioning in daily life (Annie. E. Casey Foundation, 2010; Buisman et al., 2012).

Reading problems in first grade tend to persist into adolescence, even though most schools provide additional support when children lag behind (Ferrer et al., 2015). Research showed that even with intensive remedial instruction children have difficulties to overcome their reading difficulties, especially in reading fluency (Torgesen et al., 2001). This knowledge underlines the need for prevention of reading problems by providing additional support in kindergarten or even preschool to children at risk for reading problems (Ferrer et al., 2015; Torgesen, 2002).

1 Prevention of Reading Problems

Many early-literacy interventions have been developed to prevent reading problems (for reviews see Verhoeven et al., 2020; Wanzek & Vaughn, 2007). These interventions, provided in preschool and kindergarten, are mostly focused on the training of preliteracy skills, not reading skills. Specifically, early-literacy interventions generally target letter knowledge and phonological awareness (Ehri et al., 2001a; 2001b). Letter knowledge refers to the number of letter-sound correspondences a child knows. Phonological awareness is the ability to recognize and manipulate sounds in

words (for example rhyming or blending sounds into words). When reading a word, letter knowledge is needed to recode the letters into sounds and phonological awareness is needed to blend the sounds into a recognizable word (Elbro & de Jong, 2017). This decoding subsequently contributes to orthographic knowledge (i.e. knowledge about the written form of words) and thereby to sight word reading (de Jong & Share, 2007; Share, 2008). As such, children with more letter knowledge and/or better phonological abilities in kindergarten generally learn to read more quickly (de Jong & van der Leij, 1999; Gijsel et al., 2006; Leppänen et al., 2008; Viersen et al., 2018). Training letter knowledge and phonological awareness prior to the beginning of reading instruction can contribute to early reading (Ehri et al., 2001a; 2001b; Suggate, 2010; 2016).

Together, phonological awareness and letter knowledge in kindergarten explain around 40% of the variance in reading in first grade and predict reading performance up to fourth grade (Leppänen et al., 2008; Torppa et al., 2007). Thus, for the other 60%, reading is predicted by other factors. One of these factors is rapid (automatized) naming, i.e. the ability to quickly retrieve the names of well-known symbols (pictures, numbers, or letters) from long-term memory. It is a consistent predictor of reading fluency across languages (Landerl et al., 2018). Children's rapid naming in kindergarten can predict their reading fluency skills up to second grade (de Jong & van der Leij, 1999; Viersen et al., 2018). However, most studies show that rapid naming is hard to improve by training (e.g. de Jong & Vrielink, 2004; Eleveld, 2005; for an exception see Pecini et al., 2019). Early-literacy interventions are thus not focused on rapid naming.

2 Effects of Early-Literacy Interventions

Early-literacy interventions targeting letter knowledge and phonological awareness generally show promising results on reading, having larger effects than remedial reading interventions (Ehri et al., 2001a; Lovett et al., 2017; Wanzek & Vaughn, 2007; Wanzek et al., 2013), although this is not shown by all studies (Suggate, 2016; Wanzek et al., 2016). In the long term, however, after the intervention has finished, effects of early-literacy interventions in preschool and kindergarten tend to fade out (Suggate, 2016). This is not surprising, as the precursors of reading can only partly predict reading problems (e.g. van Viersen et al., 2018). Word reading requires more than preliteracy skills (Elbro & de Jong, 2017; Elbro et al., 2012). Therefore early-literacy interventions are often integrated with the teaching of reading itself to foster long(er) lasting effects on reading (Hatcher et al., 1994). When such interventions are

continued for two or three years, they show promising results, i.e. effects that are still noticeable beyond second grade (Connor et al., 2013; Zijlstra et al., 2021).

An example of an extensive intervention for the prevention of reading problems that is focused on both precursors and reading is the Dutch early-literacy program *Build!* (in Dutch: *Bouw!;* Regtvoort & van der Leij, 2007; Regtvoort et al., 2013; Zijlstra et al., 2021). This intervention is computer-based and covers letter knowledge and phonological awareness, as well as word reading accuracy and fluency. The program starts in kindergarten, before formal reading instruction begins. It is not provided to all children, but to those who are considered at risk of reading problems, based on the precursors of reading. Children exhibiting poor letter knowledge and/or phonological awareness in kindergarten are enrolled in the program *Build!*. The aim is to provide them with a head start in first grade, when reading instruction starts. To avoid fade-out effects, the program continues for two years, until the middle of second grade.

The intervention showed its effectiveness in three RCTs. In the first RCT (Regtvoort & van der Leij, 2007), conducted in kindergarten, the program was found to increase children's letter knowledge and phonological awareness to average levels (respectively a large and moderate effect). The second RCT (Regtvoort et al., 2013) showed that children who received *Build!* from first grade onwards and who completed the program before the end of second grade, reached average levels of word reading fluency by the middle of first grade, and maintained these levels at least until the end of third grade. In the third RCT (Zijlstra et al., 2021), *Build!* was effective for a subgroup of about 60% of the children. By the middle of sixth grade the number of children with reading problems within this subgroup was substantially lower in the intervention group (22%) than in the control group (54%). For the remaining 40% of the children, the intervention was ineffective. Zijlstra et al. suggest that children whose parents were less proficient in Dutch, having an immigrant background and/or lower educational level, were possibly hard to reach as tutor, which could have resulted in less exposure to the intervention and lower intervention outcomes for their children.

The program *Build!* has currently been implemented in 80% of Dutch primary schools. Although the intervention has shown its effectiveness in several RCTs, it has not been tested whether it is effective on a large scale. This dissertation is focused on the effectiveness of *Build!* when it is implemented by schools.

3 Implementation of Interventions by Schools

For three reasons lower effects can be expected in large-scale studies in which schools implement the intervention than in the initial RCTs. First, RCTs are often small-scale studies. Generally, large-scale studies produce lower effect sizes than small-scale studies (Cheung & Slavin, 2016). Large-scale studies often include larger and more diverse samples, which results in lower effect sizes than studies with specific target groups (Lortie-Forgues & Inglis, 2019). Second, some interventions need translation 'from lab to field' before they work well in natural school settings, i.e. making the materials and manuals more teacher- and student-friendly, resolving weaknesses at scale, and adopting productive adaptations made by teachers using the program (Burkhardt & Schoenfeld, 2003). Without such adaptations effects may be only small. Third, RCTs are often researcher-led. That is, the implementation of the intervention is guided by researchers (e.g. Lovett et al., 2017; Mathes et al., 2005; Zijlstra et al., 2021). They, for example, provide training and support and frequently visit schools to monitor and stimulate the implementation, resulting in a proper implementation of the intervention. In contrast, when an intervention is scaled up, schools do not always receive the necessary training and support to implement an intervention as intended (Stein et al., 2008). In all, it is a critical question whether interventions that produce impacts in small or localized trials can show similar effects in large studies in which the intervention is implemented by schools.

The extent to which the intervention is implemented as intended is called treatment integrity (Gresham et al., 2000). Treatment integrity is not reported in a large proportion of studies on (reading) interventions. If reported, relations between treatment integrity and children's outcomes are seldomly examined, while treatment integrity could be an important factor in explaining why intervention effectiveness differs among children (Capin et al., 2018; Swanson et al., 2013). Only a few studies show that, within one intervention, children reach lower outcomes if the intervention is implemented more poorly (Fogarty et al., 2014; Vadasy & Sanders, 2009; Wolgemuth et al., 2014; Zijlstra et al., 2014). In large-scale studies on reading interventions, the examination of treatment integrity is even more rare (for an exception see Stein et al., 2008), while variation in treatment integrity, and thereby its influence, might be larger when more schools are included and schools have more freedom to implement the intervention. As treatment integrity might be key to the success of intervention at scale, more studies are needed that investigate what aspects of treatment integrity determine intervention outcomes.

When evaluating the scale-up of an intervention, schools and policy makers might not only want to know whether early-literacy interventions are effective at the individual level, which is mostly examined (Suggate, 2010; 2016), but also at the school level. That is, whether implementing the intervention results in a reduction of reading problems in schools. Effects at the individual level do not necessarily result in the reduction of reading problems at the school level, especially when only small effects can be expected (Lortie-Forgues & Inglis, 2019; Thomas et al., 2018).

In large-scale studies, effects might not be visible right after the intervention is implemented, but only after a few years. Achieving full treatment integrity may take schools more than a year. Furthermore, Harn et al. (2013) suggested that schools make modifications during implementation which lead to a better fit with the needs of staff and children. Thus the program needs to be adapted and become institutionalized or part of the school routine before reaching its potential. In line with these ideas, Torgesen (2009) found that *the response to intervention model* became more effective over three years of large-scale implementation. However, since this study lacked a control group, it was unclear whether only the intervention was responsible for the results. Therefore, it remains unknown whether an intervention really becomes more effective when schools use it for a longer time.

4 Effects of Early-Literacy Intervention For Specific Subgroups

Another question related to intervention effectiveness is whether early-literacy interventions are also effective for specific subgroups. Several reviews have shown that children with poor literacy skills (e.g. phonological skills, rapid naming, and letter knowledge), as well as children with special needs (learning, attention, or behavior problems) tend to be less responsive to early-literacy interventions (Al Otaiba & Fuchs, 2002; Lam & McMaster, 2014; Nelson et al., 2003). Generally, little is known about background and family characteristics related to responsiveness to interventions, because child and family characteristics (e.g., ethnicity, primary language, socioeconomic status) are mostly unreported (Manz et al., 2010). Moreover, vulnerable groups of children, e.g. children from low-income and ethnic-minority families, are mostly underrepresented. Thus, more research is needed on the effects of such family characteristics on the outcomes of early-literacy interventions to inform schools which interventions are most effective for which children. This way, all children at risk for reading problems can be provided with the most effective support.

The next question is why interventions are less effective for specific subgroups. Perhaps, the reduced effectiveness is attributable to the way the intervention is

delivered to these subgroups, i.e. to lower levels of treatment integrity. Alternatively, when the intervention is delivered in the same manner, it could be that children simply progress more slowly through the program. Few studies have investigated such questions.

An exception are studies on the effectiveness of early-literacy interventions for children with familial risk of dyslexia. Familial risk for dyslexia refers to the existence of reading problems in the family. If one or both of the parents has/have reading problems, the child is likely to shows delays in the development of phonological awareness and letter knowledge and is three to four times more likely to develop reading problems than children without such a familial risk (Snowling & Melby-Lervåg, 2016). Moreover, studies on the effectiveness of early-literacy interventions for children with and without familial risk showed that kindergartners and preschoolers with familial risk responded less well to interventions targeting letter knowledge and/or phonological awareness, than non-at-risk kindergartners (Elbro and Petersen, 2004; Hindson et al., 2005). Hindson et al. (2005) showed that at-risk-children required more teaching sessions to reach the same outcomes. Similarly, Zijlstra et al. (2021) found that an early-literacy intervention can be equally effective for children with and without familial risk, if the at risk children received this additional practice. As these studies were conducted on a small scale, the question remains whether additional practice is provided to vulnerable groups of children in natural school settings and how familial risk affects intervention outcomes.

5 This Dissertation

This dissertation focused on the effectiveness of the large-scale implemenation of the early-literacy intervention *Build!*. It addresses three questions: 1) Is dosage, a dimension of treatment integrity, related to intervention outcomes when an intervention is implemented on a large scale and schools implemented the intervention? 2) Do familial risk for dyslexia and parental education affect treatment integrity and outcomes of early-literacy interventions? 3) What are the effects of early-literacy interventions at the school level and does an intervention become more effective when schools use it for a longer time?

The first question is addressed in Studies 1 and 2 (Chapters 2 and 3). These studies are focused on a specific aspect of treatment integrity, i.e. dosage. Dosage refers to the amount of practice children receive (Dane & Schneider, 1998). Natural variations in dosage across children and schools are examined in relation to intervention outcomes, with the aim to understand whether this specific aspect of treatment

integrity contributes to the success of an early-literacy intervention when implemented in natural schools settings. Study 1 (Chapter 2) is focused on kindergarten and answers the question whether variations in dosage are related to children's gains in letter knowledge and phonological awareness. Dosage is investigated in a detailed way. That is, multiple aspects of dosage are distinguished, such as the number of sessions per week and the duration of sessions. Distinguishing among various aspects of dosage might give more insight in what kind of practice is most effective, for example frequent but short sessions or infrequent but long sessions. Study 2 (Chapter 3) includes multiple intervention periods between kindergarten and the middle of second grade to investigate whether dosage is not only related to preliteracy skills, but also to reading skills. It is examined whether this aspect of treatment integrity matters during the full length of the intervention.

In Study 2 (Chapter 3), it is also investigated whether two family characteristics, i.e. familial risk for dyslexia and parental education, affect intervention outcomes, either directly or indirectly via treatment integrity. Thereby, more insight is created in whether the intervention was equally effective for children whose parents have a lower and higher educational level, as well as for children who have a parent with reading problems and who have not. Moreover, it is examined whether a potential difference in effectiveness could be explained by the way the intervention was delivered to these children, especially with respect to the amount of practice children received. For children with familial risk, the question is whether they received the extra practice they needed in natural school settings.

The third question is addressed in Study 3 (Chapter 4). This is a quasi-experimental study in which it is examined whether there was a reduction of reading problems and/or an increase in the average reading ability at the school level from the moment the intervention *Build!* was implemented. Effects during the intervention (mid- and end-Grade 1), effects at post-test (mid-Grade 2), and effects at follow-up (end-Grade 2 and mid-Grade 3, i.e. 0.5 to 1 year after the intervention was finished) were determined. The results are informative for schools and policy makers who (want to) implement the program *Build!* to determine whether the benefits of the intervention outweigh the costs. Furthermore, it was examined whether the effects became stronger when schools used the program for a longer time. Thereby, more insight is created into the role of implementation when evaluating scale-ups of (reading) interventions.

Finally, in the General Discussion (Chapter 5) the results of the previous studies are further discussed. In this General Discussion, also reach is considered, that is the proportion of the target population that participates in the intervention. Three issues

of reach were addressed: 1) the challenge of predicting in kindergarten which children are in need of the intervention based on their preliteracy skills, 2) schools' decisions to include or exclude children from the intervention, and 3) the identification of false positives during the intervention, i.e. children who receive the intervention but might not need it (anymore). *T*o these ends the selection procedure that was used by part of the schools in the first and second study was evaluatie.


SAMSUNG

CHAPTER 2

Dosage Explains Individual Differences in the Outcomes of a Prevention Program for Literacy Problems

Abstract

We investigated whether dosage (the amount of practice with an intervention) was related to the outcomes of the computer-based early-literacy intervention *Build!* as implemented by schools. Progress within the intervention (the number of lessons completed) was examined as mediator. Three aspects of dosage were distinguished: Session frequency, session length, and the number of intervention weeks. Participants were 226 kindergartners from 45 schools. Letter knowledge and phonological awareness were assessed before and after the intervention. Findings showed that, controlling for performance at pre-test, the relation between dosage and preliteracy at post-test was completely mediated by progress within the intervention. Frequency of intervention sessions showed the strongest relation with progress and literacy outcomes, both at the child level and the week level. Session length and number of intervention weeks had smaller effects. Overall, the findings underline the importance of dosage for the outcomes of a literacy intervention as implemented by schools.

van der Weijden, F. A., van den Boer, M., Zijlstra, A. H., van der Leij, A., Zijlstra, B. J. H., & de Jong, P. F. (2024). Dosage explains individual differences in the outcomes of a prevention program for literacy problems [Manuscript submitted for publication]. Department of Child Development and Education, University of Amsterdam.

1 Introduction

Between 3% and 12% of the children fail to develop appropriate levels of reading ability (Fluss et al., 2009; Snowling, 2013). These children are at risk of academic failure and lowered academic self-esteem (Bear et al., 2002; Luyten & Bruggencate, 2011; Mol & Bus, 2011; Poskiparta et al., 2003). To reduce reading problems and prevent these negative outcomes, several reading interventions have been developed (Ehri et al., 2001a). Early-literacy programs, starting in kindergarten or first grade, are found to be effective (Scammacca et al., 2007; Suggate, 2016). Following positive evaluations, some programs are being implemented on a large scale and implemented by schools, in contrast to often small-scale researcher guided Randomized Controlled Trials (RCTs). An important question is to what degree treatment integrity is maintained, and especially, whether that has implications for the outcomes of the intervention.

Treatment integrity–also known as treatment fidelity, fidelity of implementation, or intervention fidelity–refers to the degree to which the intervention is implemented as intended (Gresham et al., 2000; Swanson, Wanzek et al., 2013). In the current study we examined the relation of treatment integrity with intervention outcomes for the Dutch computer-assisted early-literacy intervention *Build!*. Several RCTs have shown that with *Build!* children can reach average levels of word reading fluency and that this can last for at least one to four years after the intervention (Grade 3-6; Regtvoort et al., 2013; Zijlstra et al., 2021). Moreover, the intervention shows transfer effects to text reading and spelling (Regtvoort et al., 2013; Zijlstra et al., 2021). *Build!* is currently being implemented at a large scale by Dutch primary schools and therefore provides a well suited context to evaluate the role of treatment integrity.

Dane and Schneider (1998) have distinguished five dimensions of treatment integrity: adherence, quality, dosage, participant responsiveness, and program differentiation. *Adherence* is the extent to which essential intervention elements are implemented as intended. *Quality* refers to the instructional quality in delivering the intervention elements. *Dosage* is the amount of practice with the intervention a child is provided with. *Participant responsiveness* refers to the extent to which children are engaged by and involved in the intervention activities. *Program differentiation* is the extent to which the intervention is different from the control condition.

Most studies on treatment integrity have been focused on adherence, quality, and dosage (Capin et al., 2018; van Dijk et al., 2023). The present study is centered around dosage. With the increase in computer-based interventions (Bautista et al.,

2024), dosage is of specific interest, because other dimensions of treatment integrity, such as adherence and quality, are at least to a considerable extent covered by computer programs. In contrast, the amount of practice children receive could still differ across children and schools, especially in large-scale implementations of interventions.

A second reason to study dosage are the inconsistent findings of studies in which the relation between dosage and intervention outcomes have been examined. Several meta-analyses (Suggate, 2010; 2016; Wanzek & Vaughn, 2007; Wanzek et al., 2016) have indicated that dosage is not directly related to intervention outcomes, while empirical studies have shown that dosage is important when considering differences in effectiveness of one particular intervention (Nunnery et al., 2006; Wolgemuth et al., 2014; Zijlstra et al., 2014). An important difference between these types of studies is the unit of analysis. Empirical studies indicate that *children* receiving a smaller or larger dose of the same intervention differed in intervention outcomes, while in the meta-analyses *interventions* including a larger or smaller number of sessions were compared. Time spent on the intervention does not seem to make some interventions more effective than others, but it might still be the case that within the same intervention, more time spent on the intervention is associated with better outcomes.

Inconsistent findings were also found in the systematic review of Van Dijk et al. (2023), describing 27 studies in which the relation between dosage and outcomes of literacy interventions was examined. Of all 93 different literacy outcomes across studies, 53 outcomes were not related to dosage, 29 were positively related to dosage, six were negatively related to dosage, and for five outcomes the effect of dosage was unclear. This inconsistency was possibly due to the various ways dosage is defined. For example, in a study that found a negative relation between dosage en literacy, dosage was defined as the difference in recommended and actual dosage. Thus, a lower dose than prescribed was related to lower levels of literacy. Other studies defined dosage as time spent on the intervention (e.g. the number of months, days, hours, sessions or minutes spent on the intervention), the extent to which a schedule of sessions was followed (e.g. the percentage of sessions provided at the right date), the amount of intervention provided (e.g. the number of readings completed), or the performance within the intervention (e.g. number of words read correctly) (for an overview of all definitions see supplemental materials of van Dijk et al., 2023).

To understand the inconsistent findings, we distinguish several definitions and aspects of dosage and relate them to intervention outcomes. First, we distinguish progress within the intervention (i.e. exposure to the intervention content) from

dosage (i.e. time spent on the intervention). There are studies showing a relation between dosage and intervention outcomes (e.g. Wolgemuth et al., 2014) and studies showing a relation between progress and intervention outcomes (e.g. Hsin et al., 2023). However, in previous studies, dosage and progress are not clearly distinguished. It seems evident, but has in fact not been tested directly, that it is not hours spent on the intervention (dosage) per se, but the amount of content exposed to during the intervention (progress), that in turn determines intervention outcomes. Progress itself, however, may not depend on dosage alone. For example, the computer-based intervention *Build!* consists of numerous lessons to be completed by a child under supervision of a tutor. The number of new lessons that a child can complete within a certain amount of time, i.e. progress through the program, depends on the time it takes a child to complete a lesson and, in adaptive programs, the number of lessons that have to be reviewed. Accordingly, children with the same amount of dosage can still differ in the amount of exposure to the material of the intervention program. In the present study, we therefore distinguish dosage, i.e. the number of hours spent on the intervention, from progress, i.e. the number of intervention lessons or program parts that a child has completed. Previous research has provided some indication that the two are not identical, specifically for *Build!* (Regtvoort et al., 2013; Zijlstra et al., 2014), and also for a mathematics intervention (Muñez et al., 2022).

Moreover, three aspects of dosage are distinguished: (1) frequency, the number of sessions per week, (2) length, the average length of a session, and (3) duration, the number of weeks spent on the intervention (Dane & Schneider, 1998; Marulis & Neuman, 2010; Tran et al., 2011). Most studies into differences in dosage of the same intervention have focused on only one of the aspects of dosage (van Dijk et al., 2023). The distinction is particularly interesting in light of the well-known difference between massed and spaced practice (Donovan & Radosevich, 1999; Dunlosky et al., 2013). Surprisingly, these principles of consecutive (long) versus distributed (frequent, short) learning opportunities have only recently been extended to literacy development in a small-scale experimental study on orthographic learning (Wegener et al., 2022). To our knowledge it is unknown whether these principles are also applicable to responsiveness to early-literacy interventions.

1.1 *Current study*

1.1.1 *The Intervention Build!*

In the current study we investigated how dosage (frequency, length, and duration) is associated with the responsiveness to the evidence-based early-literacy intervention

Build! (in Dutch: *Bouw!*). This intervention is a two-year early-literacy program, starting in kindergarten –before formal reading instruction begins–and continuing until the middle of second grade. It is a computer-based program covering the precursors of reading (phonological awareness and letter knowledge) and the accurate and fluent reading of simple and complex words. The program consists of twelve program parts. Each program part (around 40 lessons) is followed by a built-in test. If children score below mastery level (i.e. less than 80% correct), review lessons are provided. Children are assisted by a tutor, either a professional (i.e. teacher) or non-professional (i.e. parent, volunteer, or older child; Zijlstra et al., 2014). Per week three to four sessions of 10-15 minutes should be completed, partly at school and partly at home. The proportion of lessons at home differs per school.

Three randomized controlled trials (RCTs) showed that *Build!* can effectively reduce reading problems. In the first RCT, conducted in kindergarten, the program was found to increase children's letter knowledge and phonological awareness to average levels (respectively a large and moderate effect; Regtvoort & van der Leij, 2007). The second RCT showed that, children who received *Build!* from first grade onwards and who completed the program before the end of second grade, reached average levels of word reading fluency, and maintained these until at least third grade (a small effect; Regtvoort et al., 2013). The program showed transfer effects to text reading accuracy, text reading fluency, and reading comprehension, but not to spelling. The third RCT showed that the intervention was highly successful for part of the children, that are the children whose parents responded to the questionnaire about the incidence of dyslexia in the family; Zijlstra et al., 2021). In this subgroup, large and long-lasting effects of *Build!* were found on children's word reading fluency: practicing with *Build!* reduced the number of children with reading problems from 61% to 28% by the middle of third grade and from 54% to 21% by the middle of sixth grade. Moreover, there were transfer effects to text reading fluency and spelling, but not to reading comprehension. Taken together, *Build!* could be seen as an evidence-based early-literacy intervention. The next step would be to investigate the implementation of the program in natural school settings.

1.1.2 *Research Questions*

The current study focused on a school-based implementation of *Build!*, a program for the prevention of reading problems. The first goal of the study was to investigate–on a weekly basis–the effect of dosage, i.e. frequency and length of intervention sessions, on the progress within the intervention. The second goal was to examine–at the child

level–the relation between dosage and intervention outcomes and the potential mediating role of progress within the intervention. To keep the study concise and comprehensible, we focused on only one part of the intervention, the part that focused on preliteracy skills, which was provided in kindergarten. During this period, the learning content in *Build!* is mostly new to children as formal literacy instruction has not yet started (see paragraph on the Dutch Educational System).

Two characteristics of *Build!* can affect the relation between dosage (i.e. frequency, length, and duration) and progress within the program (i.e. the amount of content exposed to). One characteristic concerns the review lessons provided for children with poor performance within *Build!*. The number of review lessons will affect progress. A second characteristic is that the intervention sessions in kindergarten are partly provided at school and partly at home. As the number of review lessons and the proportion of practicing at home might affect how fast children progess through the program content, these two factors were included as additional predictors of progress, next to dosage.

Our study differs in several respects from earlier studies in the field. First, a distinction is made between dosage and progress within the intervention. We hypothesized that the relation between dosage and intervention outcomes would be mediated by the amount of exposure to the content of the intervention as indicated by progress. Second, we considered various aspects of dosage, that is frequency and length of sessions per week, as well as duration of the intervention (in number of weeks). Following evidence for a benefit of spaced over massed practice for reading acquisition (Wegener et al., 2022), we expected that frequency and duration would have a stronger association with progress and literacy gains than the length of sessions. Finally, we considered the relation between dosage and progress both within and between children. That is, we expected not only that children who practice more make more progress within the intervention, but also that children's progress is larger in weeks in which dosage is higher. The former could be purely a correlational finding, that could be affected greatly by other variables, such as abilities and motivation at the beginning of the study. The latter (week-to-week relations between dosage and progress) increases the credibility of a causal relationship, as children act as their own control. Note that the current study did not involve a no-intervention control group. Inclusion of a control group would provide stronger evidence for the effect of the intervention, but cannot be used to examine the relation between dosage and responsiveness to the intervention, as a control group obviously has zero dosage.

We would like to note that the current study was followed up by another study of van der Weijden et al. (2024b), also on the implementation of *Build!* in natural

school settings. In that study, children were followed during a longer period (from kindergarten until halfway second grade) and associations between dosage, progress, and word reading accuracy and fluency were examined at the child level. Dosage was defined as the number of hours spent on the intervention. The current study provides a more detailed view on dosage by including different aspects of dosage and examining relations between dosage and progress at both the week level and child level. To minimize the number of potential confounds, we focused on the purely preventive part of the program (i.e. before the beginning of classroom instruction), which is provided in kindergarten and primarily focused on preliteracy skills.

2 Method

2.1 *Participants*

The sample consisted of 226 Dutch kindergarten children at risk for reading difficulties from 45 schools, located in two school districts. The children were selected from a larger sample of 768 children who received the intervention. Informed consent was obtained from parents and provided for 420 of these children (55%). Of these 420 children, 188 children were excluded because they did not start in time with the intervention, that is between January and April of the second kindergarten year (see also The Dutch Educational System). Finally, six children were removed because of a disability that hindered the child in learning to read (e.g. low IQ, brain damage, hearing problems). Descriptive statistics of children's sex, nationality, and socio-economic status are displayed in Table 2.1. The children were on average 5.30 years old (SD = 0.39) at pretest, halfway through the second kindergarten year (see The Dutch Educational System).

The children were selected for the intervention by the schools halfway the second year of kindergarten. Schools and teachers selected children who they deemed eligible for the intervention. Schools in District 1 followed another protocol than schools in District 2, although procedures were quite similar. Schools in District 1 (194 children; 32 schools) were advised by a local project group to adopt a two-step selection procedure based on measurement occasions in October and January. In October of the second kindergarten year, the children should be tested on productive letter knowledge and phonological awareness (see Measurements). Schools were advised to provide extra instruction to children with low scores (lowest 30%). In January, children were to be tested again on productive letter knowledge and phonological awareness and children who still scored low (lowest 25-30%) would be eligible for the intervention. The local project group also provided the schools in District 1

with the test materials and cut-off scores that could be used for the selection of the children for the intervention. All schools had to administer the tests enabling selection, but were free to use the cut-off scores and could include other children for their own reasons. Schools in District 2 could use the Dutch protocol for the prevention of reading problems in kindergarten (Druenen & Koning, 2017). This protocol is widely available to schools to enable early identification of children at risk of reading difficulties. The protocol contains a screening list that needs to be filled out in kindergarten, twice a year (January and June). To fill it out, teachers need to observe the child and to administer norm-referenced tests to measure pre-literacy skills, like letter knowledge, phonological awareness, and rapid naming. The protocol advises schools to provide *Build!* to children who do not meet the norms. Most schools in District 2 (90%) reported that they indeed used teacher observations and pre-literacy tests (letter knowledge, phonological awareness, rapid naming) to select the children for *Build!*.

Table 2.1

Sample Descriptives of Sex, Nationality, Home Language, and Parents' Educational Level

Variable	Subgroup	%
Sex	Boys	61.67
Nationality	Dutch	94.67
Home language	Dutch	92.38
Educational level mother[a]	Low	12.78
	Average	37.44
	High	32.16
	Unknown	17.62
Educational level father[a]	Low	11.89
	Average	27.75
	High	20.26
	Unknown	40.09

[a]low = primary or secondary education; average = vocational education; high = higher professional education and university.

Pre-test scores were available for 79% of the children (there was data missing from District 2; for more information, see Missing Data). Among them, approximately half of the selected children (60%) were children we would indicate to be 'at-risk', i.e. children who scored low on phonological awareness (≤25th percentile)

and/or letter knowledge (knowing ≤6 letters). A second group of children could also be considered 'at risk' (13%), because they scored below average on both phonological awareness (26-50th percentile) and letter knowledge (7-8 letters). The remaining children (27%) had average to good phonological awareness and letter knowledge, and were thus selected by the schools for other reasons.

2.1.1 *The Dutch Educational System*

As this study was conducted in the Netherlands, some remarks about the Dutch educational system are in place. Education in the Netherlands is mandatory from five years old onwards, but most children go to school when they turn four. Primary school consists of two years of kindergarten (ages 4-6), followed by Grades 1 to 6 (ages 7-12). When children enter school, they differ greatly in their pre-literacy knowledge due to differences in their home-literacy environment. On the overall majority of schools, first- and second-year kindergartners are placed in the same classroom. Consequently, education in the first and second year is largely similar. Literacy education is not yet systematic or intensive literacy instruction. Instead, children learn some letters and train phonological skills, like rhyming, through game-based lessons. After one and a half years of kindergarten, the intervention *Build!* is started for children who (still) have poor pre-literacy skills. In the program, children learn letter-sound correspondences and phoneme blending. As this content is generally offered in first grade, the intervention period at the end of kindergarten could be seen as preventive of future reading difficulties. In first grade, formal reading instruction begins, i.e. systematic and intensive phonics instruction. In parallel *Build!* continues with teaching more letter-sound correspondences and letter clusters, as well as word reading of monosyllabic words. By the end of first grade, children are expected to read and spell regular monosyllabic words. The intervention *Build!* is continued until the middle of second grade.

2.2 *Design*

The full intervention takes two years: from the middle of the second year of kindergarten until the middle of second grade. In this study we followed children until the end of kindergarten, using a longitudinal design of three waves: the beginning of the second year of kindergarten (screening), the middle of the second year of kindergarten (pre-test), and the end of kindergarten (post-test). All tests were administered by the teacher, the school counselor or a trained research assistant at school, outside the classroom. Prior to data collection, approval was obtained from the Ethics Review

Board of the Faculty Social and Behavioral Sciences of the University of Amsterdam (project number 2018-CDE-8677).

2.3 *Intervention Program Build!*

Build! (in Dutch: *Bouw!*) *is an* intervention program that consists of 523 digital lessons covering pre-reading (letter knowledge and phonological awareness in kindergarten), beginning reading (decoding in first grade), and advanced reading skills (reading fluency from the middle of first grade to the middle of second grade; Regtvoort et al., 2013; Zijlstra et al., 2014; 2021). *T*he reading exercises include one-syllable and two-syllable words, containing all orthographic complexities of Dutch (see van der Leij & van Daal, 1999): single letters and digraphs (e.g. /uu/ /oe/ /ie/), letter combinations (e.g. /sch/ /oei/ /ng/), consonant clusters (e.g. /lf/ /st/ /tr/), and open and closed syllables (e.g. tonen [tones], tonnen [barrels]).

The lessons are divided into twelve program parts, of which the first two are generally provided in kindergarten. In these two program parts, children are presented with 12 out of 34 Dutch graphemes. The learning content is presented in different types of lessons. To acquire these letters, children have series of lessons. In these series, children are first taught one or two new letter-sound correspondence(s) (*Letter* lesson), then practice with blending the sounds into words (*Phono* lesson), and subsequently to read one-syllable words with the learned letters (*Build* lesson). After three series, when four letters have been acquired, children also have one to three game-based lessons involving decoding and orthographic knowledge (*Domino, Memory, and Self lessons*). In first grade, children complete Program Part 3-5 to learn the remaining 22 Dutch graphemes.

To finish a program part (on average 35 lessons), a test is administered. If children answer less than 80% of the items correctly, they have to complete review lessons before they can continue to the next program part. The review lessons are suggested by the program, based on the child's test performance. If children answer 80% to 95% of the items correctly, they can continue to the first lesson of next program part. If children answer 95% or more of the items correctly, they are directed to the test of the next program part. If they have 80% of the new items correct, the program part is skipped. Otherwise, specific lessons of that program part are suggested, based on the child's performance. For more information about the program content, see Regtvoort et al. (2013) and Zijlstra et al. (2014; 2021).

Instructions for the tutors are provided on the screen. The tutor reads aloud the instruction and provides feedback and emotional support if needed. Previous

research has shown that non-professional tutors are able to provide sufficient quality of instruction, quality of support, and program differentiation (Zijlstra et al., 2014). In the current sample, most children were tutored by more than one person. At home, the tutor was often a parent or older sibling. At school, the tutor was mostly a volunteer, school staff member or older child from grade 5 or 6. It was prescribed to provide children with 3-4 intervention sessions per week of 10-15 minutes each. Most schools aimed to provide 2-3 sessions per week at school and asked parents to provide 1-2 session(s) at home.

2.4 *Measures*

As there were two different aims in this study, we used two types of measures: week-level measures and child-level measures.

2.4.1 *Week-Level Measures*

To investigate the effect of dosage on progress within the intervention, we measured dosage and progress from week to week. Data were obtained from the computer program. The program had registered when and where each lesson had taken place in computer logs. Specifically, these logs contained information about which lesson was finished on which date, at what time, from which location (home or school), how long it took, and whether or not the lesson was completed. Note that the program registered lessons–not sessions–while our variables were based on sessions. As such, we first distinguished sessions. We assumed that lessons done on the same day at one location together constituted one session. Then, we calculated progress, frequency, length, proportion of sessions from home, and number of review lessons for each week of the intervention.

Progress

Progress was defined as the number of new lessons completed per week. It was computed by the number of lessons completed in a week, minus the number of review lessons.

Dosage

Dosage was measured in terms of frequency and length.

- *Frequency*: Frequency represented the number of intervention sessions per week.
- *Length*: Length was defined as the average duration of intervention sessions per week, in minutes.

Proportion from home
Based on the location of interventions sessions (home or school), we calculated the proportion of intervention sessions at home per week.

Review lessons
The number of review lessons per week was counted.

2.4.2 *Child-Level Measures*

To examine the relation between dosage and intervention outcomes and whether this relation was mediated by progress within the intervention, we measured progress, dosage, and literacy outcomes at the child level from the middle of the second year of kindergarten until the end of kindergarten. Proportion from home and review lessons were measured as additional predictors.

Progress
Progress at the child level was quantified as the child sum score of the week level progress over the intervention period. That is the number of new lessons a child completed during the intervention period.

Dosage
Dosage was measured in terms of frequency, length and duration. From the week level measures, we took the child average over the intervention period. For example, frequency at the child level was quantified as the average number of intervention sessions per week during the intervention period. Duration, representing the number of weeks a child practiced within the intervention period, was measured as an additional aspect of dosage at the child level.

Literacy Outcomes
We included phonological awareness and letter knowledge. An overview of instruments per measurement occasion is shown in Table 2.2.

– *Letter Knowledge*: Productive letter knowledge was assessed with the Grapheme Test (*Grafementoets*; Verhoeven, 1993). The test included all 34 graphemes in the Dutch language. Graphemes were printed on one page in two columns. Children were asked to sound out all graphemes, without time limit. The score represents the number of correct items, with a maximum of 34. Reported Cronbach's α is above 0.85 (Verhoeven, 2000).

– *Phonological Awareness:* Phonological awareness was measured with the phonological awareness task of the CELF-4-NL (Kort et al., 2008), the Dutch version of the *Clinical Evaluation of Language Fundamentals fourth edition (Semel et al., 2003*). The test consisted of 9 subtests: (1) auditory synthesis, (2) identification of last phoneme, (3) identification of middle phoneme, (4) word segmentation, (5) syllable deletion 1, (6) syllable segmentation, (7) syllable deletion 2, (8) phoneme substitution, (9) syllable deletion 3. Each subtest contained five items, preceded by two example items, and was stopped after three subsequent incorrect responses. The score represents the number of correct items, with a maximum of 45. Cronbach's alpha reliability is .85 (Kort et al., 2008).

  At the end of kindergarten, only *t*hree subtests of the CELF-4-NL were used: identification of last phoneme, identification of middle phoneme, and phoneme substitution. Reliability analyses with a larger dataset showed that reliability of the three subtests together was acceptable at the end of kindergarten (Cronbach's alpha = .70).

Table 2.2

Overview of Instruments per Measurement Occasion

Instrument	Measurement Occasion in the Second Kindergarten Year		
	Beginning (screening)	Middle (pre-test)	End (post-test)
Letter Knowledge	x	x	x
Phonological Awareness			
Full Test	x	x	
Subtest B, C, and H			x

2.5 *Analytic Strategy*

2.5.1 *Detection of Outliers*

Note that the length of a session was derived from the length of the lessons that were done. As the computer logs of *Build!* contained only information per lesson (see Week Level Measures), we had to inspect which lesson durations were realistic. This depended on the lesson type, because some lesson types were longer than others. Specifically, three criteria were used to set the boundaries per lesson type: (1) it was determined at which point there was a gap or a drop in the distribution of lesson durations, (2) long tales of distributions were cut off, so that distributions had an

acceptable skewness, i.e. larger than -2 and smaller than 2, and (3) no more than 2% of the data could be labeled as unrealistic. As a minimum lesson duration, we took either 0.50 or 1.00 minutes, depending on the lesson type. As a maximum lesson duration, we took a value between 13 and 16 minutes for short lesson types (Phono, Memory) and between 18 and 35 minutes for long lesson types (Domino, Self, Build, Letters). Unrealistic lesson durations were replaced, because deleting one lesson would translate into a missing for the full session duration. Missing lesson durations were predicted based on two-way imputation, i.e. both on the child's average speed and the average length of the specific lesson (van Ginkel et al., 2007). In this particular study, we used the following formula for imputation: $\gamma_{tlc} = (\mu_{tc} / \mu_t) \times \mu_l$, where γ_{tlc} is the duration of lesson l of lesson type t for child c, μ_{tc} is the child average duration for lesson type t; μ_t is the sample average for lesson type t, and μ_l is the overall average duration of lesson l. Next, we used lesson durations to compute the average session length per week.

Next, outliers at the week level were inspected, i.e. outliers between children and outliers within children. Outliers were defined as values with a z-score lower than -3 or higher than 3. Finally, outliers at the child level were inspected, similarly to the week level. Outliers were transformed into missings.

2.5.2 *Week Level Analysis*

To first goal of this study was to investigate, within children, the effect of dosage on progress within the intervention per week. To reach this goal, we used multilevel modeling. The weekly progress (level 1) was treated as nested within children (level 2), and children as nested within schools (level 3). This way, we accounted for dependencies between children that came from the same school and dependencies between repeated measures within the same child (Snijders & Bosker, 2012). Using the full maximum likelihood estimator, all weeks without missing values were included in the analysis. Analyses were carried out with RStudio (Rstudio Team, 2020), used package: nlme (Pinheiro et al., 2019).

Dummy Coding

Prior to analysis, frequency and length per week were transformed into dummies. This way, we could investigate the added value of an additional intervention session and of each additional five minutes of practice. For example, if a dummy for two intervention sessions is significant, on top of the dummy for one intervention session, but the dummy for three sessions is not, we know that two sessions is the optimal

number of sessions. As such, we could investigate the optimal dose-response rate, i.e. the optimal number and length of intervention sessions per week. Categories were based on previous research and prescribed practice (Regtvoort et al., 2013; Zijlstra et al., 2014). We also created dummies for two other variables, the number of review lessons and proportion from home, because they were not normally distributed (skweness or kurtosis >1 or <-1).

- *Frequency*: Practicing once a week was chosen as a reference category. Four dummies were created: practicing ≥2 times, ≥3 times, ≥4 times, and ≥5 times a week.
- *Length*: A session length of less than 10 minutes was chosen as a reference category. Three dummies were created: sessions of ≥10 minutes, ≥15 minutes, and ≥20 minutes.
- *Review lessons*: The number of review lessons ranged from 0 to 5. As a reference category, we chose 0 review lessons per week. Three dummies were created: ≥1 review lessons, ≥2 review lessons, and ≥3 review lessons per week.
- *Proportion from home*: Proportion from home ranged from .00 to 1.00. Practicing hardly at home (<.20) was chosen as a reference category. Two dummies were created: practicing partly at home (.20-.79) and practicing mostly at home (≥.80).

Week Level Predictors

Using multilevel modeling, first an empty three-level model was fitted with progress within *Build!* as the outcome variable (Model 1). The two intercepts, for children and schools, were considered random, as children and schools could differ in their average progress. This model was compared to a regular regression model, using a deviance test, to examine whether a multilevel analysis was desirable. Second, dummies for frequency were added as predictors to the model (Model 2). Third, dummies for length were added as predictors to the model (Model 3). Fourth, proportion of intervention sessions at home was added as a predictor to the model (Model 4). Fifth, review lessons was added as a predictor to the model (Model 5).

Model Evaluation

For each model, fixed effects were evaluated with *t* tests. Random effects were evaluated with deviance tests (Snijders & Bosker, 2012). We used a significance level of 5%. For each model, the proportion of explained variance at level 1 was computed following Snijders and Bosker (2012). When the explained variance (R^2) was between 2% and 13% the effect size was considered small, between 14% and 26% medium, and above 26% large (Cohen, 1988).

Higher-Level Predictors and Random Slopes

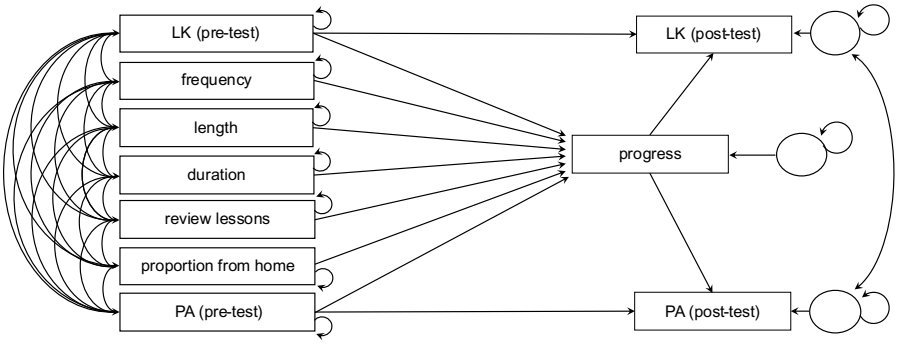
For the final model, we checked whether there were effects of dosage at the child and school levels (Model 6-13) and whether the fixed effects could be considered random for children (random slopes; Model 14-17). These checks involved 12 (8 + 4) additional significance tests. To reduce the risk of finding significant results by chance, we adjusted the significance level from .05 to .004 (Bonferroni correction).

2.5.3 *Child Level Analysis*

Our second aim was to investigate–at the child level–the relation between dosage and literacy outcomes and whether this relation was mediated by progress within the intervention. As this involved a mediation effect with multiple predictors and outcomes, structural equation modeling was used. Children (level 1) were treated as nested within schools (level 2). First, a full-mediation path model was built (see Figure 2.1). In this model, the three aspects of dosage (frequency, length, and duration), proportion from home, and review lessons were considered predictors of progress, which in turn was assumed to predict literacy skills at post-test. Literacy skills at pre-test were considered predictors of progress and literacy skills at post-test. Covariances between the exogenous variables were freely estimated. To deal with missing values and the multilevel structure, robust standard errors and a robust estimator for model fit were used (Yuan-Bentler for incomplete data; Yuan & Bentler, 2000). Analyses were carried out in Mplus version 7.31 (Muthén & Muthén, 1998-2017).

Figure 2.1

Hypothesized Mediation Model



Note. Circles represent residual variance. LK = letter knowledge; PA = phonological awareness; pre-test = the middle of the second year of kindergarten; post-test = end of kindergarten.

Dummy Coding

The number of review lessons and proportion from home were not normally distributed (skweness or kurtosis >1 or <-1). Therefore, we created dummies for these variables.

- *Review lessons*: The number of review lessons ranged from 0.00 to 2.33. A dummy was created to distinguish children who completed on average ≤0.50 review lessons per week from children who completed >0.50 review lessons per week.
- *Proportion from home*: Proportion from home ranged from .00 to .60. A dummy was created to distinguish children who did not practice at home (.00) from children who did (>.00).

Model Evaluation

We checked whether the full-mediation model held based on the following model fit indices: Chi-square, CFI, and RMSEA. We used an alpha level of .05 for the chi-square test statistic. Here, a nonsignificant chi-square test statistic indicated that the model fits the data (Schermelleh-Engel et al., 2003). CFI should be larger than 0.95 and RMSEA smaller than 0.08 for acceptable model fit, and CFI should be larger than .97 and RMSEA smaller than .05 for good model fit (Schermelleh-Engel et al., 2003). We adjusted the model if necessary, based on modification indices. For the final model, we assessed significance of direct and indirect effects. Direct effects were evaluated with *t* tests. Indirect effects were evaluated with the 95% confidence interval for parameter estimates, using the Bayes estimator. Standardized regression coefficients of .10 to .29 were considered small, .30 to .49 moderate, and ≥.50 large.

3 Results

The results are presented in three sections. First, we report on the amount of missing data. Second, we analyzed relations between dosage and progress within the intervention from week-to-week (within children). Third, we examined relations between dosage and intervention outcomes at the child level (between children) and whether these relations were mediated by progress within the intervention.

3.1 *Missing Data*

Prior to analyses, we checked for outliers. As described in the Method section, outliers were transformed into missings. We found several outliers on variables derived from the computer logs of *Build!* (frequency, length, duration, review lessons, proportion from home). All outliers were on the right side of the distribution (extremely

high). At the week level, no more than 1% of the data was depicted as outlier per variable, except for review lessons (3%). At the child level, no more than 1% of the data was considered as outlier per variable. Among literacy outcomes, we only found outliers on letter knowledge at pre-test. There were three children who knew 21 letters or more. At the child level, no more than 2% of the data was depicted as outlier.

The score on the third subtest of the phonological awareness post-test was missing for 47% of the children, because this subtest was not administered in the first half of the study. We assumed that these scores were Missing at Random (MAR), because missingness was not related to phonological awareness itself, but to the phase of the study. Missing scores were predicted based on the two other subtests, using a larger dataset including 923 children without missing data. Using multiple imputation (IBM SPSS Statistics for Windows Version 25.0), the score was imputed five times. We took the average of these five scores as the score on the third subtest.

When predicting literacy skills, we missed 91% of the pre- and post-test data from District 2, because the study started later in this district. From District 1, we missed 4% of the data, because some children were sick or entered school later. As described in the Method section, all available data was included in the analysis using the Yuan-Bentler estimator (Yuan & Bentler, 2000). As a consequence, results on predicting literacy skills were mainly based on data from District 1.

3.2 *Week Level Results*

The first goal of this study was to investigate–on a weekly basis–the effect of dosage, i.e. frequency and length of intervention sessions, on the progress within the intervention. Across weeks, children finished on average 3 new lessons per week in 2 sessions of 14 minutes each (see Table 2.3). Progress per week was strongly correlated with frequency (.62), moderately with length (.31), and weakly negative with review lessons (-.19; see Table S2.1 in the supplemental materials Chapter 2).

The three-level empty model (see Table S2.2 in the supplemental materials Chapter 2) fitted the data significantly better than a regular regression model (χ^2 (2) = 553.36, $p < .001$), indicating that indeed we had to account for the nested structure of the data. In fact, 73.62% of the variance in progress was located at the week level, 19.75% at child level, and 6.63% at the school level. Children's progress thus strongly varied across weeks.

The final multilevel model is shown in Table 2.4. Other multilevel models are shown in Table S2.2 in the supplemental materials Chapter 2. Frequency explained 46% of the variance in progress at the week level, a large effect. Children made more

progress during weeks they completed more intervention sessions. Each extra intervention session was of additional value (see Figure 2.2A). The effect of length was smaller, explaining 14% of the variance in progress, a medium effect. Children made more progress in weeks with longer sessions. Every five minutes of additional practice was associated with additional progress (see Figure 2.2B). No variance was explained by proportion of intervention sessions at home. Like the effect of frequency, the effect of review lessons was large, explaining 36% of the variance in progress. Weeks with more review lessons were associated with less progress. Each review lesson was at the cost of approximately one new lesson.

Table 2.3

Week Level Means and Standard Deviations of Frequency, Length, Proportion from Home, Review Lessons, and Progress

Variable	*M*	*SD*
Progress per week	3.01	2.29
Frequency[a]	2.20	1.04
Length[b]	13.62	5.86
Proportion from home[c]	.21	.31
Review lessons	0.59	1.08

[a]the number of intervention sessions per week; [b]average session duration per week, in minutes; [c]proportion of intervention sessions at home.

Effects of frequency, length, proportion at home, and review lessons on progress differed across children (see Table 2.4, Random Slopes for Children). Some children made more progress than others per session (χ^2 (2) = 257.29, p <.001) and per minute (χ^2 (3) = 104.79, p <.001). Furthermore, taking review lessons had a larger impact on some children than on others, in terms of progress (χ^2 (5) = 31.42, p <.001). Among the two dummies for proportion at home, the effect of practicing partly at home had a random slope and the effect of practicing mostly at home did not. This means that children differed in the amount of progress they made during weeks they practiced partly at home (χ^2 (4) = 18.65, p <.001) and that they made similar progress during weeks they practiced mostly at home (χ^2 (5) = 6.63, p = .20). We did not check whether fixed effects could be considered random for schools, because there was too little school variance. In total, 77% of the variance in progress per week was explained by the model.

Table 2.4
Final Multilevel Model Results for Progress

Effect	Est.	S.E.
Fixed Effects		
Regression coefficients		
Intercept[a]	0.907***	.052
Frequency: ≥ 2 times a week	1.557***	.056
Frequency: ≥ 3 times a week	1.719***	.066
Frequency: ≥ 4 times a week	1.379***	.085
Frequency: 5 times a week	1.274***	.173
Length: > 10 minutes	0.666***	.049
Length: > 15 minutes	0.520***	.052
Length: > 20 minutes	0.381***	.072
Proportion from home: partly	-0.137	.078
Proportion from home: mostly	0.035	.083
Review lessons: ≥ 1 lesson	-0.933***	.053
Review lessons: ≥ 2 lessons	-0.963***	.080
Review lessons: ≥ 3 lessons	-1.002***	.103
Random Intercepts		
Variance components		
Weeks	0.845***	.026
Children	0.150*	.059
Schools	0.021	.016
R^2	.774	-

Note. Unstandardized regression coefficients are displayed. R^2 is the explained level 1 variance. All p-values are two-tailed.

*$p < .05$. **$p < .01$. ***$p < .001$.

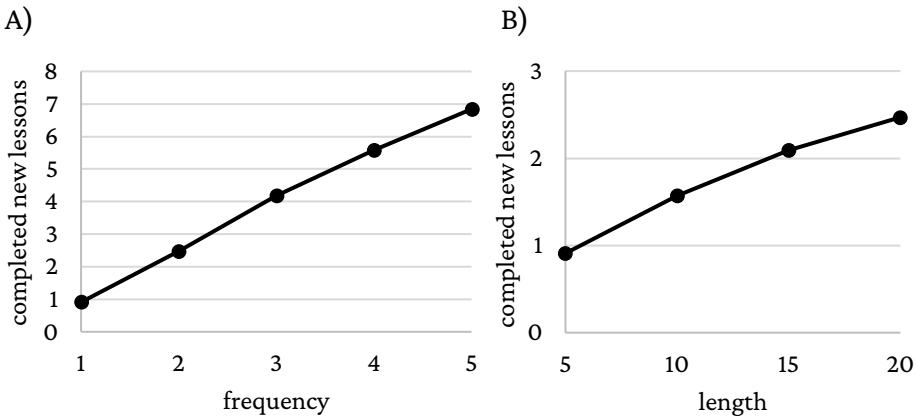
[a]The intercept parameter estimate represents the progress within the intervention for practicing once a week, less than 10 minutes, hardly at home, and without review lessons.

We checked whether the final model was similar for the subsample of children who could be considered at risk based on their scores on phonological awareness and/or letter knowledge at pretest. All fixed effects that were significant in the full sample remained significant in the subsample. Frequency, intensity, and the number review lessons explained a similar percentage of variance, i.e. 51%, 12%, and 34%

respectively (this was 46%, 14% and 36%). So, also in this subsample frequency was the most important predictor of progress.

Figure 2.2

Line Graphs for Predicting Progress by Frequency and Length of Intervention Sessions



Note. Graphs show how progress within the intervention (y-axis) increased by (A) frequency, i.e. the number of intervention sessions per week and (B) length, i.e. the average length of intervention sessions per week in minutes (x-axis). In both graphs, the first dot represents the intercept from the final multilevel model, i.e. practicing once a week for less than ten minutes. The following dots represent the intercept plus the regression coefficients for (A) ≥2, ≥3, ≥4, and 5 sessions per week, or (B) >10, >15, and >20 minutes per session. Relationships are more or less linear.

3.3 *Child Level Results*

The second goal was to examine–at the child level–the relation between dosage and intervention outcomes and whether this relation was mediated by progress within the intervention. On average, children practiced 14 out of 20 weeks, 2 times a week (while 3-4 times was prescribed), 14 minutes per session (corresponding to the prescribed 10-15 minutes), and finished around 41 new lessons during the intervention period, i.e. the second half of the second year of kindergarten (see Table 2.5). Letter knowledge and phonological awareness at pre-test were not correlated to frequency, length, and duration. This indicates that children with better pre-literacy skills at the start of the intervention practiced as much as children with poorer pre-literacy skills. Children's letter knowledge at post-test was correlated with progress within the intervention (.40), frequency (.25), and duration (.15), but not with length (-.08; see

Table S2.3 in the supplemental materials Chapter 2). Children's phonological awareness at post-test was correlated with progress within the intervention (.23), but not with frequency (.02), length (-.05), or duration (-.04; see Table S2.3 in the supplemental materials Chapter 2).

Table 2.5

Child Level Means and Standard Deviations of Predictors, Progress, and Literacy Outcomes

Variable	*M*	*SD*
Frequency[a]	2.17	.57
Length[b]	13.67	3.07
Duration[c]	14.22	2.52
Proportion from home[d]	.21	.20
Review lessons	0.60	0.58
Progress over intervention period	41.59	20.35
Letter knowledge		
Pre-test	6.72	3.37
Post-test	13.92	5.10
Phonological awareness		
Pre-test	21.20	8.30
Post-test[e]	8.40	3.82

[a]the number of intervention sessions per week; [b]average session duration per week, in minutes; [c]the number of intervention weeks; [d]proportion of intervention sessions at home; [e]score based on 3 out of 9 subtests.

The hypothesized mediation model was found to fit the data poorly. The chi-square test indicated that exact fit was rejected (χ^2 (12) = 38.95, p <.001), and RMSEA (.100) and CFI (.921) indicated poor fit. Based on the highest residual correlations and modification indices, and supported by previous research (Burgess & Lonigan, 1998; Muter et al., 2004), we added a direct effect from pre-test letter knowledge to post-test phonological awareness and from pre-test phonological awareness to post-test letter knowledge. The new model fitted the data. The chi-square test indicated that the model fitted the data (χ^2 (10) = 14.64, p = .15), and CFI (.986) and RMSEA (.045) indicated good model fit. This model was accepted as the final model.

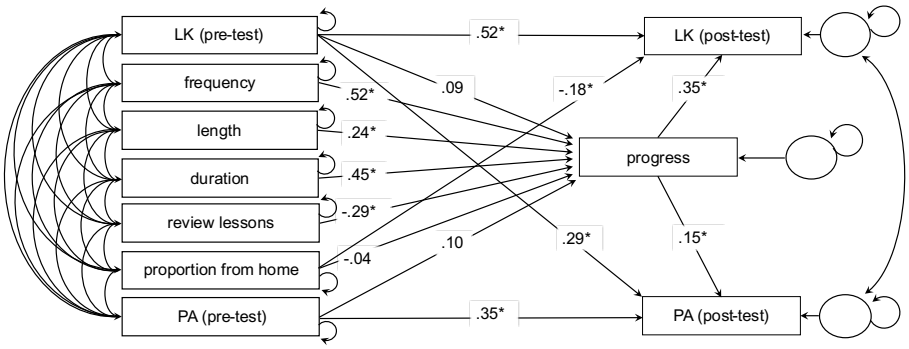
The final model is shown in Figure 2.3. Effects of dosage, i.e. frequency, length, and duration, on progress within the intervention were significant. Children made more progress, when completing more sessions per week (strong effect), longer

sessions (small effect), and more intervention weeks (moderate effect). With more progress within the intervention, children grew faster in letter knowledge (moderate effect) and phonological awareness (small effect). Indirect effects of frequency, length, and duration through progress on letter knowledge and phonological awareness were significant (see Table 2.6). In addition to separate effects, we also tested the combined indirect effects of frequency, length, and duration. The indirect effect of dosage on letter knowledge was moderate, and the effect of dosage on phonological awareness was small (see Table 2.6).

Again, we checked whether the final model was reproduced in the subsample of children who were at-risk according to their pretest scores on phonological awareness and letter knowledge. All paths that were significant in the full sample remained significant in the subsample. Among the three aspects of dosage, frequency had again the strongest indirect effect on literacy outcomes, followed by duration. Indirect effects of the three aspects of dosage together on letter knowledge and phonological awareness remained moderate and small respectively.

Figure 2.3

Final Mediation Model



Note. Circles represent residual variance. Standardized coefficients (β) are shown. LK = letter knowledge. PA = phonological awareness. Pre-test = the middle of the second year of kindergarten. Post-test = end of kindergarten.

$^{*}p < .05$.

Table 2.6

Indirect Effects of Dosage on Preliteracy Outcomes Mediated by Progress

Predictor	β	
	Letter knowledge post-test	Phonological awareness post-test
Frequency	.18*	.09*
Length	.08*	.04*
Duration	.15*	.08*
Dosage[a]	.42*	.20*

*p <.001. [a]sum of indirect effects of frequency, length, and duration

4 Discussion

The main goal of this study was to investigate whether variations in dosage, i.e. the amount of practice within an intervention, were associated with intervention outcomes. We distinguished three aspects of dosage: the number of sessions per week (frequency), the session length (length), and the number of intervention weeks (duration). Moreover, we analyzed the data at both the week level (within children) and the child level (between children).

On a weekly basis, the child's progress within the intervention varied strongly. About half of the variance in progress was explained by the number of sessions and the session length. Each extra intervention session and every five minutes of additional practice led to more progress within the intervention. The number of sessions was more important than the session length.

At the child level, we examined the association between dosage (i.e. frequency, length, *and duration*) and intervention outcomes and determined whether this relation was mediated by the progress of children within the intervention. Our findings indicated that children who practiced more, showed more growth in letter knowledge and phonological awareness during kindergarten. In addition, we found that these associations between dosage and literary outcomes were fully mediated by progress within the intervention. Similar to the week level, the number of sessions had the largest effect on progress, followed by the number of intervention weeks and the session length. Dosage had stronger associations with gains in letter knowledge than in phonological awareness.

4.1 *Relation Between Dosage and Intervention Outcomes*

Our study shows that, within the same intervention, children who spent more time on the intervention reach higher outcomes. This finding supports previous studies showing that treatment integrity in terms of dosage predicts intervention outcomes (for a review study, see van Dijk et al., 2023). Most previous studies included only one aspect of dosage. Our study shows that the relation between dosage and intervention outcomes differs for the various aspects of dosage. Specifically, we distinguished frequency and length as aspects of dosage at the week level and frequency, length, and duration at the child level. Frequency of intervention sessions was found to be more important than their length. This is in line with the theory of distributed practice showing that spaced practice is preferred over massed practice (Dunlosky et al., 2013). To our knowledge, our study is the first to show these principles are also applicable to literacy interventions.

This is not the first study that addressed how much intervention practice is necessary or sufficient. However, in previous studies dosage was measured only at the child level, while we measured dosage and progress also at the week level. This enabled us to determine a dose-response relationship. This relationship was found to be approximately linear for both frequency and length, indicating that weeks with more and longer intervention sessions were associated with more progress within the intervention (i.e. more completed new intervention lessons). The analyses at the week level also provided some support for the direction of the findings. The direction of the relation between dosage and progress cannot be determined at the child level, as children were not randomly assigned to different levels of dosage: do children who practice more proceed more rapidly through the program, or vice versa? However, at the week level, the children serve as their own control. Although probably not decisive, this provides some support for a causal effect of dosage on progress through the intervention as well as for an effect of dosage on literacy outcomes.

Another important finding of the current study is that the relation between dosage and literacy outcomes is fully mediated by children's progress within the intervention. It seems obvious that more time spent on the intervention is associated with more progress through the intervention. We are not aware of an earlier study in which such a mediation has been shown. We also found that the relation between dosage and progress is far from perfect. The distinction between dosage and progress is seldomly made. An exception is the recently reported study by Muñez et al. (2022), showing similar findings for a mathematics intervention. The imperfect relation suggests that variation in progress through the intervention is only partly due to

differences in dosage. Variation in progress may also be caused by children's ability to master the learning content. In an earlier study, Byrne et al. (2000) showed that how rapidly a child acquired phonological awareness within a preschool intervention program was highly predictive of later literacy outcomes. The additional influence of such learning ability on progress may explain why we found that progress is ultimately a stronger predictor of literacy outcomes than dosage. Thereby, the current study provides a strong case for including both dosage and progress to predict individual differences in intervention outcomes.

Similar to previous research on *Build!* in kindergarten (Regtvoort & van der Leij, 2007), we found larger associations between dosage and letter knowledge than between dosage and phonological awareness. This is not surprising, as letter knowledge was trained extensively, while only one phonological skill was trained, i.e. phoneme blending. Note that the instrument for phonological awareness used in this study (also) included other phonological skills. Thus, possibly the training of phoneme blending in *Build!* did not transfer well to other phonological skills. The question is whether it is more beneficial for reading development to extend the phonological training in the intervention. On the one hand, research has shown that phonological awareness training in kindergarten is worthwhile and has transfer effects to reading (Ehri et al., 2001; Suggate, 2010; 2016). On the other hand, there are studies showing that phonological training alone has only a small effect on word reading, at least in transparent languages such as Dutch (Bus & van IJzendoorn, 1999; Galuschka et al., 2014; van Otterloo et al., 2008). It is more effective to integrate phonological awareness training with reading, i.e. teaching children to manipulate phonemes using letters (Bus & van IJzendoorn, 1999; Ehri et al., 2001; Hatcher et al., 2004).

4.2 *Limitations and Suggestions for Future Research*

There are several limitations that need to be considered. First, we did not measure reading. The measurements in this study, measures of preliteracy skills, were most appropriate for children in kindergarten as Dutch at-risk children can hardly read any words at the end of kindergarten (Zijlstra et al., 2021). Moreover, the main goal of this study was to investigate how preliteracy intervention outcomes were influenced by dosage.

Second, schools used various procedures to select the children for the intervention. This resulted in slightly better literacy skills at the start of the intervention in the current sample, compared to previous RCTs in which only the 25% or 37% lowest

scoring children were included (Regtvoort et al., 2013; Zijlstra et al., 2021). We respected the selection of children by the schools because, in our view, this is part of everyday educational practice in the implementation of interventions by schools. Interestingly, the findings remained the same if we did select at-risk children in a similar way as the RCTs.

Third, findings of this study are based on one intervention and may not apply to all (literacy) interventions. However, it seems likely that findings do generalize to interventions that are comparable to *Build!*, that are interventions that are computer-based, adaptive, and provide the learning content in a fixed order. There are several (literacy) intervention with these characteristics, for example Lexia Reading Core5 (O'Callaghan et al., 2016) and GraphoGame (Saine et al., 2011) and, given the growing interest in computer-based interventions, it might be expected that more will follow. Fourth, this study was conducted in only two districts in the Netherlands. The sample contained relatively few urban schools, few non-Dutch children, and few children with low SES, which might have led to stronger intervention effects (Manz et al., 2010). More research is needed to investigate whether findings generalize across various school contexts.

Finally, the current study was not designed to show the effectiveness of the intervention. However, we did show that dosage had an additional effect on subsequent literacy outcomes when these abilities at an earlier time were taken into account. Following the logic of longitudinal research, such an additional effect of a factor, in this case dosage, on the growth of preliteracy skills is often taken to support a causal effect (Gollob & Reichardt, 1987), here an effect of the intervention on preliteracy outcomes. Nevertheless, without a control group it is impossible to show that such an effect is larger than the growth of preliteracy skills in children that did not follow the intervention.

4.3 *Practical Implications*

Schools are encouraged to use evidence-based interventions to reduce reading problems. An important implication of our study is that practice matters. Dosage, that is the amount of practice within the intervention, can determine to what extent children benefit from the intervention. Instead of fewer longer sessions, it seems favorable to have multiple shorter sessions per week over a longer period of time. It is helpful to provide children with as many sessions per week as possible and to extend the duration of the intervention, i.e. increasing the number of intervention weeks. The association between duration and intervention outcomes might have been

overlooked in early-literacy interventions so far, as prescribed practice mostly focuses on the number and length of intervention sessions per week.

The finding that progress within the intervention (i.e. the number of new intervention lessons completed) fully mediated the relationship between dosage (i.e. time spent on the intervention) and literacy outcomes, indicates that schools should also keep track of children's progress within the intervention, in addition to monitoring the number of intervention sessions per week. As such, schools can identify children who do not make enough progress and adjust the number and length of intervention sessions accordingly.

4.4 *Conclusion*

There are few studies that have distinguished different aspects of dosage and have considered dose-response relationships within an intervention. In the current study three aspects of dosage were included (duration of the intervention in weeks, frequency of intervention sessions, and session length) and dosage itself was distinguished from progress within the intervention (the number of new intervention lessons completed). Moreover, dose-response relationships were considered within children, on a week-to-week-basis, as well as between children. Major findings were, firstly, that dosage and progress within the intervention are related but not similar. One reason was that some children had to take more review lessons and thereby progressed more slowly through the program. They thus needed a larger dose to make the same progress. Secondly, we compared the three aspects of dosage and found that frequency of intervention sessions matters most. The number of sessions had a stronger relation with children's weekly progress within the intervention than session length. Similarly, at the child level, frequency of intervention sessions had a stronger relation with children's progress and intervention outcomes than session length and the duration of the intervention in weeks. Thirdly, the results showed that on a week-to-week basis the dose-response relationship was approximately linear. Children made more progress within the intervention in weeks with more and longer intervention sessions. Fourthly, progress within the intervention completely mediated the relationship between dosage and children's literacy outcomes. These findings shed light on earlier inconsistent findings on the relation between dosage and intervention outcomes by showing that dosage is indirectly related to intervention outcomes via progress within the intervention and that the strength of the relation between dosage and intervention outcomes is different for the three aspects of dosage.

Supplemental Materials Chapter 2

Table S2.1

Pearson's Correlations Between Frequency, Length, Location, Progress, Lessons and Review Lessons at Week Level (Within and Between Children)

	1	2	3	4	5
1. Progress[a]	-	.62*	.25*	.24*	-.31*
2. Frequency[b]	.64*	-	.02	.34*	.19*
3. Length[c]	.31*	.08*	-	.29*	.19*
4. Proportion from home[d]	.22*	.24*	.20*	-	.12
5. Review lessons	-.19*	.21*	.05*	.08*	-

Note. Below diagonal: correlations within children, above diagonal: correlations between children.

[a]the number of new lessons completed in a week; [b]the number of intervention sessions per week; [c]average session duration per week, in minutes; [d]proportion of intervention sessions at home.

*$p < .01$.

Table S2.2
Multilevel Model Results for Progress per Week

	Model 1		Model 2	
	Est.	S.E.	Est.	S.E.
	Fixed effects[a]			
Week level				
Intercept	3.02***	0.12	1.31***	.10
Frequency: ≥ 2 times a week	-	-	1.44***	.07
Frequency: ≥ 3 times a week	-	-	1.70***	.07
Frequency: ≥ 4 times a week	-	-	1.25***	.11
Frequency: 5 times a week	-	-	1.28***	.22
Length: 10-15 minutes	-	-	-	-
Length: 15-20 minutes	-	-	-	-
Length: >20 minutes	-	-	-	-
Proportion from home: partly	-	-	-	-
Proportion from home: mostly	-	-	-	-
Child level				
Frequency	-	-	-	-
Length	-	-	-	-
Proportion from home	-	-	-	-
Review lessons	-	-	-	-
School level				
Frequency	-	-	-	-
Length	-	-	-	-
Proportion from home	-	-	-	-
Review lessons	-	-	-	-
$\Delta\chi^{2}$ [b]	-		1861.84***	
Δdf [b]			4	
ΔR^{2} [b]	-		.46	
R^{2} [c]			.46	

[a]Unstandardized regression coefficients are displayed. [b]compared to the previous model. [c]explained 1-level variance compared to the previous model. [d]explained 1-level variance compared to Model 1.
[ns]$p \geq .05$. [*]$p < .05$. [**]$p < .01$. [***]$p < .001$.

Table S2.2 (continued)
Multilevel Model Results for Progress per Week

	Model 3		Model 4	
	Est.	S.E.	Est.	S.E.
	Fixed effects[a]			
Week level				
Intercept	0.60***	0.10	1.57***	0.10
Frequency: ≥ 2 times a week	1.39***	0.06	1.44***	0.07
Frequency: ≥ 3 times a week	1.62***	0.07	1.70***	0.07
Frequency: ≥ 4 times a week	1.22***	0.10	1.24***	0.10
Frequency: 5 times a week	1.23***	0.21	1.23***	0.21
Length: 10-15 minutes	0.69***	0.07	0.70***	0.07
Length: 15-20 minutes	0.66***	0.07	0.66***	0.07
Length: >20 minutes	0.38***	0.09	0.38***	0.09
Proportion from home: partly	-	-	-0.16ns	0.08
Proportion from home: mostly	-	-	0.17ns	0.11
Child level			-	-
Frequency	-	-	-	-
Length	-	-	-	-
Proportion from home	-	-		
Review lessons	-	-	-	-
School level			-	-
Frequency	-	-	-	-
Length	-	-	-	-
Proportion from home	-	-	-	-
Review lessons	-	-		
$\Delta\chi^{2\,b}$	455.89***		8.25*	
Δdf^{b}	3		2	
$\Delta R^{2\,b}$	.14		.00	
$R^{2\,c}$	.54		.54	

[a]Unstandardized regression coefficients are displayed. [b]compared to the previous model. [c]explained 1-level variance compared to the previous model. [d]explained 1-level variance compared to Model 1.
ns$p \geq .05$. *$p < .05$. **$p < .01$. ***$p < .001$.

Table S2.2 (continued)
Multilevel Model Results for Progress per Week

	Model 5		Model 6	
	Est.	S.E.	Est.	S.E.
	Fixed effects[a]			
Week level				
Intercept	0.88***	0.08	0.84***	0.21
Frequency: ≥ 2 times a week	1.63***	0.05	1.63***	0.05
Frequency: ≥ 3 times a week	1.79***	0.06	1.79***	0.06
Frequency: ≥ 4 times a week	1.37***	0.08	1.37***	0.08
Frequency: 5 times a week	1.46***	0.16	1.46***	0.16
Length: 10-15 minutes	0.72***	0.05	0.72***	0.05
Length: 15-20 minutes	0.63***	0.05	0.63***	0.05
Length: >20 minutes	0.47***	0.07	0.47***	0.07
Proportion from home: partly	-0.07ns	0.07	-0.07ns	0.07
Proportion from home: mostly	0.14ns	0.09	0.14ns	0.09
Review lessons: ≥ 1 lesson	-0.94***	0.06	-0.94***	0.06
Review lessons: ≥ 2 lessons	-0.97***	0.08	-0.97***	0.08
Review lessons: ≥ 3 lessons	-1.18***	0.10	-1.18***	0.10
Child level				
Frequency			0.02ns	0.09
Length	-	-		
Proportion from home	-	-	-	-
Review lessons	-	-	-	-
School level	-	-	-	-
Frequency				
Length	-	-	-	-
Proportion from home	-	-	-	-
Review lessons	-	-	-	-
$\Delta\chi^{2}$ [b]	1345.69***		0.04ns	
Δdf [b]	3		1	
ΔR^{2} [b]	.36		.00	
R^{2} [c]	.71		.71	

[a]Unstandardized regression coefficients are displayed. [b]compared to the previous model. [c]explained 1-level variance compared to the previous model. [d]explained 1-level variance compared to Model 1.
ns$p \geq .05$. *$p < .05$. **$p < .01$. ***$p < .001$.

Table S2.2 (continued)
Multilevel Model Results for Progress per Week

	Model 7		Model 8	
	Est.	S.E.	Est.	S.E.
	Fixed effects[a]			
Week level				
Intercept	0.89***	0.08	0.92***	0.09
Frequency: ≥ 2 times a week	1.63***	0.05	1.63***	0.05
Frequency: ≥ 3 times a week	1.79***	0.06	1.79***	0.06
Frequency: ≥ 4 times a week	1.37***	0.08	1.37***	0.08
Frequency: 5 times a week	1.46***	0.16	1.46***	0.16
Length: 10-15 minutes	0.70***	0.05	0.72***	0.05
Length: 15-20 minutes	0.62***	0.05	0.63***	0.05
Length: >20 minutes	0.46***	0.07	0.47***	0.07
Proportion from home: partly	-0.08ns	0.07	-0.05ns	0.07
Proportion from home: mostly	0.14ns	0.09	0.16ns	0.09
Review lessons: ≥ 1 lesson	-0.94***	0.06	-0.94***	0.06
Review lessons: ≥ 2 lessons	-0.97***	0.08	-0.97***	0.08
Review lessons: ≥ 3 lessons	-1.19***	0.10	-1.18***	0.10
Child level				
Frequency				
Length	0.04**	0.02		
Proportion from home			-0.22ns	0.26
Review lessons				
School level				
Frequency				
Length				
Proportion from home				
Review lessons				
$\Delta\chi^{2}$ [b]	7.10ns		0.73ns	
Δdf [b]	1		1	
ΔR^{2} [b]	.00		.00	
R^{2} [c]	.71		.71	

[a]Unstandardized regression coefficients are displayed. [b]compared to the previous model. [c]explained 1-level variance compared to the previous model. [d]explained 1-level variance compared to Model 1.
ns$p \geq .05$. *$p < .05$. **$p < .01$. ***$p < .001$.

Table S2.2 (continued)
Multilevel Model Results for Progress per Week

	Model 9		Model 10	
	Est.	S.E.	Est.	S.E.
	Fixed effects[a]			
Week level				
Intercept	0.91***	0.09	1.09***	0.35
Frequency: ≥ 2 times a week	1.63***	0.05	1.63***	0.05
Frequency: ≥ 3 times a week	1.79***	0.06	1.79***	0.06
Frequency: ≥ 4 times a week	1.37***	0.08	1.37***	0.08
Frequency: 5 times a week	1.46***	0.16	1.46***	0.16
Length: 10-15 minutes	0.72***	0.05	0.72***	0.05
Length: 15-20 minutes	0.63***	0.05	0.63***	0.05
Length: >20 minutes	0.47***	0.07	0.47***	0.07
Proportion from home: partly	-0.07ns	0.07	-0.07ns	0.07
Proportion from home: mostly	0.15ns	0.09	0.15ns	0.09
Review lessons: ≥ 1 lesson	-0.94***	0.06	-0.94***	0.06
Review lessons: ≥ 2 lessons	-0.97***	0.09	-0.97***	0.08
Review lessons: ≥ 3 lessons	-1.18***	0.10	-1.18***	0.10
Child level				
Frequency				
Length				
Proportion from home				
Review lessons	-0.05ns	0.09		
School level				
Frequency			-0.10ns	0.16
Length				
Proportion from home				
Review lessons				
$\Delta\chi^{2}$ [b]	0.39ns		0.73ns	
Δdf [b]	1		1	
ΔR^{2} [b]	.00		.00	
R^{2} [c]	.71		.71	

[a]Unstandardized regression coefficients are displayed. [b]compared to the previous model. [c]explained 1-level variance compared to the previous model. [d]explained 1-level variance compared to Model 1.
ns$p \geq .05$. *$p < .05$. **$p < .01$. ***$p < .001$.

Table S2.2 (continued)
Multilevel Model Results for Progress per Week

	Model 11		Model 12	
	Est.	S.E.	Est.	S.E.
	Fixed effects[a]			
Week level				
Intercept	0.89***	0.08	1.02***	0.12
Frequency: ≥ 2 times a week	1.63***	0.05	1.62***	0.05
Frequency: ≥ 3 times a week	1.79***	0.06	1.78***	0.06
Frequency: ≥ 4 times a week	1.37***	0.08	1.37***	0.08
Frequency: 5 times a week	1.46***	0.16	1.46***	0.16
Length: 10-15 minutes	0.71***	0.05	0.72***	0.05
Length: 15-20 minutes	0.63***	0.05	0.63***	0.05
Length: >20 minutes	0.47***	0.07	0.47***	0.07
Proportion from home: partly	-0.07ns	0.07	-0.05ns	0.07
Proportion from home: mostly	0.14ns	0.09	0.17ns	0.09
Review lessons: ≥ 1 lesson	-0.94***	0.06	-0.94***	0.06
Review lessons: ≥ 2 lessons	-0.97***	0.08	-0.97***	0.08
Review lessons: ≥ 3 lessons	-1.18***	0.10	-1.18***	0.10
Child level				
Frequency				
Length				
Proportion from home				
Review lessons				
School level				
Frequency				
Length	0.04ns	0.03		
Proportion from home			-0.70ns	0.42
Review lessons				
$\Delta\chi^{2\,b}$	2.31ns		2.81ns	
Δdf^{b}	1		1	
$\Delta R^{2\,b}$	.00		.00	
$R^{2\,c}$	.71		.71	

[a]Unstandardized regression coefficients are displayed. [b]compared to the previous model. [c]explained 1-level variance compared to the previous model. [d]explained 1-level variance compared to Model 1.
ns$p \geq .05$. *$p < .05$. **$p < .01$. ***$p < .001$.

Table S2.2 (continued)
Multilevel Model Results for Progress per Week

	Model 13	
	Est.	S.E.
Fixed effects[a]		
Week level		
Intercept	0.94***	0.13
Frequency: ≥ 2 times a week	1.63***	0.05
Frequency: ≥ 3 times a week	1.79***	0.06
Frequency: ≥ 4 times a week	1.37***	0.08
Frequency: 5 times a week	1.46***	0.16
Length: 10-15 minutes	0.72***	0.05
Length: 15-20 minutes	0.63***	0.05
Length: >20 minutes	0.47***	0.07
Proportion from home: partly	-0.07ns	0.07
Proportion from home: mostly	0.15ns	0.09
Review lessons: ≥ 1 lesson	-0.94***	0.06
Review lessons: ≥ 2 lessons	-0.97***	0.08
Review lessons: ≥ 3 lessons	-1.18***	0.10
Child level		
Frequency		
Length		
Proportion from home		
Review lessons		
School level		
Frequency		
Length		
Proportion from home		
Review lessons	-0.10ns	0.17
$\Delta\chi^{2}$ [b]	0.36ns	
Δdf [b]	1	
ΔR^{2} [b]	.00	
R^{2} [c]	.71	

[a]Unstandardized regression coefficients are displayed. [b]compared to the previous model. [c]explained 1-level variance compared to the previous model. [d]explained 1-level variance compared to Model 1.
ns$p \geq .05$. *$p < .05$. **$p < .01$. ***$p < .001$.

Table S2.3

Pearson's Correlations Between Frequency, Length, Duration, Progress, Review Lessons, and Literacy Outcomes at Child level

	2	3	4	5	6	7	8	9	10
1. Frequency[a]	.01	.38**	.20**	.34**	.63**	.09	.25**	-.03	.02
2. Length[b]	-	-.12	.17**	.29**	.13	.04	-.08	.11	-.05
3. Duration[c]	-	-	.04	.02	.57**	-.10	-.15*	-.08	-.04
4. Proportion from home[d]	-	-	-	.15*	.24**	-.09	.06	-.02	.02
5. Review lessons	-	-	-	-	-.26**	-.19**	-.20**	-.25**	-.23**
6. Progress[e]	-	-	-	-	-	.13	.40**	.12	.23**
Letter knowledge									
7. pre-test	-	-	-	-	-	-	.52**	.18*	.36**
8. post-test	-	-	-	-	-	-	-	-.05	.32**
Phonological awareness									
9. pre-test	-	-	-	-	-	-	-	-	.43**
10. post-test	-	-	-	-	-	-	-	-	-

[a]the number of intervention sessions per week; [b]average session duration per week, in minutes; [c]the number of intervention sessions per week; [d]proportion of intervention sessions at home; [e]the number of new lessons completed in the intervention period

*$p < .05$. **$p < .01$.

CHAPTER 3

A School-Based Implementation of an Early-Literacy Intervention: Relations Among Dosage, Familial Risk, Parental Education, and Reading Acquisition

Abstract

When schools implement an evidence-based intervention, the level of treatment integrity affects intervention outcomes. This study centered on one aspect of treatment integrity: dosage (time spent on the intervention). Schools implemented the computer-based early-literacy intervention *Build!*. It was examined whether natural variations in dosage were associated with literacy outcomes, through progress within the intervention (the number of new intervention lessons completed). Furthermore, it was investigated whether parental education and familial risk for dyslexia were related to children's outcomes directly, and/or indirectly through dosage and progress within the intervention. Children with poor preliteracy skills (n = 396, 50 schools) received the intervention from kindergarten through first grade. Findings show that a higher dose was associated with more progress within the intervention, and, in turn with more letter knowledge and better phonological skills at the end of kindergarten, better word reading accuracy at the beginning of first grade, and better word reading fluency, but not better word reading accuracy in the middle of first grade. After the middle of first grade, there were no additional effects of dosage on reading development. Parental education was not directly nor indirectly related to literacy outcomes. Children with familial risk reached lower literacy outcomes and made less progress within the intervention, despite the same dose. Results emphasize the importance of dosage in school-based implementations. In computer-based interventions the relation between dosage and children's outcomes seems to be mediated by progress within the intervention. Consequently, this progress could be used to adjust the dose to children's needs.

van der Weijden, F. A., 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 [Manuscript submitted for publication]. Department of Child Development and Education, University of Amsterdam.

1 Introduction

Between 3 and 10% of all children has severe difficulties to acquire sufficient reading skills within primary school (Fluss et al., 2009; Snowling, 2013). These children need more support in order to learn to read. To provide adequate support, schools can choose an intervention from a large number of reading programs (Scammacca, Vaughn, Roberts, Wanzek, Torgesen, & Instruction, 2007; Suggate, 2016). The effects of these programs have mostly been evaluated in *efficacy* studies, often randomized controlled trials (RCTs), examining the potential of interventions under favorable conditions (Earle et al., 2013; O'Donnell, 2008; Streiner, 2002). Relatively few interventions have subsequently been evaluated in studies in the field, when interventions are implemented by schools and in natural school settings. In this study, we focused on a computer-based intervention for the prevention of literacy difficulties, *Build!* (in Dutch: *Bouw!*). This is a two-year intervention starting in the second year of kindergarten and continuing until mid-Grade 2. Beneficial effects of the intervention on reading and spelling have been demonstrated in two efficacy studies (Regtvoort et al., 2013; Zijlstra et al., 2021). In this longitudinal study we examined whether the use of the intervention in a natural setting and implemented by schools was related to growth in preliteracy and reading skills.

A key issue in the implementation of an intervention is whether *treatment integrity* can be realized, i.e. whether the intervention is implemented as prescribed (O'Donnell, 2008). Insufficient treatment integrity is assumed to decrease the effect of an intervention. In efficacy studies, treatment integrity is often warranted by a high involvement of the researchers who tightly monitor the quantity of intervention sessions, thereby ensuring an overall high dosage (Earle et al., 2013; Zijlstra et al., 2021). However, sufficient treatment integrity might not always be reached when an intervention is scaled-up and schools become responsible for its proper implementation. For example, a large scale study by Stein et al. (2008) showed that schools do not always receive adequate training and support to implement an intervention, resulting in lower levels of treatment integrity which, in turn, can dampen intervention effects. Because the current study involved a computer-based intervention, we focused on one particular dimension of treatment integrity: dosage. Dosage is the amount of intervention provided, also sometimes referred to as exposure (Capin et al., 2018). The first question of the current study was whether dosage during the intervention was associated with children's reading development.

A second question concerned the child characteristics that are associated with the response to the intervention. Such characteristics become particularly relevant

when interventions are scaled-up and the population included in the intervention becomes more heterogeneous. In a meta-analysis, Stuebing et al. (2015) showed that reading level at the start of an intervention was the best predictor of response to the intervention and that other reading-related-cognitive skills did not add substantially to this prediction. In the current study we considered two family characteristics of the child: familial risk for dyslexia and educational level of the parents. Family risk for dyslexia is important as the current intervention started in kindergarten, a period in which reading ability is hard to determine reliably. Several studies have shown that familial risk for dyslexia predicts later reading ability when various precursors of reading have been taken into account (van Bergen et al., 2015; van Viersen et al., 2018). Educational level of parents was included because an earlier study suggested that the intervention might be less effective for children from families with a lower socio-economic status (Zijlstra et al., 2021).

1.1 *Dosage and Intervention Outcomes*

Dosage is mostly defined as the number of hours of intervention (Wanzek et al., 2016). Many interventions have a prescribed frequency and intensity of sessions (Wolgemuth et al., 2014). When schools are responsible for the implementation, local conditions will differ and schools might not always be able to fully implement the intervention as intended (Savage et al., 2010; Stein et al., 2008), resulting in lower dosage, and presumably lower intervention outcomes.

The effect of dosage on intervention outcomes has been evaluated in several meta-analyses. Surprisingly, in these meta-analyses dosage is generally hardly associated with intervention effectiveness: the number of sessions, hours of intervention and weeks of intervention were not significantly associated with intervention outcomes (Tran et al., 2011; Wanzek et al., 2013; Wanzek & Vaughn, 2007). However, it should be noted that in these meta-analyses, conclusions about dosage are based on the comparison of *a variety of interventions*. Interventions did thus not only differ in dosage, but also in other ways. As such, the finding that interventions with varying doses did not differ in effect size may be explained by other differences among interventions that may or may not be related to dosage. Imagine, for example, that dosage is positively related to intervention outcomes and group size; the latter having a negative relation with intervention outcomes. Then, there are two competing processes, i.e. a positive direct effect and a negative indirect effect on intervention outcomes, which can result in a total effect that is close to zero (Shrout & Bolger, 2002). So, the relation between dosage and intervention outcomes can better be studied among

children following *the same intervention*, receiving larger and smaller doses of this intervention.

Only a few studies on reading interventions have directly compared children's outcomes when receiving larger or smaller doses of the same intervention (Capin et al., 2018). Findings vary across studies. In an RCT on the early literacy intervention *TAILS*, Al Otaiba et al. (2005) compared three groups of kindergartners: practicing four days a week, practicing two days a week, and not practicing at all. The group that practiced four days a week showed stronger gains in word reading than the group that practiced two days a week and children in the control group. In contrast, Wanzek and Vaughn (2007) found that dosage did not matter. In their RCT on a reading intervention for first-grade low responders, children who practiced once and twice a day showed stronger improvement in reading than students who did not practice at all, but there was no difference between practicing once or twice a day. Small effects of dosage were shown by Wolgemuth et al. (2014) and Zijlstra et al. (2014), who treated dosage as a continous variable and related it to reading outcomes of their computer-based literacy interventions. Across diverse reading outcomes, Zijlstra et al. found that dosage predicted 8-24% of the variance in intervention outcomes and Wolgemuth et al. 0-11%. In sum, studies on the relation between dosage and intervention outcomes show mixed findings. Importantly, these were all small-scaled efficiacy studies, conducted in 4 to 14 schools and with strong researcher guidance. Under these conditions, the variety in dosage might be suppressed, making it difficult to detect a relation between dosage and intervention outcomes. On a larger scale, when the intervention is implemented by schools, the variety in dosage might be larger, which could result in a stronger relation with intervention outcomes.

As said, dosage, time spent on the intervention, is generally regarded as an indicator of the amount of exposure to the intervention (Capin et al., 2018). However, dosage does not fully capture exposure to the content of the intervention. Children with the same amount of dosage can differ in the number of exercises or lessons that are successfully completed. Thus dosage is not a pure measure of exposure (see Muñez et al., 2022 for a similar argument). When literacy interventions are not computer-based, it is difficult to acquire information about what an individual child has done within the time spent on the intervention. However, in most computer-based interventions, as in the current study, a distinction can be made between dosage and progress within the intervention, that is the number of lessons or program parts that a child has finished during a given period. The information on progress is often automatically registered in the logs of these computer programs.

Obviously, progress will be substantially affected by dosage, but the strength of this empirical relationship is, to our knowledge, largely unknown and might even depend on the intervention. Moreover, we expect that the relation is not perfect. Progress will likely be affected by how easily a child acquires the skills that the intervention aims to foster (e.g. van Uittert et al., 2022). With respect to dosage it is hard to predict how it is affected by the capabilities of the child. Children with relative better reading(-related) skills at pretest might practice more. But if the intervention, as in the current study, is mostly done in school with the help of a tutor, the poorer children at pretest might practice equally often or even more. In case progress is partly determined by the ability of the child to learn to read, and dosage is not, a further expectation is that progress might have a stronger relation with the development of reading, the outcome of the intervention. Moreover, the effect of dosage is probably indirect, that is the relation between dosage and reading outcomes is fully mediated by progress within the intervention.

1.2 *Family Characteristics and Intervention Outcomes*

The second goal of this study was to investigate the relation between two family characteristics and the response to the intervention. The first characteristic was familial risk for dyslexia. Children with familial risk have one or more family members with severe reading problems. These children are three to four times more likely to develop reading problems than their peers from non-risk families (Snowling & Melby-Lervåg, 2016). Children with familial risk tend to have lower preliteracy skills in kindergarten and more difficulty in attaining phonological skills and reading skills between kindergarten and second grade (Snowling & Melby-Lervåg, 2016; van Viersen et al., 2018).

In only a few studies, it has been examined how children with a familial risk respond to early literacy interventions. Hindson et al. (2005) found that children with familial risk for dyslexia, receiving the early literacy intervention *Sound Foundations*, showed lower gains in phoneme identification and letter knowledge during preschool than children without familial risk, which in turn affected reading and spelling in kindergarten. In contrast, Zijlstra et al. (2021) found that children with and without familial risk made similar gains in word reading fluency during first and second grade. However, in the study of Hindson et al., children with familial risk received a similar dose of the intervention as children without familial risk, while, in the study of Zijlstra et al., children with familial risk received 50% more sessions than children not at-risk. Thereby, the conclusion in both studies is actually the same:

children with familial risk need more practice than children without familial risk to reach a similar level of reading. Put differently, children with a familial risk for dyslexia proceed more slowly through the intervention. It remains to be seen whether children with familial risk for dyslexia are provided with an extra dose in a school-based implementation of a reading intervention.

The second family characteristics considered in this study, was parental education. Children whose parents have a lower educational level tend to have fewer books at home and their parents are on average less involved in literacy activities (Hamilton et al., 2016; Hemmerechts et al., 2017; Phillips & Lonigan, 2009). A poorer home literacy environment in turn is associated with lower oral language skills and poorer emergent decoding in preschool and kindergarten, as well as lower word reading skills in first grade (Hamilton et al., 2016; Petrill et al., 2005; Storch & Whitehurst, 2001).

To the best of our knowledge, parental education has not been studied in relation to effectiveness of reading interventions, but a related factor, i.e. parental income, has. A meta-analysis on summer reading interventions shows larger benefits for children whose parents have a low income than for children whose parents have a middle or high income (Kim & Quinn, 2019). Possibly, these children had missed learning opportunities (at home) and supplemental interventions can compensate for that. However, there are also studies that show that literacy interventions are less effective for children whose parents have a lower income (Manz et al., 2010). These are interventions where parents are involved in the delivery of the intervention. Possibly, these parents experience more barriers in engaging in reading-related activities with their children, such as limited time, resources, or knowledge (Wang et al., 2016). Such barriers can result in a lower dosage and, in turn, lower intervention effects.

1.3 *The Current Study*

The first aim of the current study was to examine whether dosage (time spent on the intervention) is related to the development of reading from the middle of the second year in kindergarten until the end of second grade, through progress within a literacy intervention (the number of new intervention lessons completed). As there is no dosage when children do not implement the intervention, the current study did not include a no-intervention control group. In line with other longitudinal studies (de Jong & van der Leij, 1999; Wagner et al., 1993), we examined whether dosage has an *additional effect* on the growth of preliteracy and reading skills after prior levels of these skills, that is autoregressive effects, have been taken into account. To make a

potential causal interference of such an effect, two principles of causality need to be taken into account (Gollob & Reichardt, 1987). First, a variable is only predicted by factors earlier in time, resulting in a longitudinal design with a time lag between dosage and intervention outcomes. Second, a variable is always predicted by prior levels of that variable, requiring longitudinal models that include a pre-test and post-test and can therefore control for autoregressive effects (Cole & Maxwell, 2003). When, in such models, an additional effect of dosage on intervention outcomes is found, this supports the effect of the intervention.

With respect to this first goal, this study should partly be regarded as a follow-up of a previous study by van der Weijden et al. (2024b) on the implementation of *Build!* in natural school settings. This previous study was solely focused on the association of dosage of the intervention with the development of preliteracy skills during kindergarten. It included a twofold detailed analysis of the effects of dosage. Firstly, three aspects of dosage were distinguished, number of sessions per week, length of sessions, and duration of the intervention in weeks. Secondly, the data were analyzed both within children, that is from week to week, and at the child level. The main results were that frequency and duration of the intervention were more strongly associated with growth in preliteracy skills than session length, and that, as predicted, the association of dosage with growth in preliteracy skills was fully mediated by progress within the intervention.

In the current study, we had a larger sample of kindergartners (369 instead of 226 children) and followed the children during a longer period. As a result, we could examine the associations between dosage, progress, and word reading accuracy and fluency. Moreover, dosage and progress were assessed during three periods. The first period was, as in our earlier study (van der Weijden et al., 2024b), the second half of the second kindergarten year. In this period, the intervention was mostly focused on the acquisition of letter knowledge and phonological awareness, as well as on decoding monosyllabic words with the learned letters. In this period, letter knowledge and phonological awareness were assessed. The second period concerned the first half year of first grade. In this period, the intervention is mainly concerned with further letter-sound learning and decoding monosyllabic words, both accurately and fluently. During this period, word reading accuracy and by the end also fluency were assessed. In the third period, the intervention focused on letter clusters and decoding words with two syllables, both accurately and fluently. In this period, reading fluency was measured. In each period we examined whether dosage and progress explain additional variance in outcome measures after the initial level of these skills was

controlled. For each period, however, we focus on slightly different aspects of reading development.

The second aim of the current study was to examine the association of two family characteristics, familial risk for dyslexia and parental education, with dosage and progress, and with the development of reading skills. For children with familial risk for dyslexia we expected lower preliteracy skills in kindergarten and smaller growth in word reading throughout second grade (van Viersen et al., 2018), as well as less progress within the intervention (Hindson et al., 2005; Zijlstra et al., 2021). With respect to dosage, expectations could go both ways. Children with familial risk might show a lower dosage, as they might avoid reading-related activities, especially when they are practicing at home (Petrill et al., 2005). However, following Zijlstra et al. (2021) children with familial risk might receive extra support to maintain sufficient progress through the program. For children whose parents have a lower educational level, we expected lower preliteracy skills at the beginning of kindergarten (Hamilton et al., 2016) and a lower dosage (Hemmerechts et al., 2017; Phillips & Lonigan, 2009).

2 Method

2.1 *Participants*

The children in this study were part of a larger sample of children who were selected for the intervention: 1150 Dutch children from 55 schools, located in two Districts in the Netherlands. The sample consisted of three cohorts, starting in kindergarten in three subsequent school years (i.e. 2017-2018, 2018-2019, and 2019-2020). Schools in District 1 joined the project from the beginning and all children were followed from kindergarten onwards. Schools in District 2 joined the project in school year 2019-2020. Children in kindergarten as well as in Grade 1 and Grade 2 were followed from that moment onwards. Thus, only Cohort 3 was followed from kindergarten onwards in this district. In both districts, Cohort 1 and 2 were followed until the middle of second grade, and because the project ended in July 2021, Cohort 3 could be only followed until the end of first grade.

Children were selected for the intervention by the schools. The two districts used different selection procedures. District 1 followed a procedure including two screening waves. The first wave took place in October, near the beginning of second kindergarten year. Children's letter knowledge and phonological awareness were measured (see Measurements). During the next eight weeks, the 30% lowest scoring children were provided with additional instruction in small groups within the classroom. The second wave took place in January. Only the children who scored below-

average on the first wave participated, including all children who received 8 weeks of additional instruction. Based on the second wave, children were selected for the intervention program, if they had: (1) low score(s) on either phonological awareness (≤25th percentile) or letter knowledge (≤6 letters), or (2) below-average scores on both phonological awareness (26-50th percentile) and letter knowledge (7-8 letters). Among children with informed consent in District 1 (see Figure 3.1), 50% of the children included in the intervention met these criteria, while 36% had average or above average scores, and 14% missed the second wave but was still selected for the intervention.

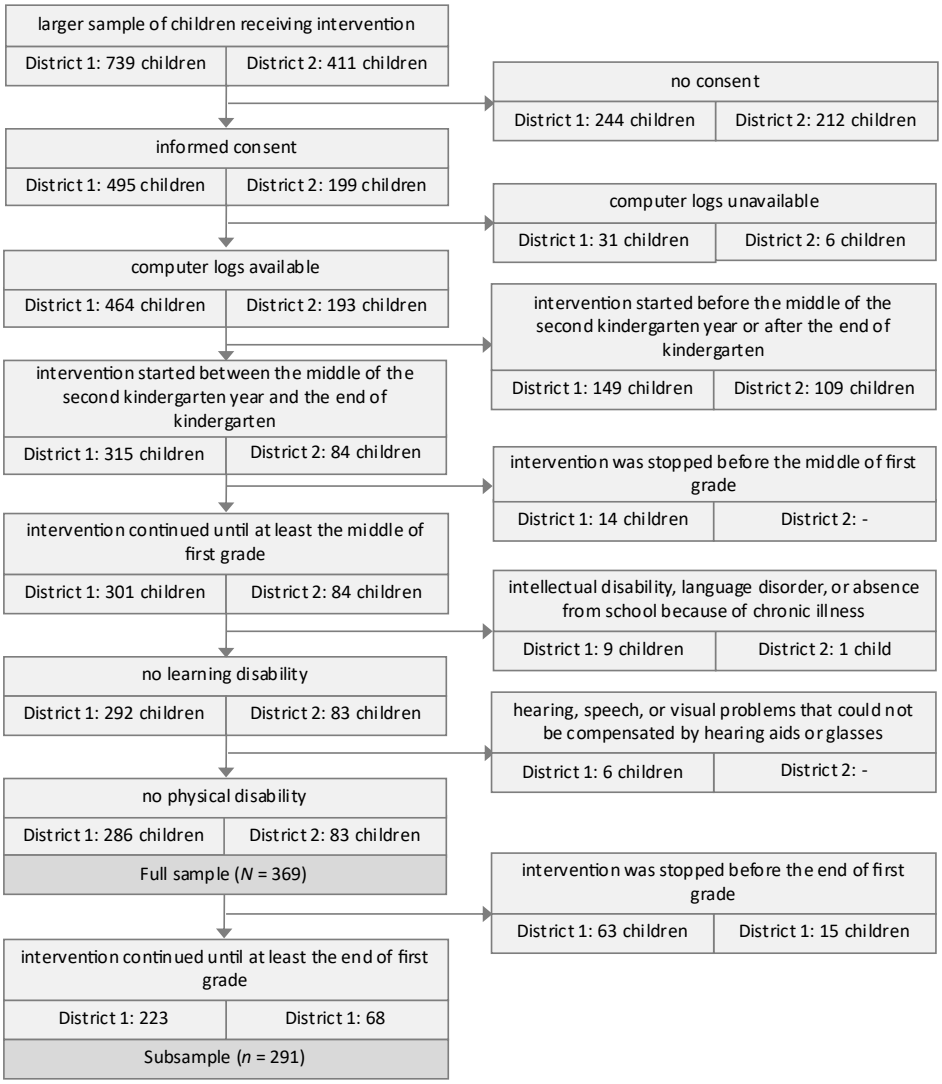
In District 2 there was more variation in selection procedures across schools. Some schools based the selection on the national protocol for dyslexia (van Druenen & Koning, 2017), including a similar response-to-intervention procedure as in District 1, while other schools based the selection on teacher observations, curriculum-based measurements, and information from the parents about dyslexia in the family. To have an indication of children's skills before the start of the intervention, schools were asked to administer the same tests as in District 1 for letter knowledge and phonological awareness (see Measures) just before the child would start the intervention. The pre-test could only be administered within the third cohort. Among the children with a pre-test, 62% met the selection criteria of District 1.

Figure 3.1 contains a schematic display of the selection procedure and inclusion and exclusion criteria for this study. The procedure consisted of multiple steps: (1) informed consent, (2) availability of computer logs of the intervention program, containing information about dosage, progress within the intervention, and the intervention start and end date, (3) intervention start, (4) intervention end, (5) learning disabilities or (6) physical disabilities. Regarding consent (Step 1), approval was obtained from Ethics Review Board of the University of Amsterdam (project number 2018-CDE-8677). Regarding the intervention start (Step 3), many children started the intervention before or after the prescribed moment, i.e. the middle of the second kindergarten year. We included children who started the intervention around the prescribed moment, i.e. between January and the end of June in the second kindergarten year. Regarding the intervention end (Step 4), children did not always continue the intervention until it was finished, i.e. around the middle of second grade. A small group of children stopped already before the middle of first grade, while others stopped during first grade or before the beginning of second grade. We do not know why. We included all children who continued the intervention until at least the middle of first grade (i.e. January). We denoted this group as the *Full sample*. In addition,

we included a subsample of children who continued the intervention until at least the end of first grade, referred to as the *Subsample*.

Figure 3.1

Selection of Children From a Larger Sample For This Study



The Full sample consisted of 369 kindergartners from 50 schools. Per cohort, between 99 and 136 children were included. The Subsample consisted of 291 kindergartners from 49 schools. At the start of the intervention, i.e. in the middle of the

second kindergarten year, children were on average 5.56 years old (*SD* = 0.37 years). In the Netherlands, elementary school starts at the age of 4. Children follow two years of kindergarten, before going to first grade, where formal reading instruction begins. We refer to the second kindergarten year as 'kindergarten' in the remaining part of the paper. Most children were born in the Netherlands (Full sample: 95%; Subsample: 96%). There were more boys (Full sample: 59%; Subsample: 55%) than girls.

Within the Full sample, pre-test scores (K_{mid}) were available for 79% of the children. Children scored between the 0.1st and 95th percentile (*M* = 44.78, *SD* = 19.67) on phonological awareness and knew between 0 and 24 (out of 34) letters at the start of the intervention. Among them, 59% had low scores on phonological awareness (≤25th percentile) and/or letter knowledge (knowing ≤6 letters), 14% had below-average scores on both phonological awareness (26-50th percentile) and letter knowledge (7-8 letters), and 27% had average or above average scores on both measures.

2.2 *Design*

Children in the Full Sample were followed from the intervention start, i.e. the middle of kindergarten, until the middle of first grade, including four measurement waves: the middle of kindergarten (K_{mid}; pre-test), the end of kindergarten (K_{end}), the start of first grade ($G1_{start}$), and the middle of first grade ($G1_{mid}$; post-test). The Subsample was followed until the middle of second grade, including two additional measurement waves: the end of first grade ($G1_{end;}$ post-test), and the middle of second grade ($G2_{mid}$; follow-up).

2.3 *Intervention*

The computer-based intervention *Build!* (in Dutch: *Bouw!*) is a prevention program for children at risk for reading problems. The intervention is intended to start in the middle of kindergarten, before formal reading instruction begins, and takes approximately two years to complete. It covers the precursors of reading, i.e. letter knowledge and phonological awareness, and continues with reading accuracy and fluency of one- and two-syllable words. The program includes the main orthographic complexities of Dutch (i.e. digraphs, consonant clusters, compound words, and open and closed syllables).

The program consists of 523 digital lessons, divided over twelve program parts. To finish a program part, children have to complete a test. Based on the performance on the test, the program suggests which lessons the child should continue with.

When performance is poor, certain lessons of the current program part should be reviewed before taking the test again and moving to the next program part. When performance is sufficient, the child is directed to the first lesson of the next program part. When performance is excellent, the child is directed to the test of the next program part and, based on the performance on that test, directed to lessons in that program part. See Regtvoort et al. (2013) for more information about the program.

Children are assisted by a tutor who reads aloud the instruction provided by the program, administers the tests, and provides support, if needed. The tutor could be a professional (e.g. a teacher) or non-professional (e.g. a parent, volunteer or older student). It was prescribed to provide children with 3-4 intervention sessions per week of 10-15 minutes each. Most schools aimed to provide 2-3 sessions per week at school, and asked parents to provide 1-2 session(s) at home.

2.4 *Measures*

2.4.1 *Dosage and Progress*

Data on dosage and progress were obtained from the computer logs of *Build!*, which contained information about the lessons a child had completed within the program: the lesson number, date, duration etc.

Dosage

Dosage was defined as the number of hours spent on the intervention. We took the sum of the lesson durations, saved in the computer logs. Very short and very long lesson durations were detected by visual inspection of distributions. What was seen as 'very short' and 'very long' depended on the lesson type (there were six lesson types, see Regtvoort et al. (2013) for more information). The minimum duration was either 0.50 or 1.00 minutes. The maximum lesson duration was between 13 and 16 minutes for short lesson types (Phono, Memory) and between 18 and 35 minutes for long lesson types (Domino, Self, Build, Letters). No more than 2% of the lessons were detected as very short or very long. These lesson durations were replaced using two-way imputation based on (van Ginkel et al., 2007). Using the following formula, new lesson durations were predicted based on the child's average speed and the average length of the specific lesson: $\gamma_{tlc} = (\mu_{tc} \,/\, \mu_t) \times \mu_l$, where γ_{tlc} is the duration of lesson l of lesson type t for child c, μ_{tc} is the child average duration for lesson type t; μ_t is the sample average for lesson type t, and μ_l is the overall average duration of lesson l. Next, dosage was calculated for three intervention periods: from the middle of kindergarten until the end of kindergarten (Kmid-Kend), from the start of first grade

until the middle of first grade (G1start-G1mid), and – only for the Subsample – from the middle of first grade until the end of first grade (G1mid-G1end, see Figure 3.2). Each period started right after the measurement wave and lasted until the next wave was finished.

Progress Within the Intervention

Progress within the intervention was defined as the number of new lessons completed, determined based on the lesson numbers from the computer logs. That is, we counted each lesson that was completed and subtracted the number of lessons that had been reviewed. Progress was determined for the same periods as dosage.

2.4.2 *Family Characteristics*

Familial Risk for Dyslexia

Both parents were asked 'Do you have dyslexia?'. Answer options were: 'no', 'I think so, but it has not been diagnosed', 'yes, I have a diagnosis of dyslexia'. Based on the answers, a dummy variable for familial risk was created. If one of the parents or both parents chose 'I think so, but it has not been diagnosed' or 'yes, I have a diagnosis of dyslexia', the child was classified as having a familial risk for dyslexia. If both parents answered 'no' or if only one parent responded and answered 'no', the child was classified as not having a familial risk for dyslexia.

Parental Education

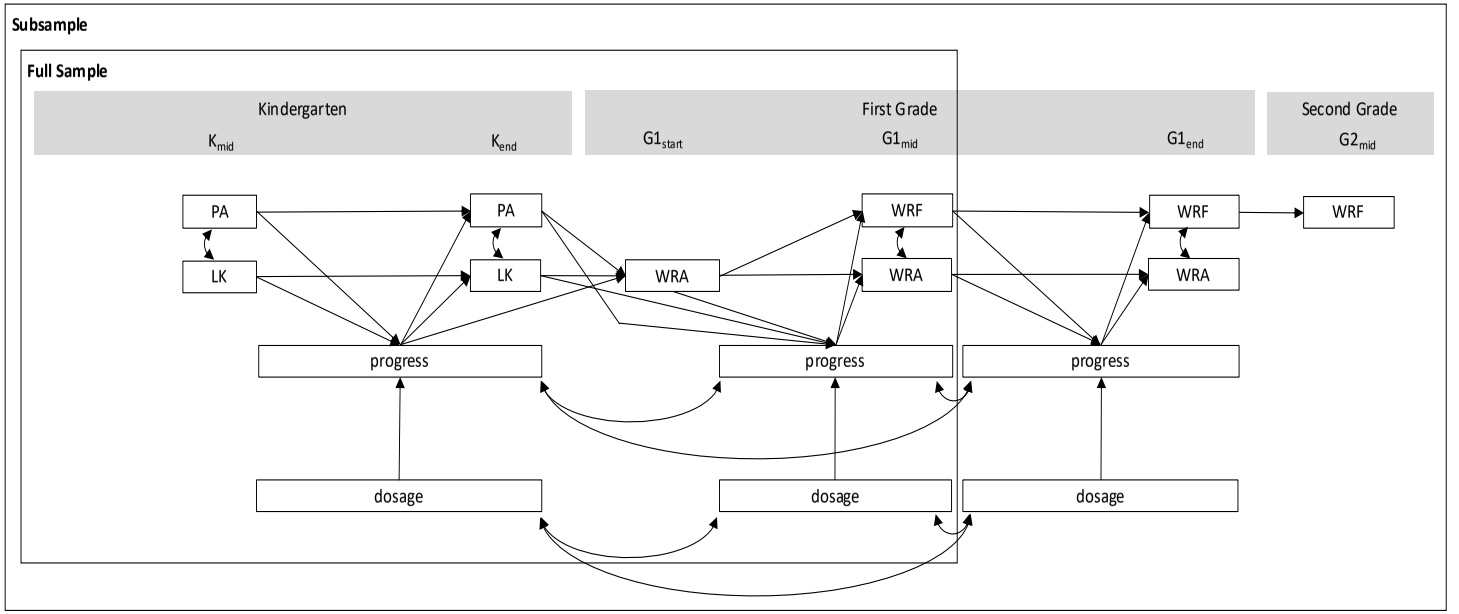
Both parents were asked 'What is the highest degree or level of schooling you have completed?'. Answer options were 1 = primary school, 2 = secondary school, 3 = vocational education, 4 = university of applied sciences, or 5 = university. If both parents answered the question, the average was taken. If only one parent had answered, that single answer was taken as an indicator of parental education.

2.4.3 *Literacy Outcomes*

Intervention outcomes were letter knowledge and phonological awareness in kindergarten, and word reading accuracy and word reading fluency in first and second grade. An overview of instruments per measurement wave is shown in Table 3.1. Tests were administered by a teacher, school counselor or trained research assistant at school.

Figure 3.2

Hypothesized Model: Effects of Dosage and Progress on Literacy Outcomes



Note. The small rectangle encloses the measurement waves included in the Full Sample model. The large rectangle encloses the measurement waves included in the Subsample model. PA = phonological awareness, LK = letter knowledge, WRA = word reading accuracy, WRF = word reading fluency.

Table 3.1

Overview of Instruments per Measurement Wave

Instrument	Measurement Wave					
	Kindergarten		First Grade			Second Grade
	K_{mid}[a]	K_{end}	$G1_{start}$	$G1_{mid}$	$G1_{end}$	$G2_{mid}$
Letter Knowledge	x	x				
Phonological Awareness						
Full Test	x					
Subtests B, C, and H		x				
Word Reading Accuracy						
Test Beginning Readers			x			
Test Advanced Readers				x	x	
Word Reading Fluency				x	x	x

[a]Pre-tests for selection.

Letter Knowledge

The Grapheme Test (*Grafementoets*; Verhoeven, 1993) was used to measure children's active letter knowledge. The test consisted of 34 Dutch graphemes, printed on one page in two columns. Children were asked to pronounce all graphemes, without time limit. The test score consisted of the number of correctly named graphemes. Cronbach's alpha is .85 (Verhoeven, 2000).

Phonological Awareness

A subtest of the CELF-4-NL (Kort et al., 2008), the Dutch version of the *Clinical Evaluation of Language Fundamentals fourth edition* (Semel et al., 2003), was used to measure phonological awareness. The subtest included nine parts: (A) phoneme blending, (B) last phoneme identification, (C) middle phoneme identification, (D) word identification, (E) last syllable deletion with two-syllable words, (F) syllable identification, (G) first or last syllable deletion with two-syllable words, (H) phoneme substitution, and (I) first or last syllable deletion with three-syllable words. All parts contained two examples, followed by five test items, and were stopped after three consecutive incorrect answers. The maximum score is 45. The score consisted of the number of correct items. Cronbach's alpha reliability is .85 (Kort et al., 2008).

At the end of kindergarten, only three subtests were administered: (B) last phoneme identification, (C) middle phoneme identification, and (H) phoneme substitution. The maximum score is 15. Reliability analysis within the larger sample of 1150

children showed that reliability of the three subtests together was acceptable (Cronbach's α = .70). For this measurement wave, the score on subtest H was missing for 32% of the children, as the test was not administered in the first cohort of kindergartners (2017-2018). We assumed that these missing values were Missing at Random (MAR), because missingness was not related to the skill itself, but to the phase of the study. Missing scores on this subtest were predicted from the scores on the two other subtests by the EM algorithm in IBM SPSS Statistics 25 (IBM, 2017). This imputation was done with the larger dataset of 1150 children who received the intervention plus 1370 from District 1 who were not selected for the intervention.

Word Reading Accuracy

Word reading accuracy was measured with two tests: a test for beginning readers (start of first grade) and a test for more advanced readers (the middle and the end of first grade). The test for beginning readers was developed by Regtvoort et al. (2013). It included ten consistent words: three CV words (e.g. aap [monkey]) and seven CVC words (e.g. bus [bus]). The words were printed in one column. Children were asked to read aloud all words without time limit. For each word, children could receive two points if the word was read correctly at once, one point if the letters were sounded out first and then blended into a word, or no points if the letters were only sounded out, or if the word was read incorrectly. There were two example items. Children who only sounded out the letters were asked to also blend them into a word. The maximum score was 20. Reliability of this test is unknown, because teachers reported only the full score.

The test for advanced readers was developed by de Jong and Wolters (2002; see also van Viersen et al., 2018). The test contained 40 words of increasing length and difficulty, including five one-syllable words (e.g. laan [lane]), twelve two-syllable words (e.g. eten [to eat]), thirteen three-syllable words (e.g. postzegel [stamp]), and ten four-syllable words (e.g. vuilnisemmer [trash bin]). The words were printed on two pages, with two columns of ten words each. Children were asked to read aloud all words, without time limit. The test was stopped after four consecutive incorrect responses. Scores represent the number of words read correctly (maximum is 40). The parallel test reliability is .83 (de Jong & Wolters, 2002).

Word Reading Fluency

Word reading fluency was measured with a standardized test: the *Three Minute Test* (Krom et al., 2010), used in most schools in the Netherlands to monitor word reading achievement during primary school. This test contained three cards of 150 words,

increasing in difficulty per card, but not within cards. The first card included one-syllable consistent words without letter clusters, e.g. tas [bag]. The second card contained one-syllable consistent and inconsistent words with letter clusters, e.g. plant [plant] or sneeuw [snow]. The third card included consistent and inconsistent words with multiple syllables, e.g. voetbal [football] or temperatuur [temperature]. Children were asked to read aloud the words as quickly as possible, while making as few errors as possible. For each card, the children were stopped after one minute. In the middle of first grade, only the first two cards were administered. At the end of first grade and in second grade, all three cards were administered. The number of words read correctly on all administered cards together was counted and converted to an ability score, whereby the scores on Card 1-2 and Card 1-3 were put on the same scale, ranging from 0 to 154 (Krom et al., 2010). In District 2, two different versions of this test were in use (version 1995 and version 2010). Because the normed tables of the oldest version were outdated, the standardized ability scores on the oldest version (i.e. version 1995) were transformed to the scale of the newest version, based on a dataset of 265 children who read both versions on the same day. Cronbach's alpha for the three cards is .88, .94, and .92 respectively (Verhoeven & van Leeuwe, 2003).

2.5 *Analytic Strategy*

The analyses were conducted in two parts. First we examined the relations of dosage, progress, and (pre)literacy development. The next part concerned the relations of these variables with familial risk for dyslexia and parental education.

2.5.1 *Dosage, Progress, and the Development of Literacy Skills*

Hypothesized Model

The first aim of this study was to investigate to what extent literacy outcomes were related to intervention dosage, via progress within the intervention. In the hypothesized model (see Figure 3.2), three intervention periods were distinguished: from the middle of kindergarten until the end of kindergarten (K_{mid}-K_{end}), from the start of first grade until the middle of first grade ($G1_{start}$-$G1_{mid}$), and – only for the Subsample – from the middle of first grade until the end of first grade ($G1_{mid}$-$G1_{end}$). *D*osage in each period was specified to predict progress in the same period which, in turn, predicted literacy outcomes at the end of the period. Progress within each period was predicted by literacy skills at the start of the period, because children with lower (pre)literacy skills might need more review lessons or more time to finish a lesson. Covariances between the exogenous variables were freely estimated.

Model Estimation and Fit

Data were analyzed in Mplus version 7.31 (Muthén & Muthén, 1998-2017). Children (level 1) were treated as nested within schools (level 2). To deal with missing values and the nested structure of the data, we used a full-information maximum likelihood estimator with standard errors and a chi-square test statistic robust to non-independence of observations (MLR; Yuan & Bentler, 2000). Model fit was evaluated based on the following model fit indices: chi-square test statistic, CFI, and the RMSEA 90% confidence interval. For chi-square, we used an alpha level of .05. Here, a non-significant chi-square test statistic indicated that the model fits the data (Schermelleh-Engel et al., 2003). A CFI value larger than .95 and a RMSEA value below .08 indicated acceptable model fit, whereas a CFI value larger than .97 and a RMSEA value below .05 indicated good model fit (Schermelleh-Engel et al., 2003).

Model Building and Trimming Approach

We used two ways to adjust the model, i.e. model building and model trimming (Kline, 2015). First, we added paths to the hypothesized model based on modification indices, until model fit was sufficient. Then, we simplified the model by eliminating non-significant paths. Using the Wald *W* statistic, we tested whether these paths could be fixed to zero.

For the Subsample, we took the adjusted Full sample model as a basis and added the later measurement waves, i.e. $G1_{end}$ and $G2_{mid}$ (see Figure 3.2). Then, we started the model building and trimming approach for the Subsample model.

Evaluation of the Effects of Dosage

For the final Full and Subsample model, significance of all paths was evaluated by t-tests. The indirect effects, i.e. the effects of dosage on literacy through progress within the intervention, were evaluated based on the 95% confidence interval for parameter estimates using the Bayes estimator. Standardized regression coefficients of ≤ .29 were considered small, .30 to .49 moderate, and ≥ .50 as large.

2.5.2 *Effects of Family Characteristics*

The second aim of this study was to investigate whether two family characteristics, i.e. familial risk for dyslexia and parental education, were related to dosage, progress, and literacy outcomes. For the Full sample, we added familial risk for dyslexia and parental education to the final model, and all measures of literacy, progress, and dosage were regressed on these two family characteristics. Then, we removed the non-significant paths by model trimming. For the subsample, we added only the paths of

familial risk for dyslexia and parental education that remained in the Full sample model to the final model of the Subsample. Based on modification indices, we checked whether any more paths were necessary.

Data, study materials, and analysis code are available on request from the first author.

3 Results

3.1 *Data Processing*

Prior to analyses, we checked for outliers on all variables (values more than 3 standard deviations from the mean). Per variable, no more than 2.1% of the data was depicted as outliers. All outliers concerned extremely high scores. Outliers were coded as missing values. After outliers were omitted, all variables had skewness and kurtosis between -2 and 2.

We had to deal with missing values. That is, we missed the score on letter knowledge and phonological awareness at Kmid for 25% of the children and at Kend for 16% of the children, because most schools in District 2 joined the research project later, when Cohort 1 and 2 were already in first and second grade respectively. In addition, the score on word reading fluency at G2mid was missing for 42% of the children, mainly because this test could not be administered in Cohort 3 as the research project ended when these children were in first grade. We assumed that these missing values were Missing at Random (MAR), because missingness was not related to the skill itself, but to the district or cohort. For the other literacy measures and for the family characteristics, no more than 10% of scores were missing. We assumed these missings were completely at random (MCAR), because they missed, for example, because children were sick or moved to another school.

3.2 *Descriptive Statistics*

Descriptive statistics of literacy outcomes, dosage, progress, familial risk, and parental education for the Full sample and for the Subsample can be found in Table 3.2. Literacy scores of the Full sample were slightly better than those of the Subsample. To investigate whether there was a difference between children who continued and stopped the intervention, we isolated the group who stopped the intervention at $G1_{mid}$ from the Full sample and compared them with the Subsample that continued. The children who stopped the intervention showed higher performance on word reading accuracy at $G1_{mid}$, t (334) = 6.42, $p < .001$, and word reading fluency at $G1_{mid}$, t (330) =

7.40, $p < .001$, and made less progress within the intervention in the period $G1_{start}$-$G1_{mid}$, t (365) = 3.02, $p = .003$ than the children who continued the intervention. On average, 60% of the children who stopped the intervention reached average levels of reading fluency at $G1_{mid}$, and another 17% reached near-average levels, whereas in the group that continued the intervention this was only 20% and 19% respectively. The higher literacy scores can thus explain why children stopped the intervention after $G1_{mid}$. The Subsample that continued did not differ from the children who stopped with respect to familial risk, χ^2 (1) = 1.687, $p = .194$, and parental education, t (335) = 0.96, $p = .338$.

Correlations between literacy outcomes, dosage, progress and parental education are shown in Table 3.3. In both the Full and Subsample, dosage was strongly related to progress in all three intervention periods. In the Full sample, progress in the period K_{mid}-K_{end} was weakly related to phonological awareness at K_{end}, moderately to letter knowledge at K_{end}, and moderately to word reading accuracy at $G1_{start}$. Progress in period $G1_{start}$-$G1_{mid}$ was weakly related to word reading accuracy at $G1_{mid}$ and moderately to word reading fluency at $G1_{mid}$. In the Subsample, correlations were slightly weaker, and progress in the period $G1_{mid}$-$G1_{end}$ was not related to word reading accuracy or fluency at $G1_{end}$. In both the Full and Subsample, parental education was hardly related to any other variable. For familial risk, it was not possible to calculate correlations, as this variable was dichotomous. Literacy scores, dosage, and progress of children with and without familial risk are shown in Supplemental Materials Chapter 3, Table S3.1.

Table 3.2

Descriptives of Literacy Measures, Dosage, Progress, Familial Risk, and Parental Educational Level for the Full and the Subsample

Measure	Wave/Period	Maximum score	Full sample: intervention until at least the middle of first grade (N = 369)		Subsample: intervention until the end of first grade (n = 291)	
			M	*SD*	*M*	*SD*
PA	K_{mid}[a]	45	22.50	8.38	21.68	8.13
	K_{end}[b]	15	8.41	3.94	8.12	4.00
LK	K_{mid}	34	6.71	3.92	6.60	3.77
	K_{end}	34	14.07	5.31	13.77	5.19
WRA	$G1_{start}$	20	9.61	4.50	9.25	4.22
	$G1_{mid}$	40	12.13	8.57	10.34	7.00
	$G1_{end}$	40	23.85	10.06	22.27	8.13
WRF	$G1_{mid}$	137	11.64	6.80	10.14	5.25
	$G1_{end}$	154	22.54	13.09	20.35	12.00
	$G2_{mid}$	154	37.57	17.88	35.60	16.91
Dosage	K_{mid}-K_{end}	-	25.54	11.93	25.50	11.82
	$G1_{start}$-$G1_{mid}$	-	27.32	9.82	26.89	9.47
	$G1_{mid}$-$G1_{end}$	-	-		33.57	14.52
Progress	K_{mid}-K_{end}	-	40.15	22.63	38.81	20.84
	$G1_{start}$-$G1_{mid}$	-	60.38	29.96	57.97	29.48
	$G1_{mid}$-$G1_{end}$	-	-		78.01	38.63
	Category		%		%	
Familial risk	Yes		23.51		25.10	
Parental education	Low		19.88		20.15	
	Average		51.33		52.09	
	High		28.78		27.76	

Note. LK = letter knowledge; PA = phonological awareness; WRA = word reading accuracy; WRF = word reading fluency; K = kindergarten; G1 = first grade.
[a]full test; [b]part of the test

Table 3.3
Correlations Between Literacy Measures, Dosage and Progress

Measure		Wave/ Period	Pearson's Correlations																
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	LK	K_{mid}	-	.50*	.22*	.35*	.18*	.12	.10	.01	.07	-.03	.06	.02	.07	.14*	-.01	.03	.04
2		K_{end}	.52*	-	.02	.39*	.43*	.31*	.25*	.27*	.24*	.17*	.26*	.04	-.04	.43*	.14*	-.06	.10
3	PA	K_{mid}[a]	.20*	-.01	-	.49*	.09	.29*	.31*	.22*	.18*	.12	-.04	.02	-.08	.12	.10	-.03	.11
4		K_{end}[b]	.34*	.34*	.46*	-	.25*	.35*	.30*	.21*	.17*	.05	.01	-.04	-.13*	.19*	.08	-.08	.09
5	WRA	$G1_{start}$	.17*	.40*	.07	.21*	-	.39*	.43*	.40*	.34*	.21*	.19*	.15*	-.11	.34*	.23*	-.09	.04
6		$G1_{mid}$	.10	.28*	.24*	.31*	.27*	-	.61*	.68*	.55*	.46*	.09	-.02	-.28*	.25*	.16*	-.15*	.11*
7		$G1_{end}$	.09	.21*	.29*	.29*	.38*	.56*	-	-	-	-	-	-	-	-	-	-	-
8	WRF	$G1_{mid}$	.08	.28*	.19*	.18*	.32*	.65*	.55*	-	.76*	.59*	.09	.11*	-.27*	.28*	.31*	-.14*	.06
9		$G1_{end}$	.04	.24*	.16*	.17*	.31*	.52*	.62*	.77*	-	-	-	-	-	-	-	-	-
10		$G2_{mid}$	-.04	.18*	.10	.06	.17*	.38*	.59*	.62*	.80*	-	-	-	-	-	-	-	-
11	Dosage	K_{mid}-K_{end}	-.02	.24*	-.11	-.06	.15*	.03	-.04	.08	.01	.00	-	.46*	.29*	.76*	.41*	.17*	.03
12		$G1_{start}$-$G1_{mid}$	-.05	-.03	.01	-.09	.06	-.09	-.07	.05	-.03	-.03	.45*	-	.43*	.35*	.85*	.32*	-.03
13		$G1_{mid}$-$G1_{end}$	-.01	-.06	-.04	-.12	-.03	-.19*	-.08	-.14*	-.15*	-.05	.33*	.53*	-	-	-	-	-
14	Progress	K_{mid}-K_{end}	.12	.43*	.07	.14*	.28*	.18*	.14*	.31*	.18*	.09	.76*	.35*	.25*	-	.46*	.16*	.04
15		$G1_{start}$-$G1_{mid}$	-.02	.10	.10	.05	.16*	.10	.07	.27*	.15*	.06	.38*	.85*	.39*	.44*	-	.34*	-.01
16		$G1_{mid}$-$G1_{end}$	-.02	-.08	.02	-.06	.01	-.03	.06	.03	.06	.04	.15*	.41*	.82*	.20*	.45*	-	-
17	Parental education		.10	.13	.11	.10	-.01	.04	.03	-.04	-.06	.05	.02	-.02	-.02	.01	-.02	-.07	-

Note. Upper triangle: Full sample (N = 369), lower triangle: Subsample (n = 291); LK = letter knowledge; PA = phonological awareness; WRA = word reading accuracy; WRF = word reading fluency; K = kindergarten; G1 = first grade.
[a]full test. [b]part of the test. *$p < .05$.

3.3 *Effects of Dosage*

The first aim of this study was to investigate the relations among dosage, progress, and literacy outcomes. We tested two hypotheses. Our first hypothesis was that during each of the three periods dosage and progress would have an additional effect on subsequent literacy outcomes when earlier literacy abilities were taken into account. Our second hypothesis was that the additional effect of dosage on later literacy abilities was fully mediated by progress within the intervention. These hypotheses were investigated in the Full and Subsample separately. We first present the findings for the Full Sample and then for the Subsample.

3.3.1 *Full Sample*

The hypothesized model in the Full sample (see Figure 3.2) did not fully fit the data, χ^2 (32) = 99.12, p <.001, RMSEA = .075 (90% CI [.059-.092]), CFI = .948. Based on the highest modification indices, we added an effect of letter knowledge at K_{mid} on phonological awareness at K_{end} and an effect of phonological awareness at K_{end} on word reading accuracy at $G1_{mid}$. Both adjustments seem theoretically understandable (Muter et al., 2004). Model fit was satisfactory after these adjustments, χ^2 (30) = 65.89, p <.001, RMSEA = .057 (90% CI [.038-.076]), CFI = .975. For reasons of clarity, non-significant paths of preliteracy skills on progress within the intervention were removed from the model, W (3) = 3.77, p = .29. The resulting model is shown in Figure 3.3 and fitted the data sufficiently, χ^2 (33) = 69.97, p <.001, RMSEA = .055 (90% CI [.037-.073]), CFI = .971.

In accordance with our hypothesis, the final model (see Figure 3.3) shows that progress during the first period had significant additional effects on the growth in phonological awareness and letter knowledge as well as on word reading accuracy, when prior preliteracy skills were controlled. Progress during the second period (the first half year of first grade) had a significant effect on reading fluency, but not on reading accuracy, when word reading accuracy at the start of this period was taken into account.

The model also shows that, as predicted, in each of these periods the relation between dosage and literacy outcomes was fully mediated by progress through the intervention. Children who received a higher dose made more progress within the intervention, and children who made more progress performed better on most subsequent literacy skills. The indirect effects of dosage on literacy outcomes are shown in Table 3.4. The significance of these indirect effects was evaluated with the Bayes estimator. In the first period (K_{mid}-K_{end}), children who received a higher dose had better

phonological awareness and letter knowledge at K_{end} (small and medium effect respectively) and read words more accurately at $G1_{start}$ (small effect), after controlling for preceding literacy skills. In the second period ($G1_{start}$-$G1_{mid}$), children who received a higher dose read words more fluently at $G1_{mid}$ (small effect), but not more accurately, after controlling for prior literacy skills.

Finally, it is noteworthy that children with better letter knowledge and phonological awareness halfway through the second year of kindergarten made significantly more progress within the intervention in the period K_{mid}-K_{end}. In contrast, these skills were not significantly associated with the amount of dosage. Progress during the next period, $G1_{start}$-$G1_{mid}$, was not related to children's earlier literacy skills.

Table 3.4

Indirect Effects of Dosage on Literacy Outcomes Through Progress Within the Intervention

Predictor	Outcome	β	
		Full sample (N = 396)	Subsample (n = 291)
Dosage K_{mid}-K_{end}	PA K_{end}	.10*	.06
	LK K_{end}	.31*	.30*
	WRA $G1_{start}$	.15*	.12*
Dosage $G1_{start}$-$G1_{mid}$	WRA $G1_{mid}$	.07	.03
	WRF $G1_{mid}$	.19*	.17*
Dosage $G1_{mid}$-$G1_{end}$	WRA $G1_{end}$	-	.06
	WRF $G1_{end}$	-	.03

Note. LK = letter knowledge; PA = phonological awareness; WRA = word reading accuracy; WRF = word reading fluency; K = kindergarten; G1 = first grade.

*$p < .05$

3.3.2 *Subsample*

For the first two periods, we fitted the same model in the Subsample as obtained in the Full sample, and we added the later measurement waves, i.e. $G1_{end}$ and $G2_{mid}$, to the Subsample model. The model did not fit the data sufficiently, χ^2 (85) = 204.01, p <.001, RMSEA = .069 (90% CI [.057-.082]), CFI = .939. Based on the highest modification indices, we added two paths, one from word reading fluency at $G1_{mid}$ to word reading accuracy at $G1_{end}$, and one path from word reading accuracy at $G1_{start}$ to word reading accuracy at $G1_{end}$. With these two adjustments, the fit of the model was

satisfactory, χ^2 (83) = 174.36, p <.001, RMSEA = .062 (90% CI [.049-.074]), CFI = .953. Non-significant effects of literacy skills on progress within the intervention were removed from the model, W (2) = 5.23, p = .073. The final model is shown in Figure 3.4 and fitted the data sufficiently, χ^2 (85) = 178.24, p <.001, RMSEA = .061 (90% CI [.049-.074]), CFI = .952.

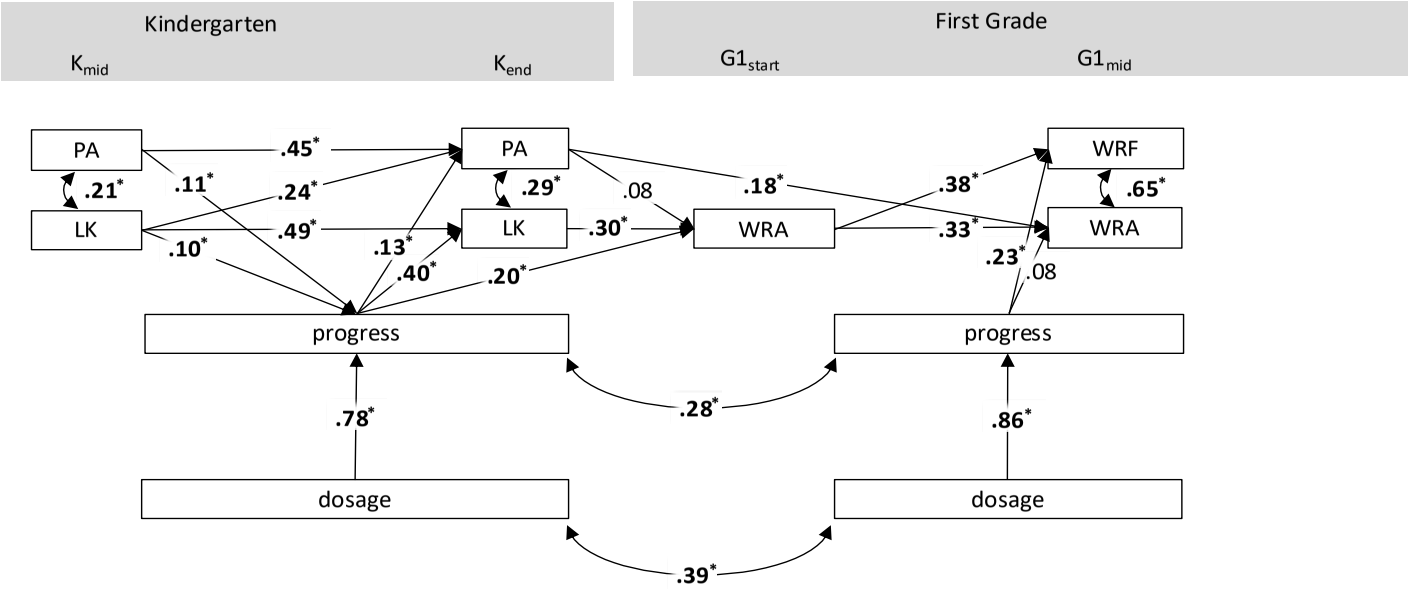
The final model (see Figure 3.4) shows that the effects of progress on subsequent literacy skills in the first two periods were highly similar to the effects observed in the Full Sample. However, in contrast to the first periods, progress no longer had an additional effect on reading development during the third period ($G1_{mid}$-$G1_{end}$). As found in the model of the Full sample, the effects of dosage on literacy outcomes were significant and fully mediated by progress within the intervention. An exception was the effect of dosage in the first period (K_{mid}-K_{end}) on phonological awareness at the end of the period (K_{end}), which was not significant in the Subsample (see Table 3.4). Note that the effects of dosage were slightly stronger in the Full sample than in the Subsample. Like progress, dosage during the additional intervention period ($G1_{mid}$-$G1_{end}$), did not have an additional effect on children's word reading accuracy and fluency at the end of the period ($G1_{end}$). Finally, also the effects of literacy skills on progress within the intervention were comparable to the effects observed in the Full sample. However, in the additional intervention period ($G1_{mid}$-$G1_{end}$), children's progress within the intervention was not affected by their prior literacy skills at $G1_{mid}$.

3.4 *Effects of Family Characteristics*

The second aim of this study was to investigate the direct and indirect relations (through dosage and progress) of familial risk for dyslexia and parental education with the development of preliteracy and reading skills. To this end we added familial risk for dyslexia and parental education to the model of the Full Sample (see Figure 3.3) by regressing all measures of literacy, progress within the intervention, and dosage on these two family characteristics. Model fit was sufficient, χ^2 (33) = 65.45, p <.001, RMSEA = .052 (90% CI [.033-.070]), CFI = .977. Again, non-significant paths were removed from the model, W (16) = 17.61, p = .34. The trimmed model fitted the data well, χ^2 (49) = 79.05, p <.001, RMSEA = .041 (90% CI [.023-.057]), CFI = .978. The final model is shown in Figure 3.5. Note that the effect of parental education on word reading accuracy at $G1_{mid}$ was initially significant, but became non-significant when other paths were removed.

Figure 3.3

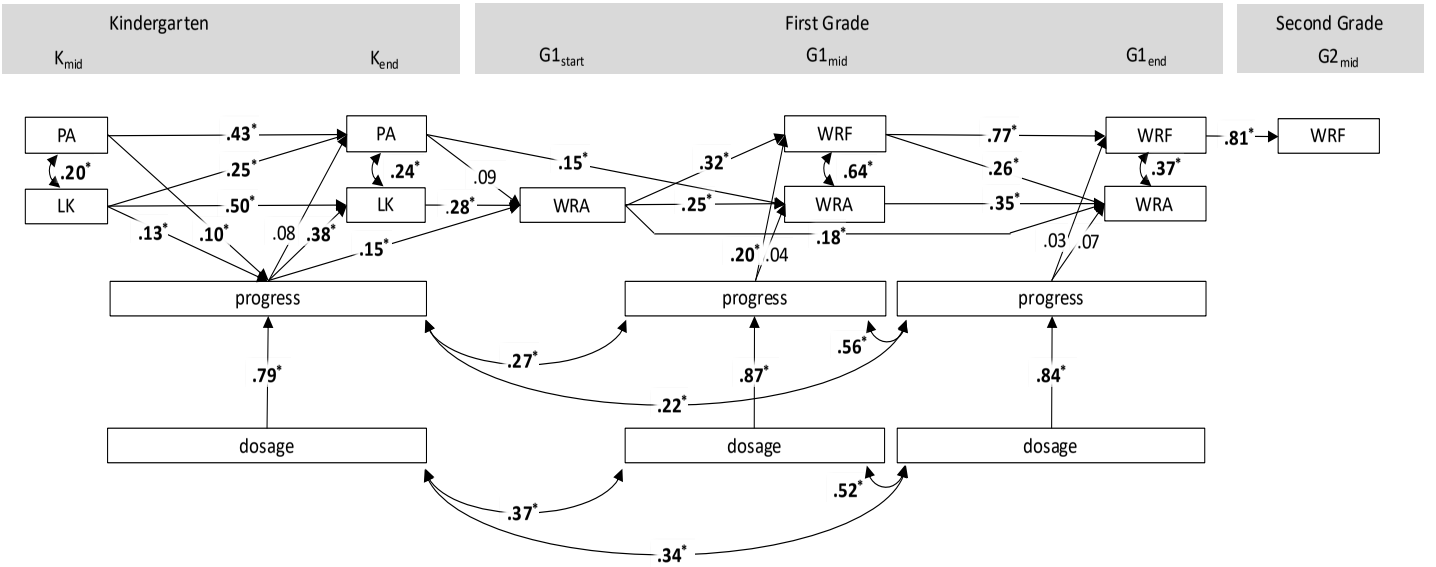
Full Sample: Effects of Dosage and Progress on Literacy Outcomes



Note. Standardized regression coefficients are shown. PA = phonological awareness; LK = letter knowledge; WRA = word reading accuracy; WRF = word reading fluency.

Figure 3.4

Subsample: Effects of Dosage and Progress on Literacy Outcomes

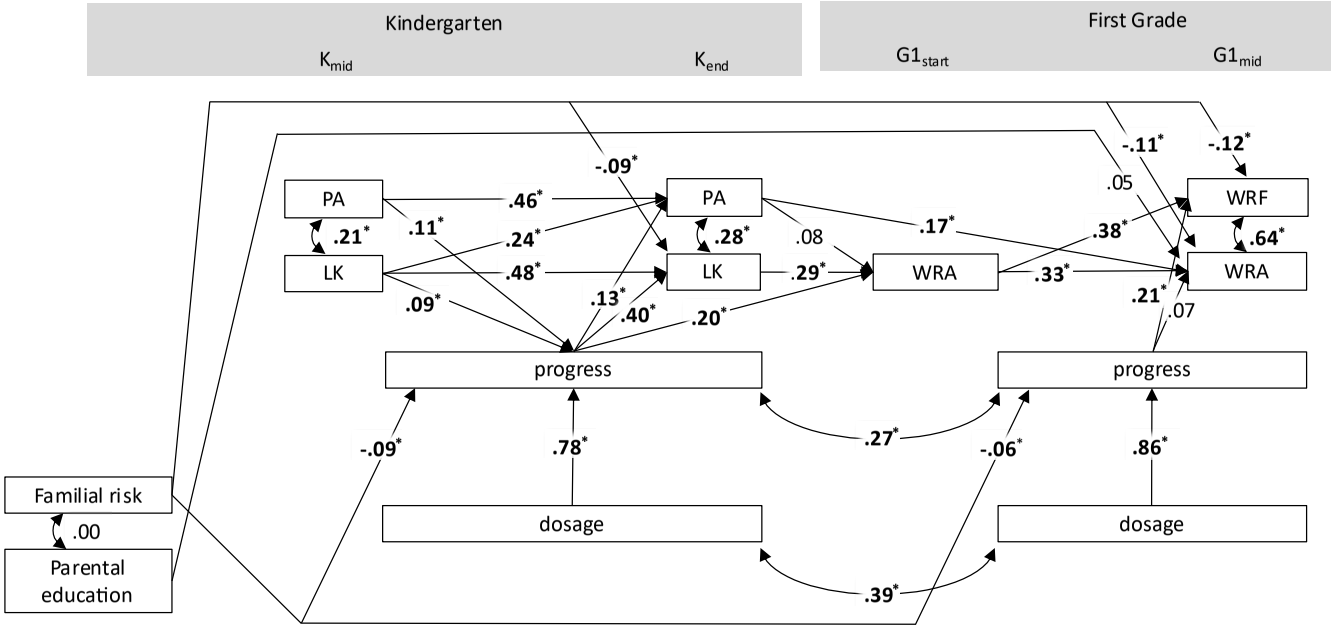


Note. Standardized regression coefficients are shown. PA = phonological awareness; LK = letter knowledge; WRA = word reading accuracy; WRF = word reading fluency.

*$p < .05$.

Figure 3.5

Full Sample: Effects of Familial Risk and Parental Education on Intervention Outcomes



Note. Standardized regression coefficients are shown. PA = phonological awareness; LK = letter knowledge; WRA = word reading accuracy; WRF = word reading fluency.

$^{*}p < .05$.

Several results are noteworthy. First, as expected, during the first period (kindergarten) the acquisition of letter knowledge of children with familial risk was slower than in children without familial risk, even after controlling for their earlier letter knowledge (see Figure 3.5). Moreover, also after taking prior literacy skills into account, familial risk had an additional negative effect on word reading accuracy and fluency by $G1_{mid}$. That is, children with familial risk read words less accurately and less fluently. Second, familial risk also affected literacy outcomes through its small but significant relation with progress through the intervention. In both periods, the progress of children with familial risk was smaller than in children without familial risk. Third, familial risk was not significantly related to dosage, indicating that children with and without familial risk received a similar amount of practice. Finally, there were no significant effects of parental education on children's literacy skills, progress within the intervention, and dosage.

The same analysis was done in the Subsample by adding the paths from familial risk and parental education in Figure 3.5 to the final Subsample model presented in Figure 3.4. This model had sufficient model fit, χ^2 (111) = 195.35, p <.001; RMSEA = .051 (90% CI [.039-.063]), CFI = .958. Findings were highly comparable to the results in the Full sample, except that the effect of familial risk on progress in the second period ($G1_{start}$-$G1_{mid}$) was no longer significant, β = -.03, t (1) = -1.226, p = .220. Next, we also added paths from familial risk and parental education to reading skills, progress, and dosage in the last period ($G1_{end}$ and $G2_{mid}$). None of these paths were significant.

4 Discussion

In this study, children who were considered to be at risk for reading difficulties received the intervention *Build!* to foster their preliteracy and reading development. The first aim of this study was to examine whether the amount of time spent on the intervention (dosage) was related to the growth of literacy skills, that is preliteracy skills and word reading. The second aim was to investigate to what extent dosage and literacy development were related to two major family characteristics, familial risk for dyslexia and parental education.

4.1 *Dosage and the Development of (Pre)literacy Skills*

We examined the relation between dosage and literacy development during three periods. The results showed that dosage had an additional effect on subsequent (pre)literacy skills from the middle to the end of the second kindergarten year and the beginning to the middle of first grade. In particular, findings show that children

who spent more time on the intervention in the second half of kindergarten, knew more letters and had more phonological awareness at the end of kindergarten, and also read words more accurately at the start of first grade. Children who spent more time on the intervention during the first half of first grade, read words more fluently, although not more accurately, in the middle of first grade.

The additional effects of dosage in the third period, mid to end of first grade, could only be examined in a subsample of children, because a group of children stopped the intervention after the middle of first grade. In this group, 77% of the children had reached above average or near-average levels of word reading fluency, suggesting that the schools deemed further intervention unnecessary. The children of the subsample who continued the intervention in the third period were the relatively poorer readers. In this subsample, we found largely the same effects as in the Full Sample for the first two periods. In the third period, we did not find additional effects of dosage on subsequent literacy outcomes. At first glance, this seems to contradict findings of RCTs on *Build!*, which showed that the difference between the intervention group and control group further increased during this period (Regtvoort et al., 2013; Zijlstra et al., 2021). However, these studies do not show which intervention periods caused this acceleration.

The absence of an additional effect of dosage on reading development in the third period does not necessarily indicate that the intervention period after the second half of first grade can be skipped. It remains possible that continuation of *Build!* might have helped children to maintain their increased level of reading ability as reached in the first half of first grade. Another possibility is that as the classroom instruction moves on, the content of the program and the regular classroom instruction have started to overlap. As a result, the program might have changed during the third period from a preventive to a more remedial intervention. Generally, remediation is less effective than prevention and therefore additional effects of dosage might be more difficult to observe (Ferrer et al., 2015; Lovett, 2017; Wanzek et al., 2013). A third possibility is that the program was implemented with less fidelity in the second half of first grade, which can dampen children's outcomes (Stein et al., 2008; Wolgemuth et al., 2014; Zijlstra et al., 2014). In support of this explanation we found that the percentage of available school weeks spent on the intervention was higher in the first half of first grade than in the second half (92% versus 80%). Taken together, however, we have to admit that it is yet unknown how long extensive literacy interventions like *Build!* should be continued to maintain their effects.

An important and novel feature of the current study was the distinction between dosage and progress. Dosage was believed to reflect the amount of practice, whereas

progress was conceived as a pure measure of exposure to the content of the intervention. As expected, dosage had a strong relation with progress, because time spent on the intervention obviously leads to the completion of more novel intervention lessons. However, it is important to note that dosage and progress were not identical, sharing approximately 60-75 percent of their variance. A remarkable difference between dosage and progress was that, although they both predicted literacy skills (directly or indirectly), prior literacy skills predicted progress but were not associated with dosage. While in previous studies it was assumed there is a one-way effect of progress on literacy skills (Regtvoort et al., 2013), our findings suggest there might be an upward spiral of effects: children with better initial (pre)literacy skills complete more new intervention lessons; because of completing these lessons, their (pre)literacy skills improve more and so on.

The relations between dosage and intervention outcomes, through progress, in this study suggest that dosage, as a dimension of treatment integrity, was related to the extent to which children benefitted from the early literacy intervention. In contrast, several meta-analyses showed no relation between dosage and intervention effectiveness (Tran et al., 2011; Wanzek et al., 2013; Wanzek & Vaughn, 2007). However, as said in the introduction, these meta-analyses examined relations of dosage with literacy outcomes across studies. Studies comparing children who received larger and smaller doses of the same intervention, rather than different interventions, like this study, do show that dosage affects intervention outcomes (Al Otaiba et al., 2005; Wolgemuth et al., 2014; Zijlstra et al., 2014). Although this was found for both digital and nondigital literacy interventions, the fact that the current intervention was computer-based might have provided us with a better chance to find relations between dosage and intervention outcomes. In digital interventions, several dimensions of treatment integrity such as adherence and quality of instruction, are covered by the program. As a result, dosage might play a more prominent role. Another reason for the effects of dosage in this study might be that we focused on a school-based implementation of an early literacy intervention, whereas previous studies were researcher-led. When the intervention is implemented by many schools, the variety in dosage becomes larger, which could result in a stronger relation with intervention outcomes.

As we had no control group, this study does not provide straightforward and conclusive evidence that the intervention was effective in improving children's literacy skills. However, the predictors preceded the outcomes, and autoregressive effects were taken into account. The additional effects on dosage therefore support the effect of the intervention (Gollob & Reichardt, 1987). Moreover, we found that the

effects of dosage on literacy outcomes were fully mediated by progress within the intervention. As such, it is unlikely that literacy growth can be explained by other aspects of the intervention, such as the individual attention the child received during the one-to-one intervention sessions. With respect to the first research question, our findings are a first step towards scientific support for the effectiveness of the current early literacy intervention in natural school settings.

4.2 *Effects of Family Characteristics*

The second aim of this study was to investigate to what extent literacy outcomes were associated with two family characteristics, familial risk for dyslexia and parental education. Specifically, we examined whether these family characteristics affected children's reading outcomes directly and/or indirectly through dosage and progress within the intervention.

4.2.1 *Familial Risk for Dyslexia*

Children with familial risk for dyslexia had on average poorer literacy skills than children without such risk. They knew fewer letters at the end of kindergarten, and reached lower reading skills in first grade. On top of that, familial risk had an additional negative effect on word reading accuracy and fluency in the middle of first grade, indicating that the effect of familial risk on reading was not fully mediated by children's preliteracy skills in kindergarten. Interestingly, this was exactly what was found in a longitudinal study by van Viersen et al. (van Viersen et al., 2018). They also found that familial risk had additional effects on children's reading accuracy and fluency in second grade, after controlling for its effect on children's preliteracy skills in kindergarten. Based on genetic research (Byrne et al., 2009), van Viersen et al. argued that additional effects of familial risk on later literacy development can indicate that there are new genes involved in the development of word reading that are not yet active during the development of preliteracy skills.

There was also a difference between our study and previous studies: the effects of familial risk in our study were generally smaller (Snowling & Melby-Lervåg, 2016; van Viersen et al., 2018). This might be due to the fact that previous studies were based on unselected samples of children, whereas in our study only children at risk for reading problems were selected, resulting in a restriction of range. Although the effects were small, findings indicate that children with familial risk reach lower literacy skills than children without familial risk, even when receiving an intensive

evidence-based early literacy intervention in a one-to-one setting (see also Zijlstra et al., 2021).

Children with a familial risk for dyslexia, in fact a proxy for the ability to learn to read (van Bergen et al., 2017), showed less progress through the intervention during the first and second period, but had on average the same dosage as their peers without familial risk. These results align with the finding that progress during an intervention period, but not dosage, was related to preliteracy skills *prior* to the intervention period (see above). This supports once more the distinction between dosage and progress. Progress is related to children's capabilities to learn to read, whereas dosage is not.

Unlike the current results, previous research has found that children with familial risk received about 50% more intervention sessions (Hindson et al., 2005; Zijlstra et al., 2021). Zijlstra et al. might have reached this higher dose by providing schools with advice for each child individually. On a regular basis, they advised on the quantity of sessions and the individual progress within the intervention, based on the computer logs of the intervention (Zijlstra personal communication). In contrast, schools in our study implemented the intervention themselves, and children with and without familial risk apparently received the same dose. Under natural conditions, schools might not be aware that children with familial risk need to practice more and/or do not closely monitor children's progress to adjust the dose for each child accordingly.

4.2.2 *Parental Education*

Parental education was not related to literacy outcomes. This contradicts the findings of previous studies showing that children whose parents have a lower educational level tend to grow up in a poorer home literacy environment (Hamilton et al., 2016; Hemmerechts et al., 2017; Phillips & Lonigan, 2009), which in turn is associated with lower literacy skills in preschool, kindergarten and first grade (Hamilton et al., 2016; Petrill et al., 2005; Storch & Whitehurst, 2001). However, the direct relation between parental education and children's literacy skills in previous studies was small, .10 to .30 (Krijnen et al., 2020; Leseman & de Jong, 1998; Segers et al., 2016) and it is likely that in our sample, the effect might have been underestimated, because of our selected sample of children at risk for reading problems (restriction of range). Alternatively, the absence of an effect of parental education could also indicate that the intervention was equally effective for children whose parents have a higher or lower educational level. This is supported by the fact that we did not find

relations between parental education and dosage or progress within the intervention either, indicating that children from various backgrounds spent the same amount of time on the intervention and finished the same number of new intervention lessons. In contrast to previous research (Manz et al., 2010), our findings suggests that an intensive individually tailored intervention might improve the reading ability of all children at risk for reading difficulties, irrespective of the educational level of their parents.

4.3 *Limitations and Suggestions for Future Research*

The current findings indicate that the intervention *Build!* seems to foster children's literacy development. Our findings do not imply that it is more effective than the regular classroom instruction, as this study was not an RCT. Note that previous RCTs on *Build!* do support this conclusion, at least for a subgroup of children (Regtvoort et al., 2013; Regtvoort & van der Leij, 2007; Zijlstra et al., 2021).

Another limitation of this study is the focus on only one dimension of treatment integrity, dosage, although we would argue that it is a very important dimension in computer-assisted interventions. Given the small number of studies on school-based implementations of computer-assisted reading interventions, our study can be seen as an important first step in investigating the role of treatment integrity in such implementations. Future research could include other dimensions of treatment integrity as well.

As a third limitation of this study we have to mention that it was conducted in only two districts in the Netherlands. The sample contained relatively few urban schools, few non-Dutch children and many parents with lower and average educational levels compared to what is representative in the Netherlands (CBS, 2022a; 2022b). Research with more diverse samples is needed to investigate the generalizability of our findings.

Regarding avenues for future research, our study suggests that high levels of dosage can contribute to the success of a reading intervention in natural school settings. However little is yet known about the support schools need to reach high levels of dosage. Not for dosage, but for adherence, the study of (Stein et al., 2008) showed that teachers need multiple sources of support (training, meetings with other implementers, and on-site assistance) to implement an intervention with fidelity, contributing to higher reading outcomes. It would be interesting to see whether these findings also apply to dosage, and what other school factors can contribute to reaching a sufficient dose.

It has often been suggested that early literacy interventions should be continued for several years to avoid that effects of early intervention periods fade way (Bailey et al., 2017; Regtvoort & van der Leij, 2007). However, we did not find that more prolonged practice led to higher (maintenance of) reading levels after the middle of first grade. This finding raises the question how long long-lasting interventions should be continued. Findings are mixed. Some earlier research has shown that prolonged interventions with 100 sessions or more are effective in reducing reading problems in the long run (Scammacca et al., 2007). Moreover, there might be an accumulating effect of intervention years. Connor et al. (2013) showed that children who received individualized reading instruction for three subsequent years (grade 1 to 3) had better reading skills by the end of third grade than children who received fewer years of such instruction. However, there was no difference between children who received one or two years of individual reading instruction. Similarly, the meta-analysis of Wanzek and Vaughn (2007) indicates that early literacy interventions lasting for more than one school year are equally effective as shorter interventions. Given these inconsistent findings and the few studies on this topic, more research is needed to examine for how long a reading intervention should be continued to establish and maintain effects.

4.4 *Practical Implications*

This is one of the first studies to examine the relation between dosage of an intervention program and the development of literacy outcomes in natural school settings. Findings suggest that schools can reach higher literacy levels when providing children with more time within the intervention. The study also clearly shows the importance of distinguishing between dosage and progress. Especially children with familial risk for dyslexia need to spent more time on the intervention to make the same progress and reach the same levels of literacy skills as children without familial risk. Thus, in addition to monitoring the number of intervention sessions per week, schools should also focus on children's progress within the intervention. Schools can use information on progress to adjust the number of intervention sessions for children whose progress lags behind. Program developers could assist schools by providing the desired amount of progress per month and facilitating an easy comparison between this standard and the individual child's performance.

4.5 *Conclusion*

In this study, we evaluated the relation between dosage and the outcomes of a school-based implementation of the early literacy intervention *Build!*. More exposure to the intervention was associated with completing more new intervention lessons, which in turn was associated with more letter knowledge and better phonological awareness at the end of kindergarten, better word reading accuracy at the beginning of first grade, and better word reading fluency at the middle of first grade. The findings indicate that the relation between dosage and intervention outcomes is mediated by progress within the intervention. Therefore, dosage clearly appears a relevant dimension of treatment integrity and it is worthwhile for schools to monitor and stimulate frequent practice when implementing an intervention at school. The results also showed that parental education was unrelated to intervention outcomes. Children with familial risk, however, showed less progress within the intervention, and performed more poorly on preliteracy skills, as well as later reading skills. These children need more practice to finish the same amount of new intervention lessons. Schools might not be aware of this special need yet, and could improve intervention outcomes by providing children with familial risk with more practice than their classmates without such risk.

Supplemental Materials Chapter 3

Table S3.1 contains descriptive statistics of children with and without familial risk: literacy scores, dosage, and progress within the intervention, per measurement wave or intervention period.

Table S3.1

Differences between Children With and Without Familial Risk in the Full Sample

Measure	Wave/Period	No familial risk			Familial risk				
		n	*M*	*SD*	*n*	*M*	*SD*	*p*	*Cohen's d*
LK	K_{mid}	187	6.92	3.94	53	6.78	4.18	.814	0.04
	K_{end}	215	14.67	5.35	63	12.94	5.27	*.024*	0.33
PA	K_{mid}[a]	193	23.23	8.25	54	21.93	7.51	*.297*	0.16
	K_{end}[b]	216	8.83	3.78	62	7.72	4.06	*.047*	0.29
WRA	$G1_{start}$	232	9.54	4.72	69	9.28	3.92	*.673*	0.06
	$G1_{mid}$	233	12.98	8.92	71	10.39	7.46	*.027*	0.30
	$G1_{end}$	231	24.94	9.95	73	21.53	9.56	*.010*	0.35
WRF	$G1_{mid}$	231	12.33	6.93	69	10.17	6.49	*.022*	0.32
	$G1_{end}$	226	23.48	13.14	71	19.53	12.04	*.025*	0.31
	$G2_{mid}$	163	39.39	17.76	51	31.43	16.21	*.005*	0.46
Dosage	K_{mid}-K_{end}	254	6.21	3.24	79	6.24	3.18	*.930*	-0.01
	$G1_{start}$-$G1_{mid}$	254	5.99	2.60	79	6.13	2.67	*.693*	-.0.05
Progress	K_{mid}-K_{end}	252	41.56	23.73	79	37.33	20.30	*.154*	0.18
	$G1_{start}$-$G1_{mid}$	255	62.35	30.54	79	59.49	26.40	*.455*	0.10

Note. LK = letter knowledge; PA = phonological awareness; WRA = word reading accuracy; WRF = word reading fluency; K = kindergarten; G1 = first grade.
[a]full test. [b]part of the test.


Mijn Groot
KAREL KLUS
Surprise
lees

CHAPTER 4

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

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 poor readers and improved mean reading skills at the school level. Transfer effects to spelling and reading comprehension were also examined. During six years schools not implementing *Build!* (61-126 schools, depending on the outcome measure) were compared to 72-145 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 two 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.

van der Weijden, F. A., van den Boer, M., Zijlstra, B. J. H., & de Jong, P. F. (2024c). Implementation takes time: Reduction of literacy problems in schools implementing an early-literacy intervention. *Journal of Research on Educational Effectiveness*, 1–33. http://dx.doi.org/10.1080/19345747.2024.2384365

1 Introduction

In primary school, 3% to 10% of the children develops 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 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 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 poor readers.

Build! (in Dutch: *Bouw!*) is a 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!* thereby 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 beginning 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 10-15 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 mid-Grade 1 to mid-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 that completed the intervention had better abilities in word reading and reading comprehension than the no-intervention control group at post-test and one year after the intervention had finished. The effect of the *Build!* intervention, from the second year of kindergarten through mid-Grade 2, was tested in an RCT by Zijlstra et al. (2021, see also Zijlstra, 2015). The intervention and control group were followed until the end of second grade, half a year after the intervention had finished. The results showed that the intervention was only effective in a subgroup of children. This subgroup concerned children whose parents had provided information about the prevalence of dyslexia in the family, 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, that is four years after the intervention had finished. Also 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 begun 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 5000 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.

1.1 *Effects of Early-Literacy Interventions*

Current evidence suggests that interventions for children at risk for or with reading diffuclties 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, i.e. around one year after the intervention had finished (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 asked for, as reading difficulties might continue or even emerge after second grade (Simmons et al., 2008; Torppa et al., 2015).

It seems plausible that it might 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 pre-literacy 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 appeals to time, resources and leadership (Durlak & DuPre, 2008) and it might take more than one year to reach full treatment integrity, i.e. to implement the intervention as intended (Gresham et al., 2000).

To our knowledge, there are hardly any studies that have considered the effects of a literacy intervention over time, that is in subsequent cohorts. As an exception, Torgesen (2009) showed that in a large-scale implementation of *the response to intervention instructional model*, intervention effects increased during three 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!*.

1.2 *Difference-in-Difference 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. Apart from the costs of large-scale RTCs, schools are inclined to implement a promising intervention and do not want to wait 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 whether schools showed an improvement in the mean reading ability and/or a decrease in the number of poor readers 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 a 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 & 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, but also at the school level. The difference-in-difference is defined as the difference between pre- and posttest in the intervention group minus the difference between pre- and posttest in the control group. It is literally a 'difference of differences' (Fredriksson & Oliveira, 2019). The design assumes that the gains in the control group and intervention group are similar in the absence of the intervention.

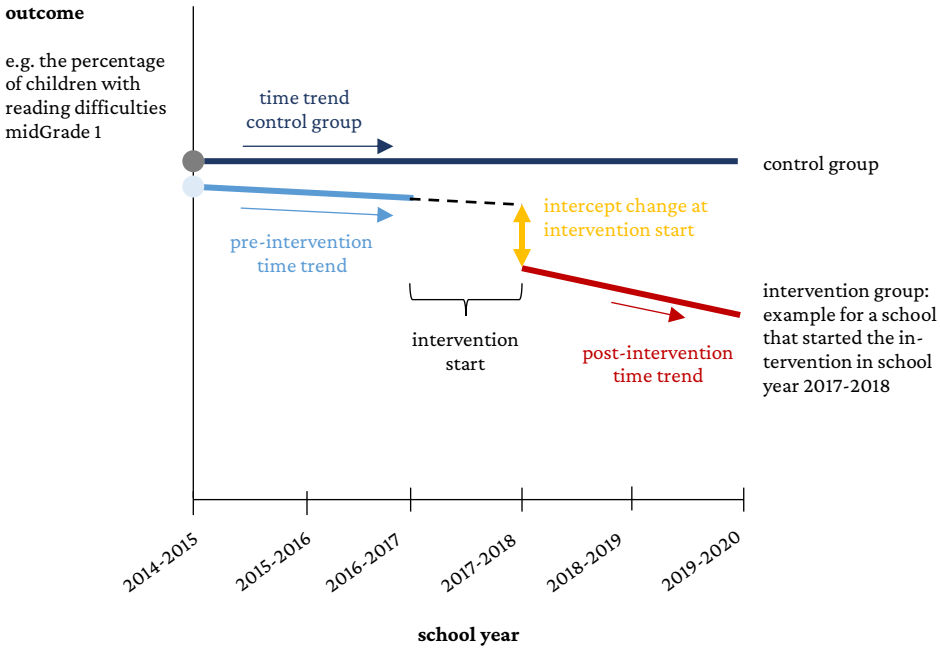
The DiD design can be extended by the addition of multiple pre- and posttests, resulting in a Comparative Interrupted Time Series (CITS) design (Jacob et al., 2016). This CITS design controls for the potential threat that larger gains in the intervention group are due to differences in pre-intervention 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 4.1.

A potential threat to the CITS-design is that the control and intervention group are not sufficiently comparable, for example when co-occuring 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 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-occuring events that differently influence the intervention and control group.

Figure 4.1

Difference-in-Difference Design With Multiple Pre- and Posttests



Note. A difference-in-difference model with multiple pre- 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).

1.3 *Current Study*

The main question of this study was whether the early-literacy intervention *Build!* led to an improvement in reading skills and a reduction of poor readers after implementation. We examined short-term effects in first and second grade: that is during the intervention (mid-Grade 1, end-Grade 1) and at the planned end of the intervention (mid-Grade 2). In addition, we investigated follow-up effects half a year and one year after the planned end of the intervention (end-Grade 2 and mid-Grade 3 respectively). Furthermore, it was investigated whether the effects were visible after schools had been using the program for one 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. First, we investigated whether the effects of the intervention would transfer to spelling, because 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 in Regtvoort et al. (2013). We also examined effects on an unrelated skill, i.e. 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 of school achievement.

2 Method

2.1 *Design*

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

2.2 *Participants*

Three hundred and eighteen schools, clustered in 26 school boards and 4 geographical locations, were asked to participate in the study. Schools were located in five (out of twelve) provinces in the west and middle parts of the Netherlands. Half of the schools were located in villages, and half of the schools in cities (mostly large cities, i.e. 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., 2024a). 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 socio-economic 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 differ 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 the national population (M = 29.84, SD = 3.91), t (6568) = 1.51, p = .131, and the 55 schools (M = 28.76, SD = 2.72), t (337) = 1.02, p = .301. The average school size of the 318 schools (M = 249.52, SD = 139.28) was somewhat larger than the average school size of the 55 schools (M = 198.23, 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 was 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 for which cohorts *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.

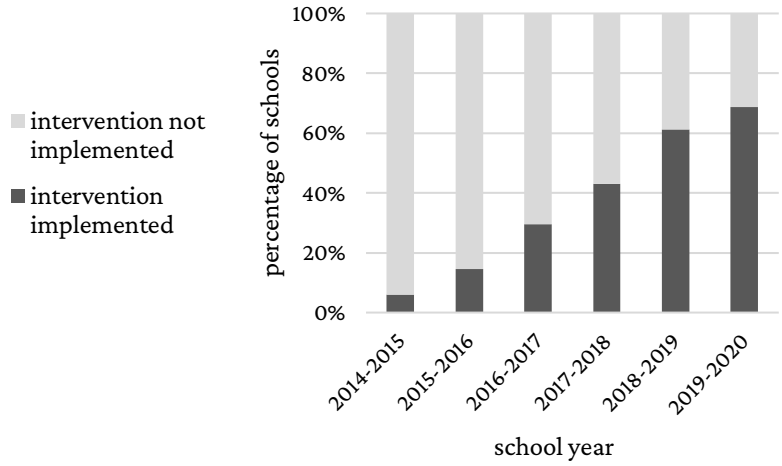
The average school weights of the final samples varied from 29.70 to 29.80 (*SDs* varied from 4.65 to 4.67), which were similar to the average school weight of the 318 approached schools (M = 29.48, SD = 4.92). The average school sizes of the final samples, varying from 253.60 to 259.59 (*SDs* varied 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 two to three 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 onwards. On all other occasions, the number of schools for reading comprehension varied between 143 and 206. Data on mathematics was 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 4.2.

Figure 4.2

Percentage of Schools Implementing Build! over Time



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 received 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 mid and end Grade 1 that had received 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; mathematics: 61-107). In the control group there were always 48-58 schools that did not implement the intervention at all. All schools of this part of the control group and of 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 group (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 of six cohorts of children should be 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.

2.3 *Intervention*

The intervention program *Build!* (in Dutch: *Bouw!*) 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*, zeezout - *sea salt*). In Parts 11 and 12, children learn to read two-syllable words with open syllables (e.g. letter - *letter*, rozen - *roses*) and closed syllables (e.g. winter - *winter*). From Part 2 onwards, there are reading exercises with and without time limit.

The instruction is characterized by (a) direct instruction (Stockard et al., 2018), (b) direct feedback (Wisniewski et al., 2020), and (c) the minimal pairing technique (McCandliss et al., 2001). This technique requires children to begin 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 non-professional (e.g. a parent, volunteer, or older student). Previous research has shown that the majority of professional and non-professional 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 two-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 four years old and follow two 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 of 10 to 15 minutes each.

2.4 *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).

2.4.1 *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 until 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: (1) whether the intervention was

used in the particular cohort, and (2) 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 onwards, 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 recieved *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 more previous years.

2.4.2 *Outcome Measures*

Within schools, we determined per cohort, outcome, and measurement occasion (1) the percentage of children with difficulties (i.e. children scoring below the 25th percentile based on national norms), and (2) 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 can 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 (for example 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 (COTAN): reliability and validity were sufficient for all tests (Egberink et al., 2009-2023). Schools administered the tests twice a year, i.e. between mid-January and mid-February and between mid-May and the end of June. Reading comprehension was measured from end-Grade 1 onwards. 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 to 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 one minute, without making errors.

In the middle of Grade 1, only Cards 1 and 2 were administered. End Grade 1 and in Grade 2, all three cards were used. From Grade 3 onwards, 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 level D and E into 'difficulties in reading fluency' and level A, B, and C to '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 per version. A dummy was created to account for test version.

Spelling

Spelling was measured with *Cito Spelling* (Cito, 2014b; de Wijs, 2010). It consisted of two parts, administered at a different (part of the) day. 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 two 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* (Feenstra et al., 2010; Cito, 2014a). The test consisted of two parts, administered at a different (part of the) day. 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, i.e. 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, 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, 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* (Janssen et al., 2010; Cito, 2013). The test consisted of two to three parts, administered at a different (part of the) day. 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.

2.5 *Analyses*

2.5.1 *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 (mid and end). For each outcome we analyzed the data per grade and measurement occasion. Data was not analyzed for end-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). In Grade 3, that was particularly problematic because in this grade few cohorts had received *Build!* from kindergarten or Grade 1 onwards while the schools had been using *Build!* for two years or longer (needed to estimate the post-intervention time trend; see Difference-in-Difference Models).

2.5.2 *Difference-in-Difference Models*

Our main interest concerned time trends, that 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 onwards. For mathematics we did not foresee such a change in time trend.

We used a difference-in-difference model to estimate changes in outcomes over time (Mascha & Sessler, 2019). The model is presented in Figure 4.1. 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 Mascha and Sessler's (2019) eighth formula, we used the following model to estimate all the parameters of interest:

$$y_t = \beta_0 + \beta_1\, test\ version + \beta_2\, intervention\ group + \beta_3\, school\ year\ + \beta_4\, intervention\ group * school\ year + \beta_5\, implementation\ of\ the\ intervention + \beta_6\, number\ of\ years\ that\ Build\ is\ used,$$

in which β_0 is the intercept for schools in the control group, β_2 is the difference in intercepts for schools in the intervention group compared to those in the control group, β_3 is the general time trend in the control group, β_4 is the difference in pre-intervention time trend in the intervention group compared to the control group, β_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 β_6 is the difference in the post-intervention time trend in the intervention group from the second intervention year onwards (measuring the effect of the number of years that *Build!* is used) compared to the pre-intervention time trend. Note that the colors in this formula correspond to the colors in Figure 4.1, which provides an illustration of the model and its parameters.

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, β_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 over- 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 fitted the data better for one of the measurement occasions (e.g. mid-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

First, it was evaluated whether there was a change in test scores in the first year *Build!* was introduced, that is whether β_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 β_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 β_6. Cohen (2013) suggested that a R^2 of .01 can be considered small, a R^2 of .09 can be considered medium and, a 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 did not only include test version as a predictor, but also allowed the models to estimate the variance for each test version separately, and ran the models for one test version too.

3 Results

3.1 *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 poor readers from Grade 3 onwards. As a result, some cohorts had more than 80% poor readers, 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 mid-Grade 3 included fewer schools. Across school years, the number of schools varied between 136 to 195 at the middle of each grade (except for reading fluency mid-Grade 3: 41 to 160 schools) and between 147 to 195 at the end of each grade.

3.2 *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, school year 2018-2019 at the end of each grade) are shown in Table 4.1, per outcome and measurement occasion and split into schools that did not use *Build!*, used *Build!* for one or two years, and used *Build!* for three 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 4.2.

Table 4.1

Descriptive Statistics for the Percentage of Children with Difficulties at the End of the Study

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

[a]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).

Cohen's *d*: [b] Difference between schools without *Build!* and schools that used *Build!* for 1 or 2 years, divided by the pooled SD. [c] 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 schools without *Build!*.

Table 4.2

Descriptives for the Mean Ability at the End of the Study

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

[a]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).

Cohen's *d*: [b] Difference between schools without *Build!* and schools that used *Build!* for 1 or 2 years, divided by the pooled SD. [c] 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 schools without *Build!*.

Regarding reading fluency, schools that had used *Build!* for one or two 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 three or more years showed fewer children with difficulties and a higher mean ability than schools without *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 one or two years (small difference) and schools that used *Build!* for three or more years (medium difference). These descriptive statistics give the impression that *Build!* may also have affected mathematics, but Tables 4.1 and 4.2 do not show whether the differences between groups already existed at the beginning of the study (before the intervention had started). Therefore, Tables 4.1 and 4.2 do not show whether any differences between groups emerged during the study and might be related to the introduction of *Build!*. We need the difference-in-difference models (controlling for pre-existing 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!*.

3.3 *Difference-in-Difference Models*

We ran separate difference-in-difference 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 (β_2) and (2) the difference between the pre-intervention time trend of the control group and the intervention group (β_4). Especially, 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 5 (end of a grade) or 6 (middle of a grade) cohorts. The number of pre-intervention cohorts in the intervention group varied between 1 and 4 at the end of a grade, and between 1 and 5 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!* (β_5)

and (2) the change in time trend when *Build!* had been used for two or more years (β_6). Our main results concern the second parameter, i.e. the post-intervention time trend. For reasons of clarity we present the parameter estimates for this parameter in one table (Table 4.3). We present them for each test version separately and the two test versions together. In the Supplemental Materials Chapter 4, all parameter estimates of the difference-in-difference models are provided (see Table S4.1-S4.4, Model 1). We also ran analyses with one of the test versions. The results are also listed in the Supplementary Material Chapter 4 (Table S4.5-S4.6). In what follows, we describe patterns per outcome, rather than significant differences per measurement occasion, and summarize findings across test versions.

Table 4.3
Post-Intervention Time Trends: Changes in Test Scores After Two Years Using Build!

	Percentage of Children With Difficulties					Mean Ability				
	$G1_{mid}$	$G1_{end}$	$G2_{mid}$	$G2_{end}$	$G3_{mid}$	$G1_{mid}$	$G1_{end}$	$G2_{mid}$	$G2_{end}$	$G3_{mid}$
Reading Fluency										
Test Versions Together	**-1.39****	-0.84	**-1.10***	-1.27	-0.38	0.30	**0.70***	**0.62***	**1.19****	0.83
Old Test Version	-2.05	**-2.51***	**-2.44***	**-2.72***	-4.74	**0.90***	**1.68****	**1.84****	**1.94****	2.04
New Test Version	**-2.07****	-0.75	**-1.36***	-1.46	-0.51	**0.39***	**0.83***	**0.99****	**1.54****	1.14
Spelling										
Test Versions Together	**-1.44***	**-1.76***	-1.22	-1.82	**-2.07***	**1.84***	1.75	**3.05****	**4.26****	**3.34****
New Test Version	**-1.65****	**-2.06****	**-1.50***	-1.61	**-2.27***	**3.30****	**2.95****	**2.84****	**3.11***	**3.19****
Reading Comprehension										
Test Versions Together	-	**-2.26***	-1.17	**-2.17***	**-2.08***	-	1.48	1.11	**1.61***	0.52
New Test Version	-	**-3.15****	-1.58	**-2.38***	**-2.67***	-	**2.23****	1.30	**1.72***	1.33
Mathematics										
Test Versions Together	-0.08	-1.50	-0.08	0.08	-1.50	-0.04	0.96	0.14	0.97	0.82
New Test Version	-0.57	**-1.66***	-0.06	0.18	-1.67	0.06	0.84	0.17	0.40	1.06

$^{**}p < .01$. $^{*}p < .05$

3.3.1 *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 Chapter 4, Table S4.1, Model 1, intercept$_{int,}$ β_2). The pre-intervention time trend in the intervention and control group was similar (see Supplemental Materials Chapter 4, Table S4.1, Model 1, time trend pre$_{int,}$ β_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 Chapter 4, Table S4.1 Model 1 and Table S4.5, time trend$_{control,}$ β_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 one year (see Supplemental Materials Chapter 4, Table S4.1, Model 1, intervention start$_{int,}$ β_5). However, overall there was a post-intervention time trend: the percentage of children with difficulties decreased and the mean ability increased from two years of *Build!* onward. Depending on test version, this effect was significant from mid or end-Grade 1 to mid or end-Grade 2 (β_6, see Table 4.3). From the second year, for schools that had been using *Build!*, the percentage of children with difficulties was stable or dropped with 1% per school year, while it increased with 1% per school year before *Build!* had been introduced. Put differently, the percentage of poor readers after two years of intervention was 1% to 2% lower than expected based on the pre-intervention slope. The mean ability increased with 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 Figures 4.3A en 4.3B we provide an illustration of time trends in the percentage of children with difficulties and the mean ability mid-Grade 2, when the intervention should have been finished. Based on the model parameters in Supplemental Materials Chapter 4, Table S4.1 (Model 1) and the formula in the Method section (see Analyses, Difference-in-Difference Models), the graphs illustrate time trends for schools that did not use *Build!*, schools that used *Build!* from 2016 onwards, and schools that used the program from 2017 onwards. 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 one year. However, after *Build!* was used for two years, time trends positively changed in the intervention groups, while the control group continued to have a negative time trend.

3.3.2 *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 Chapter 4, Table S4.2, Model 1, intercept$_{int,}$ β_2). This effect is most pronounced for the new test version (see Supplemental Materials Chapter 4, Table S4.6, intercept$_{int,}$ β_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 Chapter 4, Table S4.2, Model 1, time trend pre$_{int,}$ β_4). In both groups, the percentage was stable over time (see Supplemental Materials Chapter 4, Table S4.2, Model 1, time trend$_{control,}$ β_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 Chapter 4, Table S4.6, time trend pre$_{int,}$ β_4). Here, the mean ability slightly increased over time in the control group (see Supplemental Materials Chapter 4, Table S4.6, time trend$_{control,}$ β_3), whereas it remained stable in the intervention group (adding time trend$_{control,}$ β_3 and time trend pre$_{int,}$ β_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 one year (see Supplemental Materials Chapter 4, Table S4.2, Model 1, intervention start$_{int,}$ β_5). After implementing *Build!* for two 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 (β_6, see Table 4.3). From the second year *Build!* was used, the percentage of children with difficulties dropped with 1% to 3% per school year, and the mean ability increased with around 3 points per school year. The post-intervention time trend explained 1% to 4% of the variance in spelling (a small effect).

See Figures 4.3C en 4.3D for an illustration of the effects at the planned end of the intervention, i.e. mid-Grade 2. Graphs were created in the same way as Figures 4.3A and 4.3B, but based on the parameters in the Supplemental Materials Chapter 4,

Table S4.2 (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 onwards. Nonetheless, there was a clear change in time trends, not immediately after *Build!* was introduced, but after two years. The intervention group showed more favorable outcomes at the end of the research period.

3.3.3 *Reading Comprehension*

Pre-Intervention differences

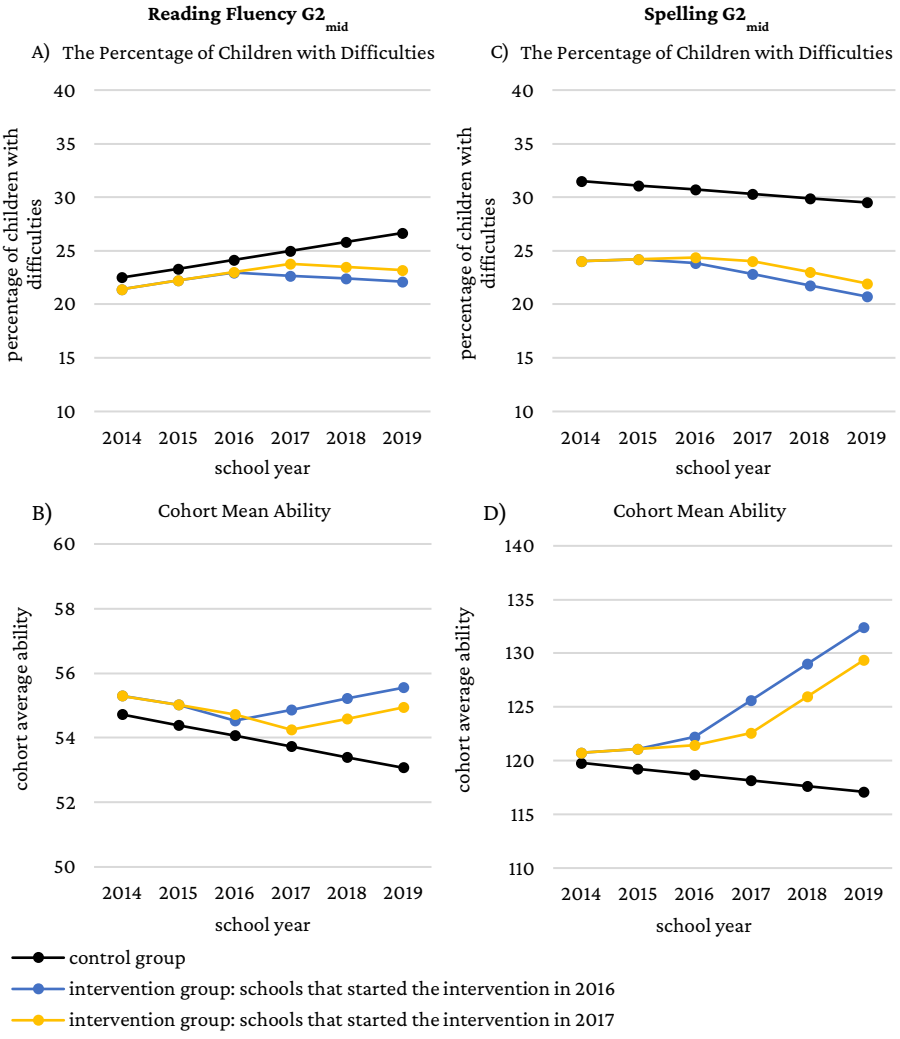
The intervention group had fewer children with difficulties and a higher mean ability on reading comprehension before the intervention had started (see Supplemental Materials Chapter 4, Table S4.3, Model 1, intercept$_{int,}$ β_2). The pre-intervention time trend in the control group and intervention group was similar (see Supplemental Materials Chapter 4, Table S4.3, Model 1, time trend pre$_{int,}$ β_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 Chapter 4, Table S4.3 Model 1 and Table S4.6, time trend$_{control,}$ β_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 one year (see Supplemental Materials Chapter 4, Table S4.3, Model 1, intervention start$_{int\,,}$ β_5). However, after implementing *Build!* for two years, the post-intervention trend for the percentage of children with difficulties changed. While it increased with 1% or 2% before *Build!* was introduced, it kept stable or decreased with 1% after implementing *Build!* for two years (post-intervention slope was around 2% lower than the pre-intervention slope). This effect was significant for most measurement occasions, end-Grade 1 to mid-Grade 3 (β_6, see Table 4.3). 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 did mostly not change after implementing *Build!* for two years.

Figure 4.3

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) mid-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 onwards, and schools that used the program from 2017 onwards. Graphs are based on the model parameters in Supplemental Materials Chapter 4, Table S4.1 and S4.2 (two test versions) and the formula in the Method section (see Analyses, Difference-in-Difference Models).

3.3.4 *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 Chapter 4, Table S4.4, Model 1, intercept$_{int,}$ β_2). The pre-intervention time trend in the control group and intervention group was mostly similar (see Supplemental Materials Chapter 4, Table S4.4, Model 1, time trend pre$_{int,}$ β_4). In both groups, the percentage of children with difficulties and the mean ability were highly stable over time (see Supplemental Materials Chapter 4, Table S4.4, time trend$_{control,}$ β_3).

Intervention Effects

As expected, there was no change in mathematics scores when the school had been using *Build!* for one year (see Supplemental Materials Chapter 4, Table S4.4, Model 1, intervention start$_{int,}$ β_5), nor after implementing *Build!* for two years (β_6, see Table 4.3).

3.4 *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 S4.5 and S4.6). However, it should be acknowledged that pre-intervention trends were more affected by schools that started later with the intervention (i.e. late-adopters) and that post-intervention trends were more affected by schools that started earlier with the intervention (i.e. early-adopters), because late-adopters had inevitably 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 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 one year, and/or the intervention effect after two 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 (β_0), test version (β_1), and school year (β_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 (β_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 years, but the post-intervention trend started from the second year of implementation. Thus, in principle we could split the intervention group into five groups, according to the number of years a school worked with *Build!*. However, the number of schools that had implemented the intervention for four, five or six years was small. Therefore, to increase the power of our analyses, we distinguished

between schools that had implemented the intervention for two years (late-adopters) and those that had used the program for three or more years (early-adopters). If the post-intervention trends would differ between these groups, we should find a difference between the groups after one and/or after two years of implementation. We took several steps to test these effects. We started to extend our original model (Model 1 in Tables A1 to A5) by splitting the post-intervention trend (β_5) into two dummy variables: two years and three years or more after implementation. Note that the effect after one 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 to 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 these tests was significant (see Supplementary Materials, Table S4.7), indicating that this model did mostly 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 three years or more were considered early-adopters, schools that used the intervention for less than three years were late-adopters. Depending on the outcome and timepoint, 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 one year) and the interaction of adopter group with the effect after two years. A significant interaction of adopter group with the immediate effect would imply a difference between early- and late-adopters after one year of implementation. More importantly, a significant interaction of adopter group with the effect after two 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 two years, being 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 three or more years of intervention, because the late-adopters had only two 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 S4.7), 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 mid-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 proceeds, were not confounded by adopter group.

4 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 poor readers 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 difference-in-difference model, specifically a Comparative Interrupted Time Series (CITS) model. In a sample of 207 schools and over six 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 one year and for two 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 two or more years. More specifically, after schools used the intervention for at least two years, there were decreases in the percentage of children with difficulties in reading fluency (mid/end-Grade 1 to mid/end-Grade 2), spelling (mid-Grade1 to mid-Grade 3), and reading comprehension (end-Grade 1 to mid-Grade 3). Moreover, the mean ability on reading fluency (mid-Grade 1 to end-Grade 2) and spelling (mid-Grade 1 to mid-Grade 3) begun to increase. A contrary trend was shown 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 pre-existing differences between schools that did and did not implement *Build!*. Schools that used *Build!* showed, before the intervention was introduced, a lower

percentage of children with difficulties and a higher cohort average performance on spelling, reading comprehension, and mathematics, but not on reading fluency, than schools that did not use the intervention. This indicates that schools with *Build!* were initially 'better' schools. However, we hardly found 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 subjects 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 due to pre-existing differences across schools.

4.1 *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.

4.1.1 *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 effect are small. However, larger effects can hardly be expected. First, the target group of *Build!* is small. The mean ability was influenced by many more children than only the small group of children who were provided with *Build!*. The intervention is meant for children who are at risk of reading problems, i.e. the 25% lowest scoring children in kindergarten or Grade 1, but the effects were examined at the school level, including performance of all the children. Second, the selection of children is never perfect, and thus not all of the 25% poorest readers might have received the intervention. Especially in kindergarten (before formal reading instruction has begun), 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 always results in under- 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 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 non-professional tutor at school), while some children need Tier 3 instruction (individual support from a specialist). Suppose that only half of the 25% lowest scoring readers

were provided with *Build!* and that it was effective for 25% of these children. Then, the percentage of poor readers would drop from 25% to 22%, which is similar to the drop of poor readers mid-Grade 1 in our study after *Build!* had been used for four years, from 25.51% to 22.45%. The cohort average ability on reading fluency would increase, for example mid-Grade 1, with around half a point, similar to the increase in our study after *Build!* had been used for three years, from 22.83 to 23.38 points. Please note that with over one million children in primary schools in the Netherlands, a drop in poor readers of 3% would refer to thousands of children for whom severe literacy problems have been prevented.

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

4.1.2 *Alternative Explanations for the Effects of Build!*

Probably, the small effects of *Build!* on literacy skills were not only caused 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 (for reading fluency and reading comprehension) or even larger effects on the mean ability than on the percentage of children with difficulties (for spelling), suggests that the implementation of *Build!* might be associated with a development within schools that affected *all* children. Possibly, the implementation of *Build!* created more attention for reading and spelling within schools and/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 could have been an incentive to provide all students with better instruction and to provide extra help to the students who do not reach the target level. Similarly, Torgesen (2009) suggested that the implementation of the RTI instructional model in his study might have caused a change of behavior within the schools. Torgesen speculated that teachers and schools could 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 poor readers. 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 suggest 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 timepoints, making it less likely that the increase in test scores was due to a particular event at a particular time. Fourth, such a larger movement can be expected to begin before the school had been using *Build!*, while we did not find that reading or spelling improved before *Build!* was used, neither after it was used for one 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.

4.1.3 *Long-Term Effects*

Few studies have examined long-term effects of early-literacy interventions, i.e. 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 end-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, i.e. mid-Grade 3, one 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 received *Build!* for two or more years. This was even more pressing for reading fluency as we had to filter out schools that tested only the poor readers. Thus it might just have been a matter of power.

Alternatively, a long-term decrease of reading difficulties might 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 took two years (covering the pre-reading and reading phase in kindergarten, Grade 1 and 2) and earlier research showed that 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 two years. A previous study (van der Weijden et al., 2024a) showed that many children stop the intervention in Grade 1 instead of Grade 2.

4.1.4 *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 of letter-sound correspondences, which has been shown to improve spelling in the short and long term (Ehri et al., 2001a; Suggate, 2016). To a smaller extent, the intervention includes phonological training, i.e. phoneme blending and reading with the minimal pairing technique. From around the middle of Grade 1, children read words with time limit. Both phonological training and reading fluency training have been shown to improve spelling skills (Ehri et al., 2001b; Suggate, 2016). Reading comprehension as well 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).

4.2 *Years Since Implementation*

The most remarkable finding of this study is that effects only emerged after *Build!* had been used for two years or longer. To the best or our knowledge, there was yet no (quasi-)experimental study to support the idea that interventions become more effective, when schools use them for a longer time. Note that many large-scale studies on K-12 interventions do not show any effects (Lortie-Forgues & Inglis, 2019), maybe 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, the intervention can 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.

4.3 *Limitations and Suggestions for Future Research*

In this study, we used a difference-in-difference 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 were 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 poor readers after three 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 and motivated to implement the intervention and did not have higher intervention outcomes in the first two years after the implementation of the intervention. 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 because we controlled for differences between schools at the beginning of the study. We measured changes within schools, and at each timepoint we studied whether the time trends changed.

Our findings suggest that future large-scale studies on evidence-based (reading) interventions, should not stop after one year of implementation. Instead, effects might only appear after two or three years of implementing the intervention. Although large-scale studies are time-consuming and costly, it would be worthwhile to include multiple cohorts and/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.

4.4 *Practical Implications*

Our findings suggest that *Build!* is slightly more effective than business-as-usual education in decreasing literacy problems. Findings indicate that schools who (want to) use an evidence-based (reading) intervention, cannot expect an immediate effect after implementing the intervention for one 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 decrease reading and spelling problems in schools.

4.5 *Conclusion*

In summary, our findings indicated that, after two 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 after schools have had several years of experience with the intervention.

Table S4.1
Difference-in-Difference Models for Reading Fluency: Test Versions Together

	Percentage of Children With Difficulties					Mean Ability				
	$G1_{mid}$	$G1_{end}$	$G2_{mid}$	$G2_{end}$	$G3_{mid}$	$G1_{mid}$	$G1_{end}$	$G2_{mid}$	$G2_{end}$	$G3_{mid}$
Model 1										
$Intercept_{control}$ (β_0)	**25.51****	**24.74****	**22.51****	**25.40****	**24.14****	**22.83****	**36.20****	**54.72****	**62.07****	**71.68****
$Intercept_{int}$ (β_2)	-1.51	-0.04	-1.09	-2.78	-0.68	**1.17***	0.54	0.57	1.32	0.08
$Intercept_{new\ test\ version}$ (β_1)	**4.90****	**-6.19****	**3.11****	**-6.13****	**-4.21****	**-6.91****	**-6.96****	**-9.59****	**-8.07****	**-7.20****
Time $trend_{control}$ (β_3)	**0.98***	**1.17***	**0.83***	**0.53**	0.75	-0.13	-0.24	-0.33	-0.2	-0.14
Time trend pre_{int} (β_4)	-0.39	-0.91	-0.02	0.40	0.03	-0.17	0.02	0.05	-0.07	-0.01
Intervention $start_{int}$ (β_5)	1.11	0.45	-0.07	0.35	0.19	-0.05	-0.08	-0.2	-0.68	-0.34
Time trend $post_{int}$ (β_6)	**-1.39****	-0.84	**-1.10***	-1.27	-0.38	0.30	**0.70***	**0.62***	**1.19****	0.83
Model 2										
Intervention $start_{int}$ (β_5)	1.20	0.32	0.34	2.09	1.49	-0.09	-0.31	-0.73	**-1.66***	-0.95
Intervention 2nd year	-0.51	0.03	-1.62	**-5.64****	-3.83	0.15	0.86	**1.78***	**3.57****	2.19
Intervention ≥ 3rd year	**-3.23***	-2.91	-3.11	-3.67	-0.70	0.88	**2.10****	**1.81***	**2.63****	2.45

Note. $Intercept_{control}$ = mean in 2014-2015 in the control group, $Intercept_{int}$ = how the mean in 2014-2015 in the intervention group differed from the one in the control group, $Intercept_{new\ test\ version}$ = how the mean in 2014-2015 with the new test version differed from the one with the old test version, $Time\ trend_{control}$ = time trend in the control group, *Time trend* pre_{int} = how the pre-intervention time trend in the intervention group differed from the time trend in the control group, *Intervention* $start_{int}$ = the change after implementing *Build!* for one year, *Time trend* $post_{int}$ = how the time trend in the intervention group changed after implementing *Build! for two years*, *Intervention* 2^{nd} *year* = the change after implementing *Build! for* two years, *Intervention* ≥ 3^{rd} *year* = the change after implementing *Build!* for three or more years.
$^{**}p < .01$. $^{*}p < .05$.

Table S4.2
Difference-in-Difference Models for Spelling: Test Versions Together

	Percentage of Children With Difficulties					Mean Ability				
	$G1_{mid}$	$G1_{end}$	$G2_{mid}$	$G2_{end}$	$G3_{mid}$	$G1_{mid}$	$G1_{end}$	$G2_{mid}$	$G2_{end}$	$G3_{mid}$
Model 1										
$Intercept_{control}$ (β_0)	**28.86****	**31.93****	**31.51****	**29.22****	**25.02****	**106.94****	**113.56****	**119.8****	**121.55****	**127.51****
$Intercept_{int}$ (β_2)	-2.58	-4.10	**-7.48****	**-5.10****	-2.57	1.30	0.93	**0.94***	0.54	-0.05
$Intercept_{new\ test\ version}$ (β_1)	-1.88	**-4.04***	-1.69	**-3.52****	0.10	**46.58****	**90.80****	**118.77****	**147.27****	**166.45****
Time $trend_{control}$ (β_3)	1.02	-0.27	-0.40	-0.05	**1.30****	**-1.52***	-1.16	-0.54	-0.11	-0.37
Time trend pre_{int} (β_4)	-0.38	0.00	0.57	0.45	-0.57	1.17	**1.67***	0.89	0.56	**0.94****
Intervention $start_{int}$ (β_5)	-0.29	-0.13	-0.49	-0.20	0.47	-0.05	0.80	0.78	1.06	-0.08
Time trend $post_{int}$ (β_6)	**-1.44***	**-1.76***	-1.22	-1.82	**-2.07***	**1.84***	1.75	**3.05****	**4.26****	**3.34****
Model 2										
Intervention $start_{int}$ (β_5)	-0.52	-0.36	-0.53	1.13	1.16	0.92	2.38	0.30	0.13	-0.53
Intervention 2nd year	0.28	-0.11	-0.44	**-5.26***	-3.59	-0.57	-1.77	**4.65***	**9.19****	**4.49***
Intervention ≥ 3rd year	-2.94	-4.13	-3.36	-4.15	-4.70	**6.16****	**6.37***	**8.91****	**10.07****	**8.63****

Note. *$Intercept_{control}$* = mean in 2014-2015 in the control group, *$Intercept_{int}$* = how the mean in 2014-2015 in the intervention group differed from the one in the control group, *$Intercept_{new\ test\ version}$* = how the mean in 2014-2015 with the new test version differed from the one with the old test version, *$Time\ trend_{control}$* = time trend in the control group, *Time trend pre_{int}* = how the pre-intervention time trend in the intervention group differed from the time trend in the control group, *Intervention $start_{int}$* = the change after implementing *Build!* for one year, *Time trend $post_{int}$* = how the time trend in the intervention group changed after implementing *Build! for two years*, *Intervention 2nd year* = the change after implementing *Build! for* two years, *Intervention ≥ 3rd year* = the change after implementing *Build!* for three or more years.
$^{**}p < .01$. $^{*}p < .05$.

Table S4.3
Difference-in-Difference Models for Reading Comprehension: Test Versions Together

	Percentage of Children With Difficulties				Mean Ability			
	$G1_{end}$	$G2_{mid}$	$G2_{end}$	$G3_{mid}$	$G1_{end}$	$G2_{mid}$	$G2_{end}$	$G3_{mid}$
Model 1								
$Intercept_{control}$ (β_0)	**29.75***	**29.91****	**30.06****	**31.39****	**-2.42***	**11.09****	**16.59****	**23.56****
$Intercept_{int}$ (β_2)	-3.37	**-5.05***	**-8.95****	**-7.08****	**2.73***	**3.27***	**4.41****	**3.21****
$Intercept_{new\ test\ version}$ (β_1)	**-4.30***	**-4.06****	**-5.27****	**-3.00***	**118.50****	**124.05****	**124.04****	**128.43****
Time $trend_{control}$ (β_3)	1.36	**2.03****	**1.35****	**0.95***	-0.66	**-1.39****	**-1.09****	-0.43
Time trend pre_{int} (β_4)	-0.05	-0.26	0.07	0.03	0.32	0.09	0.85	0.43
Intervention $start_{int}$ (β_5)	-1.87	0.37	1.44	3.33	0.97	0.38	**-2.74***	-1.06
Time trend $post_{int}$ (β_6)	**-2.26***	-1.17	**-2.17***	**-2.08***	1.48	1.11	**1.61***	0.52
Model 2								
Intervention $start_{int}$ (β_5)	-2.27	1.17	2.62	**4.72***	0.81	-0.38	**-3.57****	-1.96
Intervention 2nd year	0.23	-2.80	**-4.40***	**-5.21***	1.38	**2.80***	**2.88***	**3.02***
Intervention ≥ 3rd year	-5.20	-2.96	-5.03	-2.51	**4.41***	2.93	2.08	-0.42

Note. $Intercept_{control}$ = mean in 2014-2015 in the control group, $Intercept_{int}$ = how the mean in 2014-2015 in the intervention group differed from the one in the control group, $Intercept_{new\ test\ version}$ = how the mean in 2014-2015 with the new test version differed from the one with the old test version, $Time\ trend_{control}$ = time trend in the control group, $Time\ trend\ pre_{int}$ = how the pre-intervention time trend in the intervention group differed from the time trend in the control group, $Intervention\ start_{int}$ = the change after implementing *Build!* for one year, $Time\ trend\ post_{int}$ = how the time trend in the intervention group changed after implementing *Build! for two years*, *Intervention 2nd year* = the change after implementing *Build! for* two years, *Intervention ≥ 3rd year* = the change after implementing *Build!* for three or more years.
$^{**}p < .01$. $^{*}p < .05$.

Table S4.4
Difference-in-Difference Models for Mathematics: Test Versions Together

	Percentage of Children With Difficulties					Mean Ability				
	$G1_{mid}$	$G1_{end}$	$G2_{mid}$	$G2_{end}$	$G3_{mid}$	$G1_{mid}$	$G1_{end}$	$G2_{mid}$	$G2_{end}$	$G3_{mid}$
Model 1										
$Intercept_{control}$ (β_0)	**30.27****	**32.13****	**31.01****	**31.52****	**29.44****	**29.87****	**40.00****	**49.26****	**60.83****	**70.45****
$Intercept_{int}$ (β_2)	-4.67	**-8.99****	**-6.58****	**-9.89****	**-8.05****	2.07	**5.25****	**3.99****	**5.06****	**3.61****
$Intercept_{new\ test\ version}$ (β_1)	-0.47	**-3.83***	0.14	-1.73	0.99	**84.63****	**98.15****	**110.3****	**119.33****	**129.57****
Time $trend_{control}$ (β_3)	0.77	-0.04	0.25	-0.38	0.16	**-0.89***	-0.11	-0.07	0.31	-0.18
Time trend pre_{int} (β_4)	-0.78	0.58	-0.61	0.21	-0.29	**1.28***	0.00	0.59	0.34	**0.85***
Intervention $start_{int}$ (β_5)	0.02	0.30	1.55	0.85	2.24	-0.05	0.35	-1.09	-0.53	-0.81
Time trend $post_{int}$ (β_6)	-0.08	-1.5	-0.08	0.08	-1.50	-0.04	0.96	0.14	0.97	0.82
Model 2										
Intervention $start_{int}$ (β_5)	-0.21	0.09	1.90	0.43	3.02	0.52	1.15	-1.70	-0.70	-0.59
Intervention 2nd year	0.68	-0.22	-1.07	1.54	-3.20	-1.68	-1.41	0.76	1.46	0.34
Intervention ≥ 3rd year	0.64	-3.77	-0.97	1.89	-3.75	-0.32	2.38	-0.02	0.89	1.25

Note. $Intercept_{control}$ = mean in 2014-2015 in the control group, $Intercept_{int}$ = how the mean in 2014-2015 in the intervention group differed from the one in the control group, $Intercept_{new\ test\ version}$ = how the mean in 2014-2015 with the new test version differed from the one with the old test version, $Time\ trend_{control}$ = time trend in the control group, *Time trend* pre_{int} = how the pre-intervention time trend in the intervention group differed from the time trend in the control group, *Intervention* $start_{int}$ = the change after implementing *Build!* for one year, *Time trend* $post_{int}$ = how the time trend in the intervention group changed after implementing *Build! for two years*, *Intervention* 2^{nd} *year* = the change after implementing *Build! for* two years, *Intervention* ≥ 3^{rd} *year* = the change after implementing *Build!* for three or more years.
$^{**}p < .01$. $^{*}p < .05$.

Table S4.5
Difference-in-Difference Models for Reading Fluency: Old and New Test Version

	Percentage of Children With Difficulties					Mean Ability				
	$G1_{mid}$	$G1_{end}$	$G2_{mid}$	$G2_{end}$	$G3_{mid}$	$G1_{mid}$	$G1_{end}$	$G2_{mid}$	$G2_{end}$	$G3_{mid}$
Old Test Version										
$Intercept_{control}$ (β_0)	**26.08***	**24.37****	**21.88****	**25.06****	**25.30****	**22.91****	**36.23****	**54.58****	**61.98****	**70.40****
$Intercept_{int}$ (β_2)	-3.29	-1.47	-0.73	**-4.11***	-2.21	1.22	1.16	1.41	**2.20***	1.76
Time $trend_{control}$ (β_3)	0.21	0.84	**1.14****	0.54	0.86	0.04	-0.30	**-0.43***	-0.20	-0.29
Time trend pre_{int} (β_4)	-0.03	0.61	-0.12	1.55	1.12	-0.29	-0.32	-0.20	**-1.07***	-1.29
Intervention $start_{int}$ (β_5)	2.41	0.78	-0.48	-0.41	0.67	-0.34	-0.27	0.45	0.92	2.29
Time trend $post_{int}$ (β_6)	-2.05	**-2.51***	**-2.44***	**-2.72***	-4.74	**0.90***	**1.68****	**1.84****	**1.94****	2.04
New Test Version										
$Intercept_{control}$ (β_0)	**30.46****	**23.78****	**29.14****	**20.39****	**20.25****	**16.94****	**28.64****	**44.39****	**53.77****	**63.94****
$Intercept_{int}$ (β_2)	-0.22	-0.53	-3.92	-0.70	3.54	-0.55	-0.52	0.70	3.30	-0.78
Time $trend_{control}$ (β_3)	3.51*	-2.57	0.07	0.86	1.88	**-1.16****	0.49	-0.37	-0.62	-0.46
Time trend pre_{int} (β_4)	-1.08	0.96	1.07	-4.54	-1.38	0.27	-0.23	0.09	2.55	0.12
Intervention $start_{int}$ (β_5)	0.79	-2.30	1.94	4.24	-1.82	0.29	0.22	-1.59	**-6.17****	0.45
Time trend $post_{int}$ (β_6)	**-2.07****	-0.75	**-1.36***	-1.46	-0.51	**0.39***	**0.83***	**0.99****	**1.54****	1.14

Note. $Intercept_{control}$ = mean in 2014-2015 in the control group, $Intercept_{int}$ = how the mean in 2014-2015 in the intervention group differed from the one in the control group, *Time* $trend_{control}$ = time trend in the control group, *Time trend* pre_{int} = how the pre-intervention time trend in the intervention group differed from the time trend in the control group, *Intervention* $start_{int}$ = the change after implementing *Build!* for one year, *Time trend* $post_{int}$ = how the time trend in the intervention group changed after implementing *Build! for two years.*
$^{**}p < .01$. $^{*}p < .05$.

Table S4.6
Difference-in-Difference Models for Spelling, Reading Comprehension, and Mathematics: New Test Version

	Percentage of Children With Difficulties					Mean Ability				
	$G1_{mid}$	$G1_{end}$	$G2_{mid}$	$G2_{end}$	$G3_{mid}$	$G1_{mid}$	$G1_{end}$	$G2_{mid}$	$G2_{end}$	$G3_{mid}$
Spelling										
$Intercept_{control}$ (β_0)	**26.89****	**28.24****	**30.65****	**27.21****	**24.98****	**151.52****	**199.34****	**230.98****	**261.30****	**292.56****
$Intercept_{int}$ (β_2)	-2.18	-4.81	**-8.66****	**-7.46****	-3.12	6.86	**8.08***	**12.71****	**13.46****	**10.03****
Time $trend_{control}$ (β_3)	1.04	-0.49	-0.74	-0.60	**1.26***	-1.32	0.33	**1.30***	**2.09****	-0.40
Time trend pre_{int} (β_4)	-0.32	0.41	1.25	1.38	-0.19	-1.03	-1.12	**-1.94***	**-3.02***	-1.13
Intervention $start_{int}$ (β_5)	-0.57	0.09	-1.45	-0.97	-0.23	2.27	1.90	0.61	2.51	0.15
Time trend $post_{int}$ (β_6)	**-1.65****	**-2.06****	**-1.50***	-1.61	**-2.27***	**3.30****	**2.95****	**2.84****	**3.11***	**3.19****
Reading Comprehension										
$Intercept_{control}$ (β_0)	-	**26.20****	**25.88****	**24.96***	**28.91****	-	**114.84****	**133.95****	**139.25****	**151.57****
$Intercept_{int}$ (β_2)	-	-4.65	-5.31	**-9.81****	**-6.14***	-	**4.99***	**6.06****	**7.74****	**4.96****
Time $trend_{control}$ (β_3)	-	1.25	**1.76****	1.06	0.89	-	-0.48	**-1.15****	-0.51	-0.23
Time trend pre_{int} (β_4)	-	0.86	0.07	0.63	0.03	-	-0.52	-0.55	-0.16	-0.37
Intervention $start_{int}$ (β_5)	-	-2.74	0.62	1.25	3.45	-	0.93	0.29	**-3.25***	-1.29
Time trend $post_{int}$ (β_6)	-	**-3.15****	-1.58	**-2.38***	**-2.67***	-	**2.23****	1.30	**1.72***	1.33

Note. $Intercept_{control}$ = mean in 2014-2015 in the control group, $Intercept_{int}$ = how the mean in 2014-2015 in the intervention group differed from the one in the control group, $Time\ trend_{control}$ = time trend in the control group, $Time\ trend\ pre_{int}$ = how the pre-intervention time trend in the intervention group differed from the time trend in the control group, $Intervention\ start_{int}$ = the change after implementing *Build!* for one year, $Time\ trend\ post_{int}$ = how the time trend in the intervention group changed after implementing *Build! for two years.*
$^{**}p < .01$. $^{*}p < .05$.

Table S4.6 (continued)
Difference-in-Difference Models for Spelling, Reading Comprehension, and Mathematics: New Test Version

	Percentage of Children With Difficulties					Mean Ability				
	$G1_{mid}$	$G1_{end}$	$G2_{mid}$	$G2_{end}$	$G3_{mid}$	$G1_{mid}$	$G1_{end}$	$G2_{mid}$	$G2_{end}$	$G3_{mid}$
Mathematics										
$Intercept_{control}$ (β_0)	**30.90****	**27.82****	**31.55****	**30.02****	**30.77****	**113.05****	**137.67****	**158.33****	**178.56****	**199.28****
$Intercept_{int}$ (β_2)	**-6.52***	**-8.94****	**-7.71****	**-10.87****	**-10.4****	4.43	**6.27****	**6.61****	**8.32****	**7.13****
Time $trend_{control}$ (β_3)	0.58	0.08	0.07	-0.70	0.09	-0.64	0.09	0.34	**0.92****	-0.07
Time trend pre_{int} (β_4)	-0.08	0.56	-0.36	0.68	0.18	0.82	-0.20	-0.16	-0.5	0.06
Intervention $start_{int}$ (β_5)	-0.54	0.84	1.61	0.63	2.63	-0.36	-0.19	-1.22	-0.98	-1.36
Time trend $post_{int}$ (β_6)	-0.57	**-1.66***	-0.06	0.18	-1.67	0.06	0.84	0.17	0.40	1.06

Note. $Intercept_{control}$ = mean in 2014-2015 in the control group, $Intercept_{int}$ = how the mean in 2014-2015 in the intervention group differed from the one in the control group, *Time* $trend_{control}$ = time trend in the control group, *Time trend* pre_{int} = how the pre-intervention time trend in the intervention group differed from the time trend in the control group, *Intervention* $start_{int}$ = the change after implementing *Build!* for one year, *Time trend* $post_{int}$ = how the time trend in the intervention group changed after implementing *Build! for two years.*
$^{**}p < .01$. $^{*}p < .05$.

Table S4.7
Chi-Square Difference Tests for Models Distinguishing Early and Late Adopting Schools

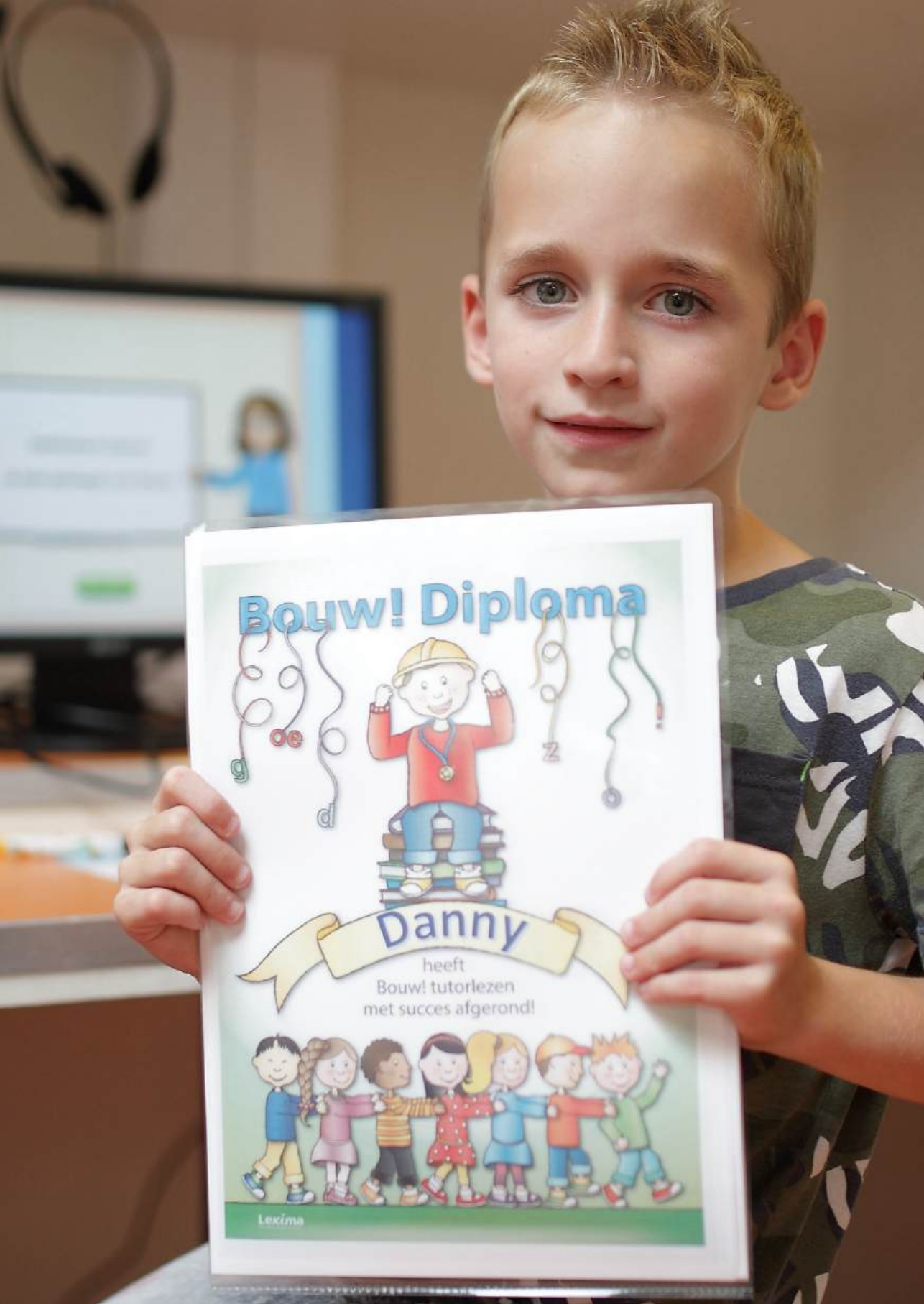
	Percentage of Children With Difficulties					Mean Ability				
	$G1_{mid}$	$G1_{end}$	$G2_{mid}$	$G2_{end}$	$G3_{mid}$	$G1_{mid}$	$G1_{end}$	$G2_{mid}$	$G2_{end}$	$G3_{mid}$
Model 1 vs. Model 2										
Reading Fluency										
$\Delta\chi^2$ (*df*)	2.69 (1)	1.97 (1)	0.39 (1)	7.16 (1)	2.89 (1)	0.04 (1)	0.76 (1)	1.89 (1)	6.67 (1)	1.65 (1)
p	.101	.160	.534	**.008**	.089	.835	.384	.169	**.010**	.198
Spelling										
$\Delta\chi^2$ (*df*)	0.83 (1)	0.06 (1)	0.49 (1)	3.56 (1)	0.16 (1)	4.20 (1)	7.47 (1)	2.79 (1)	3.90 (1)	0.51 (1)
p	.363	.803	.485	.059	.686	**.040**	**.006**	.095	**.048**	.477
Reading Comprehension										
$\Delta\chi^2$ (*df*)	-	0.17 (1)	0.82 (1)	0.76 (1)	2.41 (1)	-	1.40 (1)	2.20 (1)	0.26 (1)	6.84 (1)
p	-	.682	.365	.383	.121	-	.237	.138	.607	**.009**
Mathematics										
$\Delta\chi^2$ (*df*)	0.16 (1)	0.63 (1)	0.39 (1)	0.67 (1)	0.32 (1)	1.74 (1)	4.44 (1)	0.35 (1)	0.76 (1)	0.70 (1)
p	.686	.428	.531	.412	.573	.187	**.035**	.551	.383	.401

Note. Model 1 is the simplest model with a linear post-intervention time trend. In Model 2 the linear post-intervention time trend has been replaced by two dummies: the post-intervention effect after two years and after three or more years. Model 2 with interactions distinguished early and late adopting schools (i.e. schools using the intervention for less than three years or for three years or more) and includes two interaction terms: the interaction of adopter group with the immediate effect (after one year) and the interaction of adopter group with the effect after two years.

Table S4.7 (continued)
Chi-Square Difference Tests for Models Distinguishing Early and Late Adopting Schools

	Percentage of Children With Difficulties					Mean Ability				
	$G1_{mid}$	$G1_{end}$	$G2_{mid}$	$G2_{end}$	$G3_{mid}$	$G1_{mid}$	$G1_{end}$	$G2_{mid}$	$G2_{end}$	$G3_{mid}$
Model 2 With vs. Without Interactions										
Reading Fluency										
$\Delta\chi^2$ (*df*)	1.43 (2)	0.83(2)	1.96 (2)	3.85 (2)	2.97 (2)	1.05 (2)	0.21 (2)	6.68 (2)	7.78 (2)	2.88 (2)
p	.490	.661	.376	.146	.226	.592	.899	**.036**	**.021**	.237
Spelling										
$\Delta\chi^2$ (*df*)	4.27 (2)	0.71 (2)	2.27 (2)	0.89 (2)	3.94 (2)	4.27 (2)	3.65 (2)	3.94 (2)	0.25 (2)	3.87 (2)
p	.118	.701	.321	.639	.140	.118	.161	.140	.883	.145
Reading Comprehension										
$\Delta\chi^2$ (*df*)	-	0.79 (2)	0.63 (2)	1.06 (2)	7.26 (2)	-	1.71 (2)	0.85 (2)	3.36 (2)	14.16 (2)
p	-	.672	.729	.590	**.027**	-	.426	.653	.186	**.001**
Mathematics										
$\Delta\chi^2$ (*df*)	4.03 (2)	2.39 (2)	1.91 (2)	0.22 (2)	0.10 (2)	3.69 (2)	3.68 (2)	5.23 (2)	2.61 (2)	5.47 (2)
p	.134	.302	.385	.896	.949	.158	.159	.073	.271	.065

Note. Model 1 is the simplest model with a linear post-intervention time trend. In Model 2 the linear post-intervention time trend has been replaced by two dummies: the post-intervention effect after two years and after three or more years. Model 2 with interactions distinguished early and late adopting schools (i.e. schools using the intervention for less than three years or for three years or more) and includes two interaction terms: the interaction of adopter group with the immediate effect (after one year) and the interaction of adopter group with the effect after two years.


Bouw! Diploma
oe
g
d
z
o
Danny
heeft
Bouw! tutorlezen
met succes afgerond!
Lexima

CHAPTER 5

General Discussion

This dissertation was focused on topics related to the large-scale implemenation of the early-literacy intervention *Build!*. Three topics were adressed. First, relations between natural variations in treatment integrity and intervention outcomes were examined. Second, it was investigated whether two family characteristics were related to intervention outcomes and treatment integrity, i.e. familial risk for dyslexia and parental education. Third, the effects of *Build!* at the school level were determined and it was investigated whether the intervention became more effective when schools used it for a longer time. Each of these topics is addressed in a separate section. The final section of this General Discussion examines the reach of the intervention, that is the extent to which children at risk for reading problems—the target group for which the intervention is intended—participated in the intervention. This topic is addressed in this General Discussion, as findings in Chapters 2, 3, and 4 suggest that schools may have encountered difficulties in selecting the target group, which can influence the effectiveness of the intervention in schools (Glasgow et al., 2006). I also provide future directions and practical implications.

1.1 *Relation Between Treatment Integrity and Intervention Outcomes*

It is generally accepted that treatment integrity is key to reaching the full potential of an intervention and to evaluate intervention effectiveness (Durlak & DuPre, 2008; O'Donnell, 2008). Nevertheless, treatment integrity is reported in only about half of the studies on the effects of (reading) interventions and there are only a few studies that examined the relation between treatment integrity and intervention effectiveness (Capin et al., 2018; Swanson et al., 2013). In this dissertation, the focus was on one particular dimension of treatment integrity: dosage, which refers to the amount of intervention practice (Dane & Schneider, 1998). Studies on the relation between dosage and intervention outcomes have inconsistent findings, showing either no, a positive, or a negative relation (for a review, see van Dijk et al., 2023). We aimed to provide a better understanding of the relation between dosage and intervention outcomes by distinguishing dosage (time spent on the intervention) from progress within the intervention (the number of intervention lessons completed). Moreover, we distinguished multiple aspects of dosage. The relation between dosage and intervention outcomes was examined within the context of a large-scale implementation

of *Build!*, an intervention for the prevention of reading problems. The intervention was implemented by schools, thus a natural setting.

1.1.1 *Effects of Dosage*

In Chapters 2 and 3, the relation of dosage (i.e. the number of hours spent on the intervention) with progress within the intervention (i.e. the number of new intervention lessons finished) and with literacy outcomes was examined during three intervention periods: (1) kindergarten, (2) the first half of Grade 1, and (3) the second half of Grade 1 (Chapter 3). Findings in Chapter 2 and 3 indicate that, during kindergarten, children who practiced more made more progress, which means that they finished more new intervention lessons. Thereby, they reached higher levels of letter knowledge and phonological awareness at the end of kindergarten, as well as higher levels of reading accuracy at the beginning of Grade 1. Chapter 3 reveals that children who practiced more during the first half of Grade 1 made more progress within the intervention and, in turn, read more fluently at the middle of Grade 1. In the second half of Grade 1, a higher dosage was not associated with more progress nor with higher literacy outcomes.

For the interpretation of these relations, it is important to note that prior levels of (pre-)literacy skills were controlled. Thus, dosage was still associated with higher literacy outcomes after children's prior levels of (pre-)literacy were taken into account. Following the logic of longitudinal research, such an additional effect is often taken to support a causal effect (Gollob & Reichardt, 1987). In this case, the effect of dosage on the growth of literacy skills could thus support an effect of the intervention on literacy outcomes. Although probably not decisive, the effects of dosage on top of prior literacy skills thus provide some support for a causal effect of dosage on literacy outcomes.

It seems obvious that more practice with an intervention leads to higher literacy outcomes, but few studies have shown this relationship within the same intervention (Al Otaiba et al., 2005; Wanzek & Vaughn, 2008; Wolgemuth et al., 2014; Zijlstra et al, 2014). Actually, several meta-analyses demonstrated that there was no relation between dosage and intervention outcomes (Tran et al., 2011; Wanzek et al., 2013; Wanzek & Vaughn, 2007). However, it should be noted that in these meta-analyses, conclusions about dosage are based on the comparison of *a variety of interventions*. Interventions did thus not only differ in dosage, but also in other ways. As such, the finding that interventions with varying doses did not differ in effect size may be explained by other differences among interventions, which could have competed with

the effect of dosage on intervention outcomes. So, the relation between dosage and intervention outcomes can better be studied among children following *the same intervention*, receiving larger and smaller doses of this intervention. Results of studies that did examine the effects of dosage for children within the same intervention mostly align with those of the studies described in the previous chapters (Chapters 2 and 3): the more time is spent on the literacy intervention, the higher were the literacy outcomes (Al Otaiba et al., 2005; Wolgemuth et al., 2014; Zijlstra et al., 2014). This implies that when schools implement an evidence-based (early-literacy) intervention, effects may be dependent on the amount of practice children receive. Schools thus need to monitor and stimulate practice with the intervention to reach the full potential of the intervention.

1.1.2 *Effects of Different Aspects of Dosage*

In Chapter 2, it was examined which aspects of dosage had the strongest relation with progress within the intervention as well as intervention outcomes in kindergarten (Chapter 2). Three aspects of dosage were distinguished: the number of sessions per week (frequency), the session length (length), and the number of intervention weeks (duration). Analyses were conducted at the week level and the child level. At the week level, it was examined whether children made more progress (i.e. finished more new intervention lessons) in weeks in which they had either more or longer intervention sessions. Results indicated that frequency had a stronger impact on progress through the intervention than length. This indicates that children finished more new intervention lessons in weeks in which they practiced frequently and short than in weeks in which they practiced infrequently and long. At the child level, it was examined whether children who on average practiced more frequently, longer, or for more weeks made more progress within the intervention and in turn made larger improvements in letter knowledge and phonological awareness during kindergarten. Again, frequency had the largest impact. It was followed by the number of intervention weeks and finally the session length. Also this finding is in line with the theory of distributed practice indicating that spaced practice is preferred over massed practice (Dunlosky et al., 2013). Because analyses at the week level were conducted *within* children, whereby children served as their own control, the fact that frequency was most important at both the week and the child level, provides some support for a causal interpretation of the relation between dosage and intervention outcomes.

The three aspects of dosage were only distinguished in investigating relations between dosage and intervention outcomes in kindergarten, not in Grade 1.

Therefore, it is not clear whether the findings on the most important aspects of dosage are specific to kindergarten or pertain to Grade 1 as well. A difference between grades could be expected, as in kindergarten formal reading instruction has not yet begun, while it has in Grade 1. During formal reading instruction, actual reading problems start to arise and the literacy intervention is provided in addition to the classroom reading instruction. The response-to-intervention model suggests that children who have difficulties with learning to read benefit from intervention sessions that are not only more frequent but also longer than regular classroom practice (Bursuck & Blanks, 2010; Fuchs & Fuchs, 2006). As a result, the impact of the length of intervention sessions might be larger in Grade 1. Moreover, older children tend to have a longer attention span (Tremolada et al., 2019), meaning that longer intervention sessions may be more helpful for children in Grade 1 than in kindergarten.

These hypotheses were tested by running additional analyses; the analyses on the most important aspects of dosage in kindergarten (Chapter 2) were repeated for the intervention period in Grade 1. As some children stopped the intervention before the end of Grade 1, I took the period from the end of August until the beginning of May, including 226 children.

Similar to the results in kindergarten, 80% of the variance in progress in Grade 1 was located at the week level. In other words, progress, the number of new lessons per week, varied strongly across weeks. Table 5.1 shows, for both kindergarten and first grade, how many new intervention lessons were finished in weeks with one intervention session of 0-10 minutes per week (intercept) compared to weeks with two, three, four, or five intervention sessions of 0-10 minutes. For example, in Grade 1, in a week with three intervention sessions of 0-10 minutes, on average 5.1 (1.034 + 1.992 + 2.027) new intervention lessons were completed. Findings indicate that each extra session was associated with more progress within the intervention. But the fifth session (in which 1.6 lessons were completed) was less efficient than the second, third, and fourth session (in which around 2.0 lessons were completed).

Table 5.1 also shows, for both kindergarten and first grade, how many new intervention lessons were finished in weeks with one intervention session that lasted on average 0-10 minutes (intercept) compared to weeks with one session of on average 10-15 minutes, 15-20 minutes, or more than 20 minutes. In Grade 1, in a week with one intervention session of 15-20 minutes 2.6 (1.034 + 0.967 + 0.623) new intervention lessons were completed. Findings indicate that each additional five minutes of practice per session was associated with more progress within the intervention. However, children made more progress in the first ten minutes than in the next five minutes and, and even less progress was observed in the five minutes that followed.

Thus, the longer the session, the less efficient it was. As a result, two sessions of ten minutes (in which first-graders finished 3.0 new intervention lessons) was more efficient than one session of 15-20 minutes (in which first-graders finished 2.6 lessons), in both kindergarten and Grade 1.

Table 5.1
Regression coefficients of the Factors Predicting Progress Within the Intervention in Kindergarten and Grade 1

Predictor	Kindergarten	Grade 1
	Est.	Est.
Intercept[a]	0.907***	1.034***
Frequency: 2 times a week	1.557***	1.992***
Frequency: 3 times a week	1.719***	2.027***
Frequency: 4 times a week	1.379***	1.998***
Frequency: 5 times a week	1.274***	1.643***
Length: 10-15 minutes	0.666***	0.967***
Length: 15-20 minutes	0.520***	0.623***
Length: > 20 minutes	0.381***	0.561***
Proportion from home: partly	-0.137	0.057
Proportion from home: mostly	0.035	0.220**
Review lessons: ≥ 1 lesson	-0.933***	-1.177***
Review lessons: ≥ 2 lessons	-0.963***	-1.068***
Review lessons: ≥ 3 lessons	-1.002***	-1.327***
R^2	.774	.739

Note. Unstandardized regression coefficients are displayed. R^2 is the explained level 1 variance.
[a]The intercept parameter estimate represents the progress within the intervention for practicing once a week, 0-10 minutes, hardly at home, and without review lessons.
$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

Next to the week level, the child level was regarded as well. The means and standard deviations of the variables with respect to the various aspects of dosage in kindergarten and first grade are presented in Table 5.2. The frequency and length of intervention sessions were highly similar in kindergarten and Grade 1. Duration (the number of intervention weeks) was larger in Grade 1 than in kindergarten, as well as progress (the number of new intervention lessons completed). It was investigated

which aspects of dosage had the strongest relation with progress within the intervention at the child level. That is, do children who practice more, finish more new intervention lessons? Figure 5.1 shows that, also in Grade 1, all three aspects of dosage were related to progress within the intervention. Frequency had the strongest relation with progress. The strength of this relation was similar in kindergarten and Grade 1. More or less the same holds for length and progress. However, duration (the number of intervention weeks) had a weaker relation with progress in Grade 1 than in kindergarten. As a result, next to frequency, the length of intervention sessions was most important in Grade 1, while this was duration in kindergarten. A possible explanation for this finding could be that children who do not practice every week in Grade 1, may gain letter knowledge and reading skills by regular instruction in the classroom and thereby finish as many new intervention lessons within a session as children who practice every week and mainly gain letter knowledge and reading skills in *Build!*.

Table 5.2

Descriptive Statistics of Dosage, Review Lessons, Proportion From Home, and Progress: Kindergarten vs. Grade 1

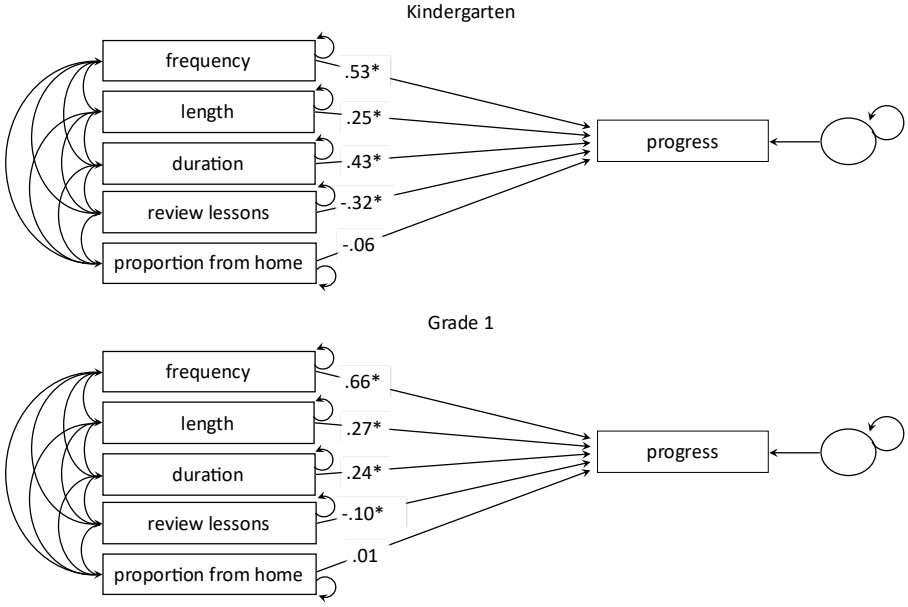
	Kindergarten		Grade 1	
	M	*SD*	*M*	*SD*
Frequency	2.17	0.57	2.41	0.60
Length	13.63	3.00	12.80	2.62
Duration	14.22	2.52	28.70	3.54
Review lessons	0.46	0.50	0.36	0.48
Proportion from home	0.65	0.48	0.81	0.40
Progress	41.59	20.35	127.73	59.25

In conclusion, the most important aspect of dosage was frequency, both in kindergarten and Grade 1. The suggestion that schools could best raise the number of intervention sessions per week rather than the length of each session to increase progress within the intervention thus holds for both grades. In kindergarten, it also seems helpful to plan as many intervention weeks as possible. In Grade 1, this effect was more difficult to determine given the smaller variation in number of intervention weeks. Findings are in line with the effective learning strategy of distributed practice, which indicates that spaced practice (short and frequent practice) is preferred over massed practice (long and rare practice; Dunlosky et al., 2013). Findings are also in

line with learning strategy of retrieval practice, indicating that learning is enhanced when children recall facts from memory on a regular basis (Carpenter et al., 2022). To our knowledge, our study is the first to show these learning principles are also applicable to (early-)literacy interventions.

Figure 5.1

Relations Between Dosage, Progress, and Intervention Outcomes: Kindergarten and Grade 1

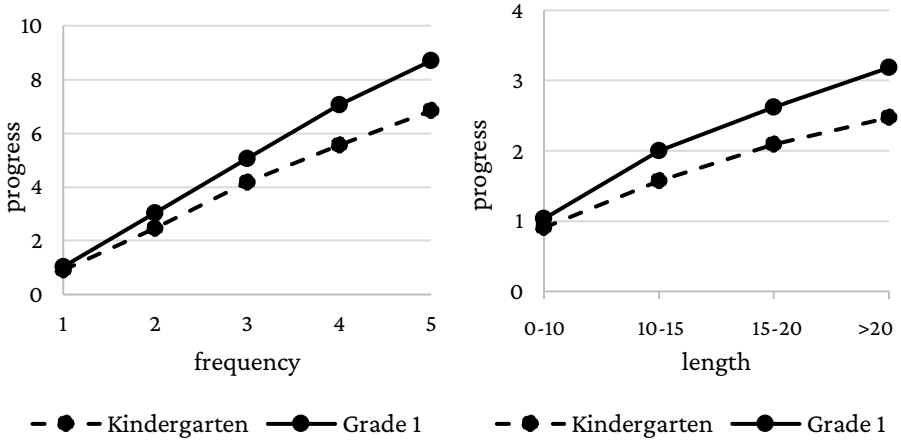


1.1.3 *Dose-Response Relationships*

In Chapter 2, dose-response relationships were investigated. While literacy interventions often have a prescribed amount of practice, the optimal dose is rarely investigated (for some exceptions see Al Otaiba et al., 2005; Wanzek & Vaughn, 2008). Previous research has shown that the highest dose is not always the best (Wanzek & Vaughn, 2008). We studied dose-response relationships for *Build!* in kindergarten (Chapter 2). Based on Table 5.1 in the previous section, it is also possible to determine the form of the dose-response relationships for Grade 1. Dose-response relationships were not linear (see Figure 5.2). In both kindergarten and Grade 1, the longer the sessions, the less efficient they were in terms of progress within the intervention. Thus, doubling the *length* of a session did not double the number of completed new intervention lessons. But doubling the *number* of intervention sessions per week did

double the number of completed new intervention lessons, up to three sessions per week in kindergarten and four sessions per week in Grade 1. To our knowledge, dose-response relationships have not been studied in literacy interventions in such detail before. However, clearly, they provide useful information for schools that implement the intervention. In this case, practicing three times a week for ten to fifteen minutes was most efficient in kindergarten and four times a week in Grade 1, but if children needed a higher dose it was most efficient to increase the number of sessions per week rather than the session length. practicing more was rewarding, especially having more intervention sessions per week.

Figure 5.2
Dose-Response Relationships in Kindergarten and Grade 1



1.1.4 *Distinction Between Dosage and Progress*

An important finding in this dissertation was that progress within the intervention fully mediated the relationship between dosage and children's literacy outcomes (Chapters 2 and 3). The improved literacy outcomes observed in children who practiced more frequently during the intervention could be attributed to their completion of a larger number of new intervention lessons. This may seem evident, but to our knowledge, this mediation has not been demonstrated in previous studies. In previous research, dosage and progress are often taken together (van Uittert et al., 2022; Zijlstra et al., 2014) or progress is not taken into account (Al Otaiba et al., 2005; Wolgemuth et al., 2013).

Findings in this dissertation reveals that dosage and progress, although strongly related, are not the same (Chapters 2 and 3). Children who spent equal amounts of time on the intervention did not always made equal amounts of progress through the program. As expected, progress was correlated with the number of review lessons. Children who required more review lessons advanced more slowly through the program, despite engaging in similar amounts of practice (Chapter 2). Moreover, children with lower pre-literacy skills at the beginning of the intervention progressed more slowly through the program, resulting in smaller gains in literacy outcomes (Chapter 3). A similar finding was reported by van Uittert et al. (2022) in their game-based intervention. Van Uittert et al. registered 'in game efficiency', defined as the average number of exercises completed correctly by a child per minute. This measure resembles our measure of progress. Van Uittert et al. found that children's literacy skills at the beginning of the intervention predicted children's 'in-game efficiency'. Similarly, in our study we found that initial preliteracy skills were related to progress but, importantly, were unrelated to dosage (Chapters 2 and 3). Thus, the results of our studies indicate that it is important to distinguish between dosage and progress. Both progress and dosage are related to intervention outcomes, but only progress is related to children's ability to learn to read.

An interesting approach is to use progress as an indicator of children's responsiveness to literacy interventions, when taking into account the time spent on the intervention. A study by Thomson et al. (2020) conducted a pioneer study in which progress data of a digital intervention, *Graphogame*, was used to predict future reading performance. They used growth curves of within-game skill mastery, i.e. the number of words read correctly within a particular play session, to predict reading performance at the end of Grade 1. They found that differences in growth predicted variation in literacy performance at the end of Grade 1, even after controlling for risk status at the beginning of Grade 1 (a combined measure of letter knowledge, rapid automatized naming, phonemic awareness, and familial risk for dyslexia). Thomson et al. suggested that children's progress through the intervention could serve as a dynamic assessment tool for the early detection of reading difficulties.

A measure of progress within the intervention could therefore be useful for literacy interventions like *Build!* to make a further distinction between children who will and will not develop reading problems. The early selection of children at risk for reading problems based on the precursors of reading is always susceptible to both over- and under-identification (Fletcher et al., 2021). Future research is needed to investigate whether children's progress within the intervention can indicate which children

need even more intensive and specialized instruction, in line with the response-to-instruction model (Fuchs & Fuchs, 2006).

1.2 *Family Characteristics Related to Treatment Integrity and Intervention Outcomes*

The second aim of the current dissertation was to investigate whether two family characteristics, family risk for dyslexia and parental education, affect treatment integrity and intervention outcomes. Previous research provide little information on the relation between family characteristics and children's responsiveness to interventions, because child and family characteristics (e.g., ethnicity, primary language, socioeconomic status) are not always reported and their relation with intervention outcomes is mostly not examined (Bautista et al., 2024; Manz et al., 2010). Research on the relation between family characteristics and individual differences in intervention effectiveness can provide insight into the effects of (early-literacy) interventions for specific subgroups. Some studies on children with a familial risk of dyslexia indicate that literacy interventions can be less effective for them and that they require more intervention sessions compared to children not at risk (Elbro & Petersen, 2004; Hindson et al., 2005). As these studies were conducted on a small scale, it is important to understand whether such additional practice is provided to children with familial risk in natural school settings and how familial risk affects progress within the intervention and intervention outcomes.

Findings in this dissertation reveal that familial risk for dyslexia was unrelated to dosage (Chapter 3), indicating that children with familial risk for dyslexia were not provided with more practice than children without familial risk. Familial risk for dyslexia did show a relation with literacy outcomes, i.e. familial risk had a negative effect on children's letter knowledge at the end of kindergarten. Interestingly, after controlling for this effect, familial risk also had a negative effect on children's reading skills at the middle of Grade 1. Children with familial risk for dyslexia thus experienced difficulties with pre-literacy skills and with aspects of reading that were not predicted by those skills. The latter finding is in line with the results of previous research (e.g. Byrne et al., 2009; van Viersen et al., 2018). Based on genetic research Byrne et al. (2009) suggested that this could indicate that genes are involved in the development of word reading that are not related to the development of preliteracy skills. Findings in this dissertation reveal another reason why children with a familial risk attained lower literacy levels after the intervention: they progressed more slowly through the program, despite spending as much time on the intervention as their

peers without a familial risk. In line with previous studies (Hindson et al., 2005; Zijlstra et al., 2021), findings indicate that children with familial risk need to spend more time practicing with the intervention to progress as quickly through the program as children without such a risk and to prevent them from reading difficulties due to this slower progress. Because the schools in the current study did not provide children at risk with this additional practice, they might not have been aware that part of the reading difficulties could have been prevented by increasing the dosage of children at risk so that they complete as many intervention lessons as children without such risk. Thus, in addition to monitoring the number of intervention sessions per week, schools should also focus on children's progress within the intervention. Schools can use information on progress to adjust the number of intervention sessions for children whose progress lags behind.

Parental education, the other family characteristic that was studied, was unrelated to either dosage, progress, or intervention outcomes (Chapter 3). Children whose parents had a higher or lower educational level spent similar amounts of time on the intervention, completed the same number of new intervention lessons, and achieved similar literacy outcomes. In other words, the intervention was equally effective for children whose parents have a higher or lower educational level. This finding contradicts those of the last RCT on *Build!* (Zijlstra et al., 2021). In that study, the intervention was ineffective for a subgroup of children, which was suggested to consist of, among others, children whose parents have a lower educational level. The finding is also not in line with studies showing that parental education affects children's literacy development (Krijnen et al., 2020; Leseman & Jong, 1998; Segers et al., 2016) or the effectivity of the intervention (Manz et al., 2010). Perhaps the effects of parental education in the current studies were underestimated, due to the selected sample of children at risk for reading problems (restriction of range). In our sample, there was some indication that restriction of range was the case as the sample contained a relatively small group of parents with a higher educational level compared to the national sample (29% vs. 48%; CBS, 2023). This could be a result of selecting children at risk of reading problems for the intervention, because parental education is positively related to children's literacy skills (Krijnen et al., 2020; Leseman & Jong, 1998; Segers et al., 2016).

1.3 *Effects of Build! at the School Level*

The third aim of this dissertation was to examine whether the large-scale implementation of *Build!* led to a reduction of literacy problems and an increase in the average

literacy skills in schools. Effects of literacy interventions are rarely investigated when implemented by schools on a large scale (for some exceptions see Denton et al., 2010; Stein et al., 2008; Torgesen, 2009). This is unfortunate. Children and schools want to know whether their work pays off, and policy makers need to determine whether the costs of these interventions are justified.

Generally, the effects of interventions implemented on a large scale will be smaller than in small-scaled RCTs. Populations are more heterogeneous and treatment integrity may vary more strongly in natural school settings than in researcher-led RCTs (Lortie-Forgues & Inglis, 2019; Thomas et al., 2018). Especially among schools that have only recently started to implement the intervention, treatment integrity might not yet be optimal. It has been suggested that interventions become more effective when schools implement them for a longer time (Harn et al., 2013; Torgesen, 2009), although, to our knowledge, this has not yet been examined.

Therefore, a quasi-experimental study was conducted to examine whether the large-scale implementation of *Build!* is associated with a reduction of the percentage of children with literacy problems and/or an increase in the average literacy skills at the school level (Chapter 4). In particular, it was investigated whether these effects become larger when schools have implemented *Build!* for a longer period of time. Reading fluency, spelling, and reading comprehension were measured, and compared to mathematics to which transfer of the effect of the intervention was deemed unlikely. Effects of the intervention were determined during the intervention (mid- and end-Grade 1), at post-test (mid-Grade 2), and at follow-up (end-Grade 2 and mid-Grade 3, i.e. half a year and one year after the intervention was finished). Schools were followed during six successive school years, in which schools in the intervention group (72-145 schools, depending on the outcome measure) introduced the intervention, and were compared to schools that did not used the intervention (61-126 schools). Although this was a quasi-experimental study, it is important to note that pre-existing differences between schools that did and did not implement *Build!* were controlled for. Changes *within* schools were estimated, and thus findings are not due to pre-existing differences across schools.

Findings showed that there was no change in literacy skills after the first year that schools had been using *Build!*. However, after schools implemented the program for two years, there were small overall decreases in the percentage of children with difficulties in reading fluency, spelling, and reading comprehension, while this was not the case for math. Moreover, the average reading fluency and spelling ability of children in cohorts receiving *Build!* begun to increase. Changes were small (only a few percent or scale points), but only small effects could be expected on a large scale,

due to heterogeneous populations and a large variety in treatment integrity. For two other reasons, more specific to *Build!*, the change in the percentage of children with reading problems was also expected to be limited. First, the target group of *Build!* is relatively small (the children with the 25-30% lowest scores on preliteracy skills) and the selection of children is never perfect. Early selection always results in under- and overidentification of children with reading problems (Fletcher et al., 2021). Thus, not all the children with the 25% lowest reading scores might have received the intervention. Second, the program might not prevent all children from developing reading problems, especially not in the children with the most severe deficits (Snowling & Melby-Lervåg, 2016). Although effects are small at first glance, a drop in poor readers of about 3% in a population of over one million Dutch primary-school children, means that in thousands of children severe literacy problems have been prevented.

The results of this study (Chapter 4) did not reveal why the implementation of *Build!* had small effects at the school level. It seems unlikely that the improvement in the average literacy abilities in schools was only due to the small group of children that received *Build!*. 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. Instead, we found the opposite pattern for some literacy skills and equally large effects on these two outcomes for other literacy skills. Thus, it is likely that the implementation of *Build!* affected the literacy level of children who did not participate in the intervention as well. The implementation of *Build!* may have increased attention for literacy education in the school or the use of *Build!* might be part of a broader school policy aimed at improving literacy education.

Furthermore, an interesting finding is that *Build!* was found to be effective only after schools had been implementing the intervention for two or more years. Though no systematic data were gathered about why it takes time to reach intervention effects, in a number of interviews with schools that used the program for several years (de Jong et al., 2023), it was revealed that these schools had made some changes over the years of implementation, which made, in their opinion, the intervention more effective. These changes included (a) improving the quality of intervention sessions by giving training, supervision, and support to tutors who provided the children with the intervention sessions, (b) increasing the frequency of practice by improving the schedule and increasing communication with tutors at school and at home, and (c) solving logistic problems at school, such as an insufficient number of computers, a noisy room, and time needed to start the computer program. These findings are in line with the assertions by Harn et al. (2013) who suggested that adaptations to the

school context can make an intervention more effective, because it better meets the needs of school staff and children. Consequently, implementing an intervention and making it effective at a particular school requires time and persistence. This might be easily overlooked by many researchers and schools.

The finding that it took time before the intervention was effective, is also in line with the theory on the phases of implementation (Bradshaw et al., 2009). These phases are *preparation* (schools are preparing to implement the intervention), *initiation* (schools are beginning to implement the intervention), *implementation* (schools are actively implementing the core components of the program), and *maintenance* (schools are integrating the interventions core aspects in organizational routines). It is possible that an intervention might not yet have an observable effect in the initial phases of implementation, but only in the later phases.

1.4 *The Reach of the Intervention*

So far, the central topics of Chapters 2, 3, and 4 were discussed: treatment integrity, family characteristics, and intervention effectiveness. Another important aspect of translating interventions into practice is reach (Glasgow et al., 2006). *Reach* is the proportion of the target population that participates in the intervention (Glasgow et al., 1999). In the intervention *Build!,* the target population is the group of children that will develop reading problems later on. The challenge for schools is to accurately predict which children will develop reading problems and to select them for the intervention. In Chapters 2 and 3 a selection procedure was described that was used by part of the schools, and it was found that schools did not only select children who were eligible for the intervention, but also children who were not eligible. In Chapter 4, a study in which more schools participated, it was suggested that only a part of the target population was possibly reached by the schools, which might explain the small intervention effects. In the study described in Chapter 3, the intervention was stopped for part of the children, likely because schools did not see a further need to continue the intervention for these children. These findings illustrate three challenges that schools face in assessing whether children should get an early-literacy intervention and, if so, how long the intervention should be continued. In this section, I will elaborate on these challenges and examine a number of additional research questions.

1.4.1 *First Challenge: Early Screening*

The intervention *Build!* is a prevention program and, like other early-literacy interventions, meant to start in kindergarten. Therefore, schools could not select the target population directly, but only children *at risk* for reading problems. They had to predict which children would and would not develop reading problems based on the precursors of reading (e.g. letter knowledge and phonological awareness). Precursors of reading can only partly predict reading (Leppänen et al., 2008; Torppa et al., 2007). Thus inevitably, some children are identified as 'not at risk' while they will develop reading problems (false negatives) and some children are identified as 'at risk' while they will develop sufficient reading skills (false positives). In case of false negatives, children are not provided with the intervention while needing it and therefore may develop (lifelong) reading difficulties with potential negative economic and social consequences (Annie. E. Casey Foundation, 2010; Bear & Minke, 2002; Buisman et al., 2012; Luyten & Bruggencate, 2011; Poskiparta et al., 2003). In case of false positives, children receive unnecessary intervention, which is a waste of schools precious time at the cost of children who really need the intervention. The first challenge for schools is to develop a screening procedure that results in a minimum number of false positives and false negatives and that can be administered in the limited time and with the limited resources that schools have available. Such brief and teacher-friendly screening measures for early-literacy interventions are scarce (for an exception see Fletcher et al., 2021).

In the Participants sections of the studies presented in Chapters 2 and 3, a selection procedure was described that was used by part of the schools. This procedure was developed to be brief and teacher-friendly. It included two screening waves, one in October and one in January, in the second year of kindergarten. In both waves, letter knowledge and phonological awareness were assessed. Additional analyses are needed to evaluate to what extent this procedure led to false negatives, i.e. how many children who were not eligible for the intervention on the screening measures in January, did develop reading problems later on.

1.4.2 *Second Challenge: Inclusion and Exclusion*

A second challenge for schools is to decide which children are provided with the intervention, after having administered the screening. There might be more children identified as 'at risk' than the number of children that can be provided with the intervention within the time and with recourses that schools have. Then, schools have to select the children who are 'most at risk'. Schools might consider them to be

children with the lowest screening scores or children with additional risk factors, such as familial risk for dyslexia. However, it is also possible that schools could provide the intervention to more children, rather than only those identified as 'at risk'. Then schools might decide to select some additional children for the intervention, children who might be in need of the intervention too, such as children with familial risk for dyslexia or those who stand out as 'at risk' by teacher observations of early literacy. Indeed, as reported in Chapters 2 and 3, schools did provide some children with the intervention when they were not eligible according to the selection procedure. There were also children who were eligible, but did not receive the intervention. It is unclear what the consequences of these decisions were. Adhering strictly to the procedure is often considered best (Capin et al., 2018; Gresham et al., 2000; O'Donnell, 2008). The current procedure included only two predictors of reading. This raises the question whether the procedure might be improved by taking additional predictors into account.

1.4.3 *Third Challenge: Identifying False Positives*

A third challenge for schools was identifying false positives during the intervention (i.e., children who received the intervention without needing it) and determining when these children can stop the intervention. Schools may want to reallocate resources when children do not need the intervention (anymore). To identify those children, schools can monitor children's reading progress in Grade 1 (Fletcher et al., 2021; Gilbert et al., 2012). However, it might be risky to make children stop the intervention, as it is hard to determine which children have reached (above) average reading levels due to the intervention and which children would have developed above average reading levels without it as well. Children at risk for literacy problems might benefit from continued literacy support (Connor et al., 2013) and could therefore better not stop. In Chapter 3, it seemed that some schools indeed used progress monitoring mid-Grade 1 to identify children who did not need the intervention anymore. They stopped the intervention primarily for above-average to good readers. A remaining question is whether these children kept sufficient reading levels after the intervention was stopped.

1.4.4 *Additional Analyses*

Taken together, some questions remain, concerning the reach of the intervention. Therefore, I investigated some additional research questions regarding the selection procedure described in the Participants section of Chapters 2 and 3:

(1) Did all children who needed the intervention, receive it?
   a) **Eligible, no *Build!*.** How many children were eligible for the intervention, but did not receive *Build!*? What characterizes these children? How many of them developed reading problems in Grade 1?
   b) **False Negatives.** How many children who were not eligible for the intervention developed reading problems by the end of Grade 1?

(2) Did children (continue to) receive the intervention while they did not need it?
   a) **Not Eligible, *Build!*.** How many children were not eligible for the intervention, but did receive *Build!*? What characterizes these children?
   b) **Stopping *Build!*.** How many children with good or above-average reading skills for whom the intervention was stopped mid-Grade 1 did develop reading problems afterwards? Did stopping or continuing the intervention make a difference for children's reading skills?

Some additional analyses were conducted to provide insight in these issues.

Eligible, no Build!

Researchers and schools in one participating district (District 1) developed a screening procedure, including two screening waves, one in October and one in January in the second year of kindergarten, in which letter knowledge and phonological awareness were assessed (see Chapters 2 and 3). After the screening wave in October, schools were advised to provide extra instruction to children with low scores (lowest 30%) and to test children with below-average scores (lowest 50%) again in January. To determine who were the 30% or 50% lowest scoring children, schools were provided with a table including children's individual scores, i.e. the percentile on the phonological awareness test and the number of correct letter sounds on the letter knowledge test. Based on this table schools were expected to determine themselves who were the 30% and 50% lowest scoring children. After testing the 50% lowest scoring children in January, schools identified which children were 'at risk' and eligible for the intervention. That were children with low scores on phonological awareness (≤25th percentile) and/or letter knowledge (knowing ≤6 letters), or below-average scores on both phonological awareness (26-50th percentile) and letter knowledge (7-8 letters).

I determined which children in District 1 were and were not eligible to start with *Build!* according to this procedure. In October, 1460 children were tested. To determine the children with the 50% lowest scores that had to be tested again in January,

I first calculated children's *z*-scores within schools for both letter knowledge and phonological awareness, then took the mean of both z-scores, and subsequently determined the children with the 50% lowest z-scores. This resulted in 730 children belonging to the lowest 50%, of which 656 children (90%) were tested again in January. Of this group, 225 children (15% of the total sample) met the selection criteria in January and were deemed eligible to start with the intervention. The remaining 431 children plus the 730 children with the 50% highest scores in October (n = 1161) were considered not-eligible. I refer to the aforementioned 431 children as the *Not Eligible At-Risk* group.

Children who were eligible had substantially poorer letter knowledge skills (n = 225, M = 4.56, SD = 2.28) than the *Not Eligible At-Risk* group in January (n = 431, M = 13.64, SD = 5.59), t (654) = 29.39, p = <.001. The eligible group also had considerably poorer phonological awareness (M = 34.59 , SD = 14.85) than the *Not Eligible At-Risk* group in January (M = 59.40, SD = 19.29), t (652) = 18.24, p = <.001. This is not surprising, as these abilities were included in the selection criteria. Table 5.3 shows the mean scores of eligible children and not-eligible children on rapid automatized naming (RAN)[1] as well as child and family characteristics. These mean scores were compared with t-tests (continuous variables) or Chi-square tests (dichotomous variables). Compared to the not-eligible children, the children who were eligible had lower scores on RAN, one of the main precursors of reading development (Landerl et al., 2018; van Viersen et al., 2018). Moreover, they grew up under less favorable family circumstances, including a poorer home literacy environment, less reading activities at home, parents with a lower educational level and lower self-reported reading skills, and more often with dyslexia in the family. The group eligible children also consisted of more boys, younger children, and more children with Dutch as a second language. Thus, children who were eligible based on only two precursors of reading (letter knowledge and phonological awareness), were also characterized by a lower performance on or presence of other factors known to be related to later reading.

Next, I determined whether eligible children indeed started with *Build!* using the logs of *Build!*. If children had started *Build!* between the beginning of January and the end of June in the second kindergarten year, they were considered to have been selected for *Build!*. For three eligible children this could not be determined, because parents did not provide consent to use the logs of *Build!*. The cross classification of eligibility (yes or no) and start of *Build!* (yes or no) is presented in Table 5.4. The

[1] RAN is the ability to quickly retrieve the names of well-known symbols (pictures, numbers, or letters) from the long-term memory.

results reveal that 73% of the children who were eligible received *Build!*, but still a fair percentage, 27%, did not receive the intervention whereas they were eligible. A quarter of this group (15 children) repeated their grade the next year, and were therefore probably not provided with *Build!*. Half of the group (31 children) came from schools and years in which more than 30% of the children was eligible for the intervention, while *Build!* is meant for 25-30% of the children. Possibly, these schools may aimed to improve the classroom instruction and provided the intervention only to the children with weakest scores, as they had many low-scoring children. Another four children came from a large school where eleven children were eligible for the intervention that year, possibly exceeding the school's capacity. For the remaining 10 children, it was unclear why schools did not provide them with *Build!*. Possibly, their parents did not agree with the intervention.

Table 5.3

Differences Between Children that Were and Were Not Eligible for the Intervention

Variable	*Maximum Score*	*M*		*p*
		Eligible (n = 225)	Not Eligible (n = 1161)	
Beginning First Grade				
RAN pictures (nr. of correct items per min)	-	46.42	52.33	<.001
RAN numbers (nr. of correct items per min)	-	53.19	64.91	<.001
Child Characteristics				
% Boys	100	60.44	50.90	.024
Age in Months	-	5.47	5.69	<.001
% Dutch Second Language	100	7.21	4.15	.021
Family Characteristics				
Home Literacy Environment[b]	5	1.83	2.36	<.001
Reading Activities at Home[c]	5	4.13	4.41	.004
Parents' Educational Level[c]	5	3.30	3.57	<.001
Parents' Self-Reported Reading Skills[c]	5	3.52	3.72	<.001
% Familial Risk for Dyslexia[d]	100	34.33	18.96	<.001

[a]percentile, [b]according to the teacher, [c]average of both parents, [d]whether one of the parents thinks he/she has dyslexia

Table 5.4

Extent to Which the Selection Procedure was Followed in District 1

Procedure	*n* (%)	
	Eligible	Not Eligible
Build!	162 (73%)	152 (13%)
No *Build!*	60 (27%)	1009 (87%)

Note. The group of children for whom it was determined whether they were eligible or not eligible, consisted of only the 50% lowest scoring children on the screening wave in October kindergarten.

The next question is which factors, in addition to the eligibility criteria, schools used to determine which eligible children should or should not be provided with *Build!*. To answer this question, the group of eligible children that did not receive the intervention (*Eligible-NoBuild!*) was compared to the group of children that was eligible and got the intervention (*Eligible-Build!*) as well as to the group that was not eligible and did not receive the intervention (*NotEligible-NoBuild!*). These comparisons can provide information about the extent to which the *Eligible-NoBuild!* group resembles the *Eligible-Build!* group, sharing the eligibility, and the *NotEligible-NoBuild!* group, both not receiving the intervention. Groups were compared on preliteracy skills, as well as child and family characteristics. In Table 5.5 the characteristics of all groups obtained by a cross classification of eligibility and start with *Build!* are provided. To test whether groups differed on continuous variables, one-way ANOVAs with Tukey HSD as post-hoc statistic were conducted. To test whether groups differed on dichotomous variables, Chi-square tests with z-tests as post-hoc statistic were conducted. Note that only the children with the 50% weakest scores in October were tested again on letter knowledge and phonological awareness in January. To compare groups on letter knowledge and phonological awareness in January, I therefore made a comparison between the *Eligible-NoBuild!* group and the children from the *Not Eligible At-Risk* group who did not receive *Build!*.

Table 5.5 shows that the *Eligible-NoBuild!* group, was better than the *Eligible-Build!* group in RAN at the start of first grade and contained younger children. Thus, eligible children who were excluded from the intervention by schools tended to be younger and had relatively better RAN performance. It was also investigated whether the *Eligible-NoBuild!* group resembled the *NotEligible-NoBuild!* group, both not receiving the intervention. Children in the *Eligible-NoBuild!* group had substantially poorer letter knowledge skills in January of the second kindergarten year (n = 60, M

= 4.25, *SD* = 2.34) than the *Not Eligible At-Risk* group that did not receive *Build!* (*n* = 316, *M* = 14.97, *SD* = 5.11), *t* (374) = 15.92, *p* < .001. The *Eligible-NoBuild!* group also had considerably poorer phonological awareness (*M* = 37.19, *SD* = 13.02) than the *Not Eligible At-Risk* group that did not receive *Build!* (*M* = 61.91, *SD* = 18.72), *t* (374) = 9.78, *p* < .001. However, Table 5.5 shows that the *Eligible-NoBuild!* was similar in many other child and family characteristics to the *NotEligible-NoBuild!* group including performance on RAN, the percentage of boys, reading activities at home, parents' educational level, and parents' self-reported reading skills. Taken together, the *Eligible-NoBuild!* group was more similar to the *NotEligible-NoBuild!* group than the *Eligible-Build!* group in RAN performance. To the best of our knowledge, RAN was not examined in January of the second kindergarten year and thereby it is unlikely that schools used RAN to determine which eligible children did not need the intervention. Instead, findings suggest that schools possibly thought that younger eligible children did not need the intervention.

To investigate whether the *Eligible-NoBuild!* group did indeed not need the intervention, it was determined which percentage of the *Eligible-NoBuild!* group developed reading problems by the end of Grade 1. Children's performance on the Three-Minute-Test (DMT), measuring word reading fluency, was used. Children scoring D or E (i.e. children scoring within the lowest 25th percentile), were considered to have reading problems. The reading score was available for 43 out of 60 children in the *Eligible-NoBuild!* group. Findings reveal that 23% of the children in the *Eligible-NoBuild!* group (10 children) developed reading problems. This percentage was lower than in the *Eligible-Build!* group, χ^2 (1) = 6.38, *p* = .012, in which 45% of the children developed reading problems (64 out of 143 children for whom the reading score was available). This indicated that the *Eligible-NoBuild!* group was indeed less at risk of reading problems than the *Eligible-Build!* group. However, there were more reading problems in the *Eligbile-NoBuild!* group than in the *NotEligible-NoBuild!* group, χ^2 (1) = 4.04, *p* = .045, in which only 13% develop reading problems (i.e. 123 out of 970 children for whom the reading score was available). Thus, the *Eligbile-NoBuild!* group was still more at risk of reading problems than the *NotEligible-NoBuild!* group.

In sum, a fair number of children did not receive the intervention although they were eligible. This decision seems to be based in part on age and/or difficulty with RAN. The *Eligible-NoBuild!* group was clearly less at risk for reading problems than the *Eligible-Build!* group. A smaller percentage of this group developed reading problems, even when they did not receive the intervention. As the aim of the intervention was to reduce the number of children with reading problems in school, it would be

more effective for schools to provide the intervention to all eligible children and then identify those who do not need it as soon as possible (see paragraph Stopping *Build!*).

Table 5.5
Differences Between Groups that Were and Were not Eligible for the Intervention and that Were and Were Not Provided With the Intervention

Variable	Maximum Score	*M*				*p*
		Eligible *(n=222)*		Not Eligible *(n=1161)*		
		Build! *(n=162)*	No *Build!* *(n=60)*	*Build!* *(n=152)*	No *Build!* *(n=1009)*	
Start First Grade						
RAN pictures (nr. of correct items per min)	-	44.58_{a}	52.85_{b}	46.34_{a}	53.26_{b}	<.001
RAN numbers (nr. of correct items per min)	-	51.10_{a}	60.66_{bc}	56.39_{ab}	66.23_{c}	<.001
Child Characteristics						
% Boys	100	61.11_{a}	56.67_{ab}	63.16_{a}	49.06_{b}	<.001
Age in Months	-	5.59_{a}	5.16_{b}	5.70_{a}	5.69_{a}	<.001
% Dutch Second Language	100	6.83_{a}	8.62_{a}	8.55_{a}	3.48_{b}	0.007
Family Characteristics						
Home Literacy Environment[2]	4	1.81_{a}	1.88_{a}	1.99_{a}	2.41_{b}	<.001
Reading Activities at Home[3]	5	4.14_{a}	4.14_{ab}	4.26_{ab}	4.43_{bc}	0.008
Parents' Educational Level[e]	5	3.27_{a}	3.35_{abc}	3.36_{ab}	3.60_{cd}	<.001
Parents' Self-Reported Reading Skills[3]	5	3.48_{a}	3.67_{ab}	3.61_{ab}	3.74_{bc}	<.001
% Familial Risk for Dyslexia[4]	1	34.41_{a}	30.77_{a}	29.89_{a}	17.49_{b}	<.001

[1]percentile, [2]according to the teacher, [3]average of both parents, [4]whether one of the parents thinks he/she has dyslexia

False Negatives

I examined which percentage of the 1161 children who were not eligible for the intervention did nevertheless develop reading problems end-Grade 1. As in the previous section, children were considered to have reading problems if they belonged, according to national norms, to the weakest 25% readers on the DMT by the end of Grade 1. Fifteen percent (163 children) had reading problems, indicating that one out of every seven children who was not eligible developed reading problems in Grade 1.

As said, the screening procedure in District 1, was based on only two precursors of reading. Nevertheless, the 15% false negatives is similar to the percentages found for other early identification procedures, even when many risk factors are taken into account (e.g. Compton et al., 2010; Gijsel et al., 2006). Inclusion of extra screening tests in kindergarten might therefore not lead to a substantial further reduction of false negatives. Instead, an additional screening in first grade, including measures of actual reading skills might be better suited to identify false negatives (Gilbert et al., 2012; Torgesen, 2002). It is recommend that schools include another selection moment in Grade 1, preferably as soon as possible, for example in the autumn of Grade 1, a few months after formal reading instruction has begun. Although the absence of an additional effect of *Build!* after mid Grade 1 (Chapter 3) raises the question whether it is beneficial to begin the intervention in Grade 1, an earlier study on *Build!* (Regtvoort et al., 2013) suggests that children who begin *Build!* in Grade 1 and who are able to finish the program before the end of Grade 2, can reach average reading skills in Grade 3. When children start *Build!* in Grade 1, it may thus important that they receive the practice needed to progress quickly through the program and finish the program before the end of Grade 2.

Not Eligible, Build!

The next group to be considered are children who were not eligible for the intervention, but did receive it. Table 5.4 shows that, as expected, 87% of the children who were not eligible did not receive *Build!*. Nevertheless 13% of the not-eligible children did receive the intervention. At first glance, this percentage seems to be small. But because the not-eligible group was larger than the eligible group, the *Eligible-Build!* group and the *NotEligible-Build!* group were almost equally large. That is, 48% of the children who were included in the intervention were not eligible based on the screening procedure.

To investigate why schools decided that some not-eligible children should be provided with *Build!*, I first compared the *NotEligible-Build!* group to the *Eligible-Build!* group. To compare groups on letter knowledge and phonological awareness in

January, I made a comparison between the *Eligible-Build!* group and the children from the *Not Eligible At-Risk* group who received Build!. Similarities between these two groups may reveal reasons why schools considered the *NotEligible-Build!* group to be similar to the *Eligible-Build!* group. Table 5.5 shows that the *NotEligible-Build!* and the *Eligible-Build!* group were similar with respect to all child and family characteristics, with one exception: as to be expected, the children from the *Not Eligible At-Risk* group who received *Build!* had more letter knowledge and stronger phonological awareness than the *Eligible-Build!* group.

Second, I compared the *NotEligible-Build!* group to the *NotEligible-NoBuild!* group. To compare groups on letter knowledge and phonological awareness in January, I made a comparison between the children from the *Not Eligible At-Risk* group that did and did not receive *Build!*. Differences between these groups may reveal reasons why schools considered some eligible children to be in need of the intervention and others not. Table 5.5 shows that compared to the children from the *Not Eligible At-Risk* group who did not receive *Build!*, the children from this subgroup who did, had lower letter knowledge and phonological awareness at the beginning of the intervention. However, their scores could be considered as (above) average. Therefore, the admission of these children to the intervention was probably not based on their preliteracy skills. The *NotEligible-Build!* group also had more difficulty with RAN, consisted of more boys and more children with Dutch as a second language, came from poorer home literacy environments (reported by the teacher), had parents with lower educational levels, and interestingly contained more children with a family risk for dyslexia. These characteristics may thus have been plausible reasons for schools to provide some of the not-eligible children with the intervention.

In the *NotEligible-Build!* group, 28% of the children developed reading problems despite the fact that they had received *Build!* (i.e. 40 out of 141 children for whom the reading score was available). The percentage is higher than the percentage of children with reading problems in the *NotEligible-NoBuild!* group (13%, i.e. 123 out of 970 children for whom the reading score was available), χ^2 (1) = 24.20, $p < .001$. This suggests that the children in the *NotEligible-Build!* group were more at risk of reading problems than the other not eligible children. However, there were fewer reading problems in the *NotEligible-Build!* group than in the Eligible-Build! group, χ^2 (1) = 8.21, $p = .004$, in which, as said, 45% developed reading problems. Thus, the NotEligbile-Build! group was not as much at risk for reading problems as the *Eligible-Build!* group. Still 72% of the *NotEligible-Build!* group (101 out of 141 children) did not develop reading problems, suggesting that the group included children who did not

need the intervention, although it is hard to tell how many would have developed sufficient reading skills in the absence of the intervention.

In sum, a fair number of children received the intervention although they were not eligible. This decision seems to be based in part on difficulty with RAN, several child characteristics (being a boy or having Dutch as a second language) and/or several family characteristics (home literacy environment, parental educational level, or family risk for dyslexia). The *NotEligible-Build!* group was clearly more at risk for reading problems than the *NotEligible-NoBuild!* group. A larger percentage of this group developed reading problems even though they had received the intervention.

Earlier research showed that kindergarten teachers' predictions of which children will develop reading problems in Grade 1 can contribute to a more accurate screening of children at risk of reading problems (Gijsel et al., 2006). It might thus be beneficial to include teacher judgements when deciding which children need to start the intervention. However, current findings suggest that this should be combined with the identification of false positives, so that negative consequences of providing children with unnecessary interventions are reduced, for children themselves, as well as their parents and the school's capacity.

Stopping Build!

When children at risk of reading problems are selected in kindergarten based on the precursors of reading, some children might be incorrectly identified as at risk (Fletcher et al., 2021; Gilbert et al., 2012). To identify those children, schools can monitor children's reading progress in Grade 1 (Fletcher et al., 2021; Gilbert et al., 2012). However, it might be risky to stop the intervention for these children. Despite this risk, some schools that participated in our studies stopped the intervention for children with good or above-average reading skills by mid-Grade 1 (i.e. between January and May Grade 1, $n = 78$, of which 63 from District 1 and 15 from District 2, see Chapter 3). Additional analyses were conducted to examine the relation between stopping the intervention and the development of reading skills.

First, it was determined how many children who stopped the intervention developed reading problems later on. Test scores on the Three-Minute-Test (measure of word reading fluency) from mid-Grade 1 to end-Grade 3 were used. Children with level A (i.e. 76th to 100th percentile) were considered good readers, children with level B (i.e. 51st to 75th percentile) above-average readers, children with level C (i.e. 25-50th percentile) below-average readers, and children with levels D and E (i.e. 1st to 25th percentile) poor readers. For 6 children the reading score mid-Grade 1 was not available leaving 72 children. Not all children could be followed until end-Grade 3, as

there were three cohorts of children for whom the research project ended in Grade 3, Grade 2, and Grade 1 respectively. Moreover, there were missings due to attrition. Table 5.6 shows a cross-classification of reading level mid-Grade 1 (good, above-average, below-average, poor) by reading problems (yes or no) end-Grade 1, mid-Grade 2, end-Grade 2, mid-Grade 3, and end-Grade 3.

Table 5.6 shows that none of the good readers by mid-Grade 1 became a poor reader. One of the above-average readers and one of the below-average readers became a poor reader. Among the eleven poor readers, nine continued to be a poor reader. Thus, it seems that the overall majority of the children with sufficient reading skills when the intervention was stopped, did not develop reading problems afterwards.

Next, I matched children who stopped the intervention at the middle of Grade 1 (i.e. between January and May Grade 1; $n = 78$) to children who continued (i.e. until at least the beginning of May Grade 1; $n = 291$) on reading fluency and accuracy mid-Grade 1, and letter knowledge and phonological awareness at the intervention start. As the children who stopped were better readers than the children who continued (see Results of Chapter 3), only 59 out of the 78 children who stopped could be matched. Among them, 27% of the children had good reading skills, 29% above-average reading skills, 24% below-average reading skills, and 20% poor reading skills. Thus, both groups contained 59 children. Table 5.7 shows that the matching was successful: the groups did not differ on matching skills.

Using independent t-tests, it was investigated whether the group that stopped the intervention differed from the group that continued. These tests revealed that the groups did not differ in letter knowledge (kindergarten until mid-Grade 1), reading accuracy (start-Grade 1 until mid-Grade 2), reading fluency (mid-Grade 1 until end-Grade 2), as well as on dosage and progress in kindergarten and the first part of Grade 1 (see Table 5.7).

In sum, findings suggest that the intervention can be stopped when children have developed sufficient reading skills mid-Grade 1. This creates the possibility to provide more children with the intervention in kindergarten and/or to provide children who need the intervention with more intervention sessions. These findings also provide an answer to the questions posed in the Discussion of Chapter 3. There, it was not clear why there was no additional effect of *Build!* after mid-Grade 1. Current findings suggest that for the better readers, continuing the intervention after mid-Grade 1 did not make a difference. Put differently, *Build!* seemed to have no effect for the better readers anymore after mid-Grade 1. Poorer readers however may still

benefit from continuing the intervention, because the results on the matched groups only included a small group of the poorer readers who continued the intervention.

Table 5.6

Reading Development of Children who Stopped the Intervention Mid-Grade 1

	Number of Children per Reading Level Mid-Grade 1			
	Good	*Above-Average*	*Below-Average*	*Poor*
Reading Problems End-Grade 1				
No	29	17	13	4
Yes	0	0	1[b]	7[c]
Reading Problems Mid-Grade 2				
No	14	9	7	1
Yes	0	1[a]	1[a]	4
Reading Problems End-Grade 2				
No	14	10	8	2
Yes	0	0	0	2[d]
Reading Problems Mid-Grade 3				
No	6	3	3	1
Yes	0	0	0	0
Reading Problems End-Grade 3				
No	7	1	3	0
Yes	0	1[b]	0	1[b]

[a]This child became a below-average reader end-Grade 2. [b] This was the last observed reading score for this child.[c] All these children showed reading problems on the last observation (i.e. for five children end-Grade 1, for one child mid-Grade 2, and for one child end-Grade 2). [d] For another child besides [c], this was the last observed reading score.

Table 5.7

Descriptive statistics of the groups that stopped or continued the intervention after mid-Grade 1

	Stopped		Continued		*p*
	M	*SD*	*M*	*SD*	
Letter knowledge					
Middle of the Second Kindergarten Year[a]	6.82	3.67	6.59	3.43	.754
End of Kindergarten	15.81	5.34	14.04	5.02	.088
Start-Grade 1	23.69	6.55	22.44	5.06	.261
Mid-Grade 1	33.22	2.28	33.03	2.04	.640
Phonological Awareness[b]					
Middle of the Second Kindergarten Year[a]	52.96	19.38	52.28	17.34	.932
Reading Accuracy					
Start-Grade 1	10.71	4.84	9.65	4.20	.218
Mid-Grade 1[a]	15.31	8.82	14.88	8.74	.794
End-Grade 1	29.04	8.62	26.29	9.67	.108
Mid-Grade 2	37.42	2.06	35.28	6.54	.086
Reading Fluency					
Mid-Grade 1[a]	15.47	7.58	15.47	7.38	1.000
End-Grade 1	28.72	13.38	28.52	14.48	.939
Mid-Grade 2	44.59	17.54	42.65	19.14	.669
End-Grade 2	53.87	14.65	51.50	20.57	.601
Dosage[c]					
Second Kindergarten Year	6.65	3.28	6.14	2.99	.387
First Part of Grade 1	6.49	2.54	5.88	2.51	.189
Progress[d]					
Second Kindergarten Year	44.08	28.22	45.09	25.32	.964
First Part of Grade 1	69.00	29.29	64.14	30.32	.379

[a]matching variable, [b]percentile, [c]the number of hours spent on the intervention, [d]-the number of new intervention lessons completed

1.4.5 *Conclusion on Reach*

The first challenge mentioned regarding the selection of children for an early-literacy intervention is that schools need to develop of a brief and accurate screening procedure. In kindergarten, screening cannot reliably identify which children will develop reading problems, even when many risk factors are taken into account (Catts et al., 2001). Inevitably, some children are identified as 'not at risk' while they will develop reading problems (false negatives) and some children are identified as 'at risk' while they will develop sufficient reading skills (false positives). Screenings may be more reliable when larger screening batteries are administered (Catts et al., 2001), but schools often lack time and/or professionals to administer them. Findings reveal that the selection procedure in District 1, which was designed to be brief and teacher-friendly, produced a similar percentage of false negatives as found in other screening procedures for kindergarten (e.g. Compton et al., 2010; Gijsel et al., 2006). Therefore, the proposed selection procedure is a good starting point for Dutch schools to identify children at risk of reading problems. However, findings also indicate that the screening procedure could be improved, as there were children who were 'missed' (false negatives). It is recommended to implement an additional screening wave, a few months after the beginning of formal reading instruction. It is important that children who start *Build!* in Grade 1 receive the practice needed to finish the program end Grade 2 (Regtvoort et al., 2013). This additional screening wave allows schools to accurately identify children who specifically struggle with reading skills.

The second challenge I mentioned regarding early intervention was that schools have to make decisions on which children are provided with the intervention, if they do not have the capacity to provide all "at risk" children with the intervention. Findings suggest that schools and/or parents indeed decided to provide the intervention to some not-eligible children and not to provide the intervention to some eligible children. Results indicate this had both positive and negative consequences for reach, i.e. the percentage of the target population that participated in the intervention. Reach can affect the effectivity of an intervention program when it is scaled up (Glasgow et al., 1999, 2006). As such, schools may want to play safe, i.e. to provide all eligible children with the intervention and to additionally select some not-eligible children who they consider to be in need of the intervention. However, selecting many children for the intervention, even when they are not eligible, has consequences for the number of false positives and thereby can constrain schools' time to provide children with frequent practice at school. Therefore, it is recommended to combine this 'play-safe approach' with the identification of false positives during the

intervention, so that schools can spent their time to only the children who really need the intervention in Grade 1.

The identification of false positives during the intervention (i.e. children who are provided with the intervention, while not needing it) was the third challenge of early intervention that I mentioned. In general, early screening leads to fairly high percentages of false positives, i.e. children who were predicted to have reading difficulties, but who turned out to be average-to-good readers (41% in Catts et al., 2001; 31-53% in Foorman et al., 2014; 23-29% in Wood et al., 2005). This constrains schools' time and resources and can be at the expense of the dosage for children who really need the intervention. One way to identify false positives is by progress monitoring. Our findings suggest that schools could stop the intervention for children who have developed adequate reading skills by mid-Grade 1. This creates more time for children who really need the intervention or to provide more children with the intervention in kindergarten.

The three challenges in optimizing reach have gained little attention in research on early-literacy interventions so far. Mostly, research is focused on the effectiveness of interventions, while reach (here: the selection of children) is of equal importance (Glasgow et al., 2006). The extent to which schools succeed in selecting the target group of children (i.e. children who will develop reading problems) can determine the success of the intervention in school. Therefore, more research on factors that contribute to accurate screening in kindergarten can help schools to better understand how to reach the target group. Moreover, research on identifying and addressing false positives and false negatives could equip schools with the knowledge to more accurately target and support children who really need the intervention.

1.5 *Future Directions*

The main focus of the studies in this dissertation was on the effects of *Build!* on literacy and reading outcomes. When implementing interventions in schools, Glasgow et al. (1999) suggest that researchers should also investigate possible negative or unintended consequences of the intervention. One potential negative consequence of programs that are provided (partly) at home, like *Build!*, is an increase in parent-child conflict during practice with the program (de Jong et al., 2022). Parents are no longer solely caregivers, but become teachers as well. Especially parents who perceive stress or household chaos and parents who are insecure about their teaching skills tend to experience parent-child conflict during schoolwork (de Jong et al., 2022). Such negative parent-child interactions can, in turn, lead to lower child self-esteem and

academic performance (Moed et al., 2017; Morrison et al., 2003; Wang et al., 2021). Another negative effect of an early-literacy intervention like *Build!* could be that children may learn that their reading ability is lower than those of other children, because they are provided with extra instruction while other children are not. This stigma may result in a lowered academic self-concept and/or reading enjoyment (Boliver & Capsada-Munsech, 2021; Campbell, 2021). Future research is needed to investigate such unintended negative consequences of early-literacy interventions, so that it can be decided whether positive and negative effects balance out.

An interesting finding in this dissertation is that *Build!* was found to be effective only after schools had been implementing the intervention for two or more years (Chapter 4). Then, there were small decreases in the percentage of children with difficulties in reading fluency, spelling, and reading comprehension, and the average reading fluency and spelling ability of children in cohorts receiving *Build!* started to increase. Why this took a few years, is yet unknown. A suggestion for future research is to examine the effects of large-scale intervention over several years and to investigate what happens during these years in the schools. Is the schools' experience with the intervention related to treatment integrity or to the phases of implementation? This dissertation also did not provide insight in whether the effects stabilized after several years: the number of children with reading problems decreased, but was this an ending trend? This dissertation stresses the importance of long-term investigations of effectiveness. Future research is needed to investigate even longer-term trends.

Another suggestion for future research is to study what kinds of support schools need to reach a sufficient level of dosage or treatment integrity. A good example is the study of Stein et al. (2008) who investigated the effect of three levels of schools support on the implementation of the early-literacy intervention *Kindergarten Peer Assisted Learning Strategies* and, in turn, on children's outcomes. The three levels of support were: (1) one-day training during which teachers learned about the program, shared ideas, and familiarized themselves with the program and its materials, (2) one-day training and two follow-up sessions with other teachers during which questions were answered, problems were discussed, and ideas were shared, and (3) one-day training, two follow-up sessions, and weekly technical support of a trained assistant in the classroom. Findings indicated that each level of support contributed to teachers' treatment integrity (adherence in their study) which, in turn, predicted children's reading achievement. More research is needed to investigate what kind of support contributes to other dimensions of treatment integrity, such as dosage. In

my opinion not only support at the school level is of interest, but also support at the regional level (groups of schools), as schools can learn from each other.

Issues on reach have gained less attention in reading intervention research so far, while it is essential for the impact of an intervention (Glasgow et al., 2006). The better schools are able to reach the target population, the larger difference the intervention can make. Research is needed to show which tests at what moment are useful and which selection criteria (cut-off points) lead to an optimal selection of children at risk, so that few children at risk are 'missed' and few children are erroneously identified as at risk. It is yet unclear whether screening procedures with multiple screening waves or dynamic assessment are best to identify at-risk children in kindergarten (Cho et al., 2017; Compton et al., 2010; Gilbert et al., 2012; Thomson et al., 2020). Moreover, such dynamic assessment procedures have to be shortened and made user-friendly for schools. More research is needed to investigate whether a short version of the intervention could be used as a form of dynamic assessment to identify children at risk for reading problems more accurately in kindergarten.

1.6 *Practical Implications*

This dissertation provides multiple implications for educators and schools that (want to) use an early-literacy intervention. First, intervention effects clearly depend on how much children practice with the intervention. The findings in this dissertation reveal that frequency of practice is most important in both kindergarten and Grade 1, more important than session length. To reach the full potential of an intervention, schools may want to monitor and stimulate frequent practice. However, a focus on frequency might not be sufficient. Some children progress more quickly through the program than others while spending as much time on the intervention, particularly children with dyslexia in the family and children with lower preliteracy skills at the start of the intervention. Schools are recommended to monitor not only children's practice, but also their progress within the intervention, and to adjust the frequency of practice to ensure sufficient progress. Program developers could help by providing an indication of the progress through the intervention the average child is expected to make. Such an indication might also be useful for schools to reach the progress within the program that is needed to maintain the preventive character of the intervention. If progress through the program is too slow, then its content might become very similar to the instruction in the classroom.

A second implication is that it may take several years of experience with the intervention before it will show its effects at the school level. The reduction of reading

problems and the improvement in average reading skills at the school level appeared after two or more years of implementation. Schools thus need to be patient and to continue optimizing the implementation of an intervention for multiple years to benefit from their efforts. Program developers could facilitate this process by providing adequate training[2], but this might not be enough (Stein et al., 2008). School boards and school leaders can contribute to the implementation of the intervention by making teachers learn from each other, i.e. by facilitating follow-up sessions with other teachers to review intervention procedures, to identify implementation issues, and to solve problems together with other teachers.

A third implication of the current studies is that the selection of children for an early-literacy intervention is challenging. Schools are often not provided with a screening procedure when they buy an intervention and they have to develop one themselves. For the studies in this dissertation, schools, researchers, and intervention developers collaborated on the development of a brief screening procedure that could be used by the schools to select children for the intervention. Findings indicate that this procedure provides a good starting point for Dutch schools to identify children at risk of reading problems. It entailed a two-step selection procedure, in which the most important predictors of reading were assessed (letter knowledge and phonological awareness, the latter with a standardized measure) in October and January of the second kindergarten year. Between these two screening waves, schools provided additional instruction to children with poor pre-literacy skills. Children who had again poor preliteracy skills at the second wave, were selected for the intervention.

Inevitably, any selection procedure in kindergarten will be imperfect. A second screening at the begining of or a few months after the beginning of formal reading instruction, might be necessary. To identify children who need to start the intervention in Grade 1, schools could measure children's reading skills in the classroom, but they could also provide (some) children with a shortened version of the intervention and select the children who do not make the expected progress within the intervention, as progress has been shown to be an indicator of reading skills in Chapters 2 and 3. Imperfect screening in kindergarten also indicates that some children are selected for the program who do not need it. Schools might want to identify those children as they want to spent their time efficiently. The results reported in this dissertation

[2] A training workshop could focus on (a) research indicating the positive effects of the intervention, (b) videos in which the intervention is implementation in diverse classrooms, (c) the key program components, and (d) role-play activities in which teachers implement the intervention (Stein et al., 2008).

indicate that children who became above-average readers by mid-Grade 1 may not benefit from continuing the intervention and did not develop reading problems afterwards. Findings indicate that children who had started the intervention in kindergarten and who developed above-average reading skills by mid-Grade 1 could stop the intervention. Children who start *Build!* in Grade 1 and reach above-average reading skills may still benefit from continuing the intervention after mid Grade 1 (Regtvoort, 2013).

1.7 *Conclusion*

This dissertation emphasized the importance of studying interventions not only in researcher-led RCTs, but also when they are implemented on a large scale, under schools' own responsibility. Findings indicate that effects of evidence-based interventions might only show up after schools have implemented the intervention for several years, that these effects may be small, and that intervention outcomes can be affected by many factors, such as the amount of practice, the extent to which schools succeed in selecting the right children for the intervention, and whether or not there are learning problems in the family. In educational research, few interventions have been studied on a large scale, while in health research this is common practice. This dissertation presents both methods and initial results on how to study the effectiveness of an intervention that is implemented on a large scale without researcher contol, and factors influencing this effectiveness. More large-scale studies on the effectiveness of interventions in natural school settings are needed to provide schools with information on how to make a success of the implementation of evidence-based interventions. This dissertation has revealed two factors that were related and may contribute to this success: providing children with *weeks of practice* and having multiple *years of experience* with the intervention.

References

Al Otaiba, S., & Fuchs, D. (2002). Characteristics of children who are unresponsive to early literacy intervention: A review of the literature. *Remedial and Special Education*, *23*(5), 300-316. https://doi.org/10.1177/07419325020230050501

Al Otaiba, S., Schatschneider, C., & Silverman, E. (2005). Tutor-assisted intensive learning strategies in kindergarten: How much is enough? *Exceptionality*, *13*(4), 195-208. https://doi.org/10.1207/s15327035ex1304_2

Álvarez-Cañizo, M., Suárez-Coalla, P., & Cuetos, F. (2015). The role of reading fluency in children's text comprehension. *Frontiers in Psychology*, *8*, Article 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

Bautista, G. F., Ghesquière, P., & Torbeyns, J. (2023). Stimulating preschoolers' early literacy development using educational technology: A systematic literature review. *International Journal of Child-Computer Interaction*, *36*, 100620. https://doi.org/10.1016/j.ijcci.2023.100620

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

Boliver, V., & Capsada-Munsech, Q. (2021). Does ability grouping affect UK primary school pupils' enjoyment of Maths and English? *Research in Social Stratification and Mobility*, *76*, 100629. https://doi.org/10.1016/j.rssm.2021.100629

Bradshaw, C. P., Debnam, K., Koth, C. W., & Leaf, P. (2009). Preliminary validation of the implementation phases inventory for assessing fidelity of schoolwide positive behavior supports. *Journal of Positive Behavior Interventions*, *11*(3), 145-160. https://doi.org/10.1177/1098300708319126

Buisman, M., Allen, J., Fouarge, D., Houtkoop, W., & van der Velden, R. (2013). *PIAAC 2012: De belangrijkste resultaten* [Summary of a large-scale Dutch study on skills that are key to work and daily life]. Expertisecentrum Beroepsonderwijs.

https://ecbo.nl/wp-content/uploads/sites/3/2013-10-PIAAC-Kernvaardigheden-voor-Werk-en-Leven-2.pdf

Burgess, S. R., & Lonigan, C. J. (1998). Bidirectional relations of phonological sensitivity and prereading abilities: Evidence from a preschool sample. *Journal of Experimental Child Psychology*, *70*(2), 117-141. https://doi.org/10.1006/jecp.1998.2450

Burkhardt, H., & Schoenfeld, A. H. (2003). Improving educational research: Toward a more useful, more influential, and better-funded enterprise. *Educational Researcher*, *32*(9), 3-14. https://doi.org/10.3102/0013189X032009003

Bursuck, B., & Blanks, B. (2010). Evidence-based early reading practices within a response To intervention system. *Psychology in the Schools*, *47*(5), 421-431. https://doi.org/10.1002/pits.20480

Bus, A. G., & Van IJzendoorn, M. H. (1999). Phonological awareness and early reading: A meta-analysis of experimental training studies. *Journal of Educational Psychology*, *91*(3), 403-414. https://doi.org/10.1037/0022-0663.91.3.403

Byrne, B., Coventry, W. L., Olson, R. K., Samuelsson, S., Corley, R., Willcutt, E. G., Wadsworth, S., & DeFries, J. C. (2009). Genetic and environmental influences on aspects of literacy and language in early childhood: Continuity and change from preschool to Grade 2. *Journal of Neurolinguistics*, *22*(3), 219-236. https://doi.org/10.1016/j.jneuroling.2008.09.003

Byrne, B., Fielding-Barnsley, R., & Ashley, L. (2000). Effects of preschool phoneme identity training after six years: Outcome level distinguished from rate of response. *Journal of Educational Psychology*, *92*(4), 659-667. https://doi.org/10.1037/0022-0663.92.4.659

Campbell, T. (2021). In-class 'ability'-grouping, teacher judgements and children's mathematics self-concept: Evidence from primary-aged girls and boys in the UK Millennium Cohort Study. *Cambridge Journal of Education*, *51*(5), 563-587. https://doi.org/10.1080/0305764X.2021.1877619

Capin, P., Walker, M. A., Vaughn, S., & Wanzek, J. (2018). Examining ow treatment tidelity is supported, measured, and reported in K-3 reading intervention research. *Educational Psychology Review*, *30*(3), 885-919. https://doi.org/10.1007/s10648-017-9429-z

Carpenter, S. K., Pan, S. C., & Butler, A. C. (2022). The science of effective learning with spacing and retrieval practice. *Nature Reviews Psychology*, *1*, 496,511. https://doi.org/10.1038/s44159-022-00089-1

Catts, H. W., Fey, M. E., Zhang, X., & Tomblin, J. B. (2001). Estimating the risk of future reading difficulties in kindergarten children: A research-based model and its clinical implementation. *Language, Speech, and Hearing Services in Schools*, *32*(1), 38–50. https://doi.org/10.1044/0161-1461(2001/004)

CBS (2022a). Bevolking; hoogstbehaald onderwijsniveau en onderwijsrichting [StatLine: Population; highest educational level and educational field]. https://opendata.cbs.nl/#/CBS/nl/dataset/85184NED/table?dl=64EF3

CBS (2022b). Leerlingen in (speciaal) basisonderwijs; migratieachtergrond, woonregio [StatLine: Students in primary school (special needs); migrant background; residential region]. https://opendata.cbs.nl/#/CBS/nl/dataset/83295NED/table?ts=1648731385428

CBS (2023). Bevolking; hoogstbehaald onderwijsniveau en onderwijsrichting, 2003-2022 [Population; highest educational level and educational program, 2003-2022]. Retrieved from https://opendata.cbs.nl/statline/#/CBS/nl/dataset/85184NED/table?ts=1688109190019

Cheung, A. C. K., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, *45*(5), 283-292. https://doi.org/10.3102/0013189X16656615

Cho, E., Compton, D. L., Gilbert, J. K., Steacy, L. M., Collins, A. A., & Lindström, E. R. (2017). Development of first-graders' word reading skills: For whom can dynamic assessment tell us more? *Journal of Learning Disabilities*, *50*(1), 95-112. https://doi.org/10.1177/0022219415599343

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. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203771587

Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology*, *112*(4), 558-577. https://doi.org/10.1037/0021-843X.112.4.558

Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Gilbert, J. K., Barquero, L. A., Cho, E., & Crouch, R. C. (2010). Selecting at-risk first-grade readers for early intervention: Eliminating false positives and exploring the promise of a two-stage gated screening process. *Journal of Educational Psychology*, *102*(2), 327-340. https://doi.org/10.1037/a0018448

Connor, C. M. D., Morrison, F. J., Fishman, B., Crowe, E. C., Al Otaiba, S., & Schatschneider, C. (2013). A longitudinal cluster-randomized controlled study on the accumulating effects of individualized literacy instruction on students' reading from first through third grade. *Psychological Science*, *24*(8), 1408-1419. https://doi.org/10.1177/0956797612472204

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

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review*, *18*(1), 23-45. https://doi.org/10.1016/S0272-7358(97)00043-3

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 Jong, P. & Wolters, G. (2002*). Fonemisch Bewustzijn, Benoemsnelheid en Leren Lezen* [Phonemic awareness, naming speed and learning to read]. *Pedagogische Studiën*, *89*(1), 53-63. https://pedagogischestudien.nl/download?type=document&identifier=617257

de Jong, P. F., & Share, D. L. (2007). Orthographic learning during oral and silent reading. *Scientific Studies of Reading*, *11*(1), 55-71. https://doi.org/10.1207/s1532799xssr1101_4

de Jong, P. F., & van der Leij, A. (1999). Specific contributions of phonological abilities to early reading acquisition: Results from a Dutch latent variable longitudinal study. *Journal of Educational Psychology*, *91*(3), 450-476. https://doi.org/10.1037/0022-0663.91.3.450

de Jong, P. F., & Vrielink, L. O. (2004). Rapid automatic naming: Easy to measure, hard to improve (quickly). *Annals of Dyslexia*, *54*(1), 65-88. https://doi.org/10.1007/s11881-004-0004-1

de Jong, P. F., Schreurs, B. G. M., & Zee, M. (2022). Parent–child conflict during homeschooling in times of the COVID-19 pandemic: A key role for mothers' self-efficacy in teaching. *Contemporary Educational Psychology*, *70*, 102083. https://doi.org/10.1016/j.cedpsych.2022.102083

de Jong, P. F., Schreurs, B. G. M., van der Weijden, F. A., & Cornelissen, F. (2023). Wat maakt de grootschalige implementatie van een evidence-based programma voor de preventie van leesproblemen tot een succes? https://www.nro.nl/sites/nro/files/media-files/eindrapport_-_evidence_based_programma.pdf

de Leescoalitie (2020). *Oproep tot een Ambitieus Leesoffensief* [Call for an Ambitious Plan to Improve Reading]. Retrieved from https://tijdvooreenleesoffensief.nl/wp-content/uploads/2021/02/Manifest-Leesoffensief.pdf

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.

Denton, C. A., Nimon, K., Mathes, P. G., Swanson, E. A., Kethley, C., Kurz, T. B., & Shih, M. (2010). Effectiveness of a supplemental early reading intervention scaled up in multiple schools. *Exceptional Children*, *76*(4), 394-416. https://doi.org/10.1177/001440291007600402

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology*, *84*(5), 795-805. https://doi.org/10.1037/0021-9010.84.5.795

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising

directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4-58. https://doi.org/10.1177/1529100612453266

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

Earle, J., Maynard, R., Neild, R. C., Easton, J. Q., Ferrini-Mundy, J., Albro, E., Cai, J., Cator, K., Fulmer, G., Gummer, E., Hamos, J., Lach, M., Lesnick, J., Okagaki, L., Kolodner, J., Pestronk, J., Ricciuti, A., Rimdzius, T., Ruby., A., … Winter, S. (2013). *Common Guidelines for Education Research and Development*. Institute of Education Sciences and the National Science Foundation. https://www.nsf.gov/pubs/2013/nsf13126/nsf13126.pdf

Egberink, I.J.L., Leng, W.E. de, & Vermeulen, C.S.M. (2009-2023). *COTAN Documentatie*. Boom Uitgevers Amsterdam. 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., 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

Elbro, C., & de Jong, P. F. (2017). Orthographic learning is verbal learning: The role of spelling pronunciations. In K. Cain, D. L. Compton & R. K. Parrila (Eds.), *Theories of Reading Development* (pp. 169–190). John Benjamins Publishing Company. https://doi.org/10.1075/swll.15.10elb

Elbro, C., & Petersen, D. K. (2004). Long-term effects of phoneme awareness and letter sound training: An intervention study with children at risk for dyslexia. *Journal of Educational Psychology*, *96*(4), 660-670. https://doi.org/10.1037/0022-0663.96.4.660

Elbro, C., de Jong, P. F., Houter, D., & Nielsen, A. M. (2012). From spelling pronunciation to lexical access: A second step in word decoding? *Scientific Studies of Reading*, *16*(4), 341–359. https://doi.org/10.1080/10888438.2011.568556

Eleveld, M. A. (2005). *At risk for dyslexia: The role of phonological abilities, letter knowledge, and speed of serial naming in early intervention and diagnosis* [Doctoral dissertation, University of Groningen]. RUG Research Portal. https://research.rug.nl/en/publications/at-risk-for-dyslexia-the-role-of-phonological-abilities-letter-kn

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. *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*, *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*, 553-576. https://doi.org/10.1007/s10648-011-9175-6

Fluss, J. M. D., 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*, *30*(3), 206-216. https://doi.org/10.1097/DBP.0b013e3181a7ed6c

Fogarty, M., Oslund, E., Simmons, D., Davis, J., Simmons, L., Anderson, L., Clemens, N., Roberts, G. (2014). Examining the effectiveness of a multicomponent reading comprehension intervention in middle schools: A focus on treatment fidelity. *Educational Psychology Review*, *26*(3), 425–449. https://doi.org/10.1007/s10648-014-9270-6

Foorman, B. R., Petscher, Y., & Schatschneider, C. (2015). Florida Center for Reading Research (FCRR) Reading Assessment (FRA): Grades 3 through 12. Technical Manual. http://files.eric.ed.gov/fulltext/ED580133.pdf

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), Article e105843. https://doi.org/10.1371/journal.pone.0089900

Gijsel, M. A. R., Bosman, A. M. T., & Verhoeven, L. (2006). Kindergarten risk factors, cognitive factors, and teacher judgments as predictors of early reading in Dutch. *Journal of Learning Disabilities*, *39*(6), 558–571. https://doi.org/10.1177/00222194060390060701

Gilbert, J. K., Compton, D. L., Fuchs, D., & Fuchs, L. S. (2012). Early screening for risk of reading disabilities: Recommendations for a four-step screening system. *Assessment for Effective Intervention*, *38*(1), 6-14. https://doi.org/10.1177/1534508412451491

Glasgow, R. E., Klesges, L. M., Dzewaltowski, D. A., Estabrooks, P. A., & Vogt, T. M. (2006). Evaluating the impact of health promotion programs: Using the RE-AIM framework to form summary measures for decision making involving complex issues. *Health Education Research*, *21*(5), 688–694. https://doi.org/10.1093/her/cyl081

Glasgow, R. E., Vogt, T. M., & Boles, S. M. (1999). Evaluating the public health impact of health promotion interventions: The RE-AIM framework. *American Journal of Public Health*, *89*(9), 1322-1327. https://doi.org/10.2105/AJPH.89.9.1322

Gollob, H. F., & Reichardt, C. S. (1987). Taking account of time lags in causal models. *Child Development*, *58*(1), 80-92. https://doi.org/10.1111/j.1467-8624.1987.tb03492.x

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

Hamilton, L. G., Hayiou-Thomas, M. E., Hulme, C., & Snowling, M. J. (2016). The home literacy environment as a predictor of the early literacy development of children at family-risk of dyslexia. *Scientific Studies of Reading*, *20*(5), 401-419. https://doi.org/10.1080/10888438.2016.1213266

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*(3), 181-193. https://doi.org/10.1177/001440291307900204

Hatcher, P. J., Hulme, C., & Ellis, A. W. (1994). Ameliorating early reading failure by integrating the teaching of reading and phonological skills: The phonological linkage hypothesis. *Child Development*, *65*(1), 41-57. https://doi.org/10.1111/j.1467-8624.1994.tb00733.x

Hatcher, P. J., Hulme, C., & Snowling, M. J. (2004). Explicit phoneme training combined with phonic reading instruction helps young children at risk of reading failure. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *45*(2), 338-358. https://doi.org/10.1111/j.1469-7610.2004.00225.x

Hemmerechts, K., Agirdag, O., & Kavadias, D. (2017). The relationship between parental literacy involvement, socio-economic status and reading literacy. *Educational Review*, *69*(1), 85-101. https://doi.org/10.1080/00131911.2016.1164667

Hindson, B., Byrne, B., Fielding-Barnsley, R., Newman, C., Hine, D. W., & Shankweiler, D. (2005). Assessment and early instruction of preschool children at risk for reading disability. *Journal of Educational Psychology*, *97*(4), 687-704. https://doi.org/10.1037/0022-0663.97.4.687

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.

Hsin, L. B., Miratrix, L., Kim, H. Y., LaRusso, M. D., & Snow, C. E. (2023). Predictable variation in the implementation of a curricular intervention—and why it matters. *The Elementary School Journal*, *124*(1), 1-30. https://doi.org/10.1086/725765

IBM. (2017). *IBM SPSS Statistics Software for Windows, Version 25*. IBM Corp.

Inspectie van het Onderwijs [Schools Inspectorate] (2019). Verschillen tussen scholen nader bekeken [Focusing on differences between schools]. Retrieved from https://www.onderwijsinspectie.nl/binaries/onderwijsinspectie/documenten/themarapporten/2019/04/10/dyslexieverklaringen-verschillen-tussen-

scholen-nader-bekeken/Rapport+Dyslexieverklaringen+verschillen+tussen+scholen+nader+bekeken.pdf

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, J. S., & Quinn, D. M. (2019). The effects of summer reading on low-income children's literacy achievement from kindergarten to grade 8: A meta-analysis of classroom and home interventions. *Review of Educational Research*, *83*(3), 386-431. https://doi.org/10.3102/0034654313483906

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

Kline, R. B. (2015). *Principles and practices of structural equation modelling* (4th ed.). Guilford Publications.

Kort, W., Schittekatte, M., & Compaan, E. (2008). *CELF-4-NL: Clinical evaluation of language fundamentals*. Pearson Assessment.

Krijnen, E., van Steensel, R., Meeuwisse, M., Jongerling, J., & Severiens, S. (2020). Exploring a refined model of home literacy activities and associations with children's emergent literacy skills. *Reading and Writing*, *33*, 207-238. https://doi.org/10.1007/s11145-019-09957-4

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.

Lam, E. A., & McMaster, K. L. (2014). Predictors of responsiveness to early literacy intervention: A 10-year update. *Learning Disability Quarterly*, *37*(3), 134-147. https://doi.org/10.1177/0731948714529772

Landerl, K., Freudenthaler, H. H., Heene, M., De Jong, P. F., Desrochers, A., Manolitsis, G., Parrila, R., & Georgiou, G. K. (2018). Phonological awareness and rapid automatized naming as longitudinal predictors of reading in five alphabetic orthographies with varying degrees of consistency. *Scientific Studies of Reading*, *23*(3), 220–234. https://doi.org/10.1080/10888438.2018.1510936

Leppänen, U., Aunola, K., Niemi, P., & Nurmi, J. E. (2008). Letter knowledge predicts Grade 4 reading fluency and reading comprehension. *Learning and Instruction*, *18*(6), 548-564. https://doi.org/10.1016/j.learninstruc.2007.11.004

Leseman, P. P. M., & Jong, P. F. (1998). Home literacy: Opportunity, instruction, cooperation and social-emotional quality predicting early reading achievement. *Reading Research Quarterly*, 33(3), 294-318. https://doi.org/10.1598/rrq.33.3.3

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. (2017). Working toward a more literate world: Reading intervention commentary. *New Directions for Child and Adolescent Development*, *2017*(155), 131-141. https://doi.org/10.1002/cad.20190

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

Luyten, H., & Bruggencate, G. (2011). The presence of matthew effects in dutch primary education, development of language skills over a six-year period. *Journal of Learning Disabilities*, *44*(5), 444-458. https://doi.org/10.1177/0022219411410289

Manz, P. H., Hughes, C., Barnabas, E., Bracaliello, C., & Ginsburg-Block, M. (2010). A descriptive review and meta-analysis of family-based emergent literacy interventions: To what extent is the research applicable to low-income,

ethnic-minority or linguistically-diverse young children? *Early Childhood Research Quarterly*, *25*(4), 409-431. https://doi.org/10.1016/j.ecresq.2010.03.002

Marulis, L. M., & Neuman, S. B. (2010). The effects of vocabulary intervention on young children's word learning: A meta-analysis. *Review of Educational Research*, *80*(3), 300-335. https://doi.org/10.3102/0034654310377087

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

Moed, A., Gershoff, E. T., Eisenberg, N., Hofer, C., Losoya, S., Spinrad, T. L., & Liew, J. (2017). Parent–child negative emotion reciprocity and children's school success: An emotion-attention process model. *Social Development*, *26*(3), 560-574. https://doi.org/10.1111/sode.12217

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

Morrison, E. F., Rimm-Kauffman, S., & Pianta, R. C. (2003). A longitudinal study of mother-child interactions at school entry and social and academic outcomes in middle school. *Journal of School Psychology*, *41*(3), 185-200. https://doi.org/10.1016/S0022-4405(03)00044-X

Muñez, D., Lee, K., Bull, R., Khng, K. H., Cheam, F., & Rahim, R. A. (2022). Working memory and numeracy training for children with math learning difficulties: Evidence from a large-scale Implementation in the classroom. *Journal of Educational Psychology*. Advance online publication. https://doi.org/10.1037/edu0000732

Muter, V., Hulme, C., Snowling, M. J., & Stevenson, J. (2004). Phonemes, rimes, vocabulary, and grammatical skills as foundations of early reading development:

Evidence from a longitudinal study. *Developmental psychology*, *40*(5), 665-681. https://doi.org/10.1037/0012-1649.40.5.665

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus User's Guide* (8th ed.). Muthén & Muthén. https://www.statmodel.com/download/usersguide/MplusUserGuideVer_8.pdf

NCEE (2008). *The Reading First Program*. Retrieved from https://ies.ed.gov/ncee/pubs/20094038/summ_a.asp

Nelson, J. R., Benner, G. J., & Gonzalez, J. (2003). Learner characteristics that influence the treatment effectiveness of early literacy interventions: A meta-analytic review. *Learning Disabilities Research & Practice*, *18*(4), 255–267. https://doi.org/10.1111/1540-5826.00080

Nunnery, J. A., Ross, S. M., & McDonald, A. (2006). A Randomized experimental evaluation of the impact of accelerated reader/reading renaissance implementation on reading achievement in grades 3 to 6. *Journal of Education for Students Placed at Risk*, *11*(1), 1-18. https://doi.org/10.1207/s15327671espr1101_1

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research*, *78*(1), 33–84. https://doi.org/10.3102/0034654307313793

Pecini, C., Spoglianti, S., Bonetti, S., Di Lieto, M. C., Guaran, F., Martinelli, A., Gasperini, F., Cristofani, P., Casalini, C., Mazzotti, S., Salvadorini, R., Bargagna, S., Palladino, P., Cismondo, D., Verga, A., Zorzi, C., Brizzolara, D., Vio, C., & Chilosi, A. M. (2019). Training RAN or reading? A telerehabilitation study on developmental dyslexia. *Dyslexia*, *25*(3), 318-331. https://doi.org/10.1002/dys.1619

Petrill, S. A., Deater-Deckard, K., Schatschneider, C., & Davis, C. (2005). Measured environmental influences on early reading: Evidence from an adoption study. *Scientific Studies of Reading*, *9*(3), 237-259. https://doi.org/10.1207/s1532799xssr0903_4

Phillips, B. M., & Lonigan, C. J. (2009). Variations in the home literacy environment of preschool children: A cluster analytic approach. *Scientific Studies of Reading*, *13*(2), 146-174. https://doi.org/10.1080/10888430902769533

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team (2019*). nlme: Linear and nonlinear mixed effects models*. R package version 3.1-140. https://CRAN.R-project.org/package=nlme

Poskiparta, E., Niemi, P., Lepola, J., Ahtola, A., & Laine, P. (2003). Motivational-emotional vulnerability and difficulties in learning to read and spell. *British Journal of Educational Psychology*, *73*(2), 187-206. https://doi.org/10.1348/00070990360626930

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/feduc.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. https://doi.org/10.1002/dys.1465

RStudio Team (2020). *RStudio: Integrated Development for R*. RStudio, PBC. http://www.rstudio.com/.

Saine, N. L., Lerkkanen, M. K., Ahonen, T., Tolvanen, A., & Lyytinen, H. (2011). Computer-assisted remedial reading intervention for school beginners at risk for reading disability. *Child Development*, *82*(3), 1013-1028. https://doi.org/10.1111/j.1467-8624.2011.01580.x

Savage, R. S., Erten, O., Abrami, P., Hipps, G., Comaskey, E., & van Lierop, D. (2010). ABRACADABRA in the hands of teachers: The effectiveness of a web-based literacy intervention in grade 1 language arts programs. *Computers and Education*, *55*(2), 911-922. https://doi.org/10.1016/j.compedu.2010.04.002

Scammacca, N., Vaughn, S., Roberts, G., Wanzek, J., & Torgesen, J. K. (2007). *Extensive reading interventions in grades K-3: From research to practice*. https://files.eric.ed.gov/fulltext/ED521573.pdf

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-

fit measures. *Methods of Psychological Research Online*, *8*(2), 23-74. https://www.researchgate.net/publication/251060246

Segers, E., Damhuis, C. M. P., van de Sande, E., & Verhoeven, L. (2016). Role of executive functioning and home environment in early reading development. *Learning and Individual Differences*, *49*, 251-259. https://doi.org/10.1016/j.lindif.2016.07.004

Semel, E., Wiig, E. H., & Secord, W. A. (2003). *Clinical evaluation of language fundamentals, fourth edition (CELF-4)*. Pearson Assessment.

Share, D. L. (2008). On the anglocentricities of current reading research and practice: The perils of overreliance on an "outlier" orthography. *Psychological Bulletin*, *134*(4), 584–615. https://doi.org/10.1037/0033-2909.134.4.584

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods*, *7*(4), 422–445. https://doi.org/10.1037/1082-989X.7.4.422

Simmons, D. C., Coyne, M. D., Kwok, O. M., McDonagh, S., Harn, 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*, *13*(1), 7-14. https://doi.org/10.1111/j.1471-3802.2012.01262.x

Snowling, M. J., & Melby-Lervåg, M. (2016). Oral language deficits in familial dyslexia: A meta-analysis and review. *Psychological Bulletin*, *142*(2), 498-545. https://doi.org/10.1037/bul0000037

Stein, M. L., Berends, M., Fuchs, D., McMaster, K., Sáenz, L., Yen, L., Fuchs, L. S., & Compton, D. L. (2008). Scaling up an early reading program: Relationships among teacher support, fidelity of implementation, and student performance across different sites and years. *Educational Evaluation and Policy Analysis*, *30*(4), 368-388. https://doi.org/10.3102/0162373708322738

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

Storch, S. A., & Whitehurst, G. J. (2001). The role of family and home in the literacy development of children from low-income backgrounds. *New Directions for Child and Adolescent Development, 2001*(92), 53-72. https://doi.org/10.1002/cd.15

Streiner, D. L. (2002). The 2 "es" of research: Efficacy and effectiveness trials. *The Canadian Journal of Psychiatry*, *47*(6), 552-556. https://doi.org/10.1177/070674370204700607

Stuebing, K. K., Barth, A. E., Trahan, L. H., Reddy, R. R., Miciak, J., & Fletcher, J. M. (2015). Are child cognitive characteristics strong predictors of responses to intervention? A meta-analysis. *Review of Educational Research*, *85*(3), 395-429. https://doi.org/10.3102/0034654314555996

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

Swanson, E., Wanzek, J., Haring, C., Ciullo, S., & McCulley, L. (2013). Intervention fidelity in special and general education research journals. *Journal of Special Education*, *47*(1), 3-13. https://doi.org/10.1177/0022466911419516

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

Thomson, J. M., Foldnes, N., Uppstad, P. H., Njå, M., Solheim, O. J., & Lundetræ, K. (2020). Can children's instructional gameplay activity be used as a predictive indicator of reading skills? *Learning and Instruction*, *68*, 101348. https://doi.org/10.1016/j.learninstruc.2020.101348

Tomesen, M., Weekers, A., Hilte, 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. (2002). The Prevention of Reading Difficulties. *Journal of School Psychology*, *40*(1), 7-26. https://doi.org/10.1016/S0022-4405(01)00092-9

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*, *43*(1), 33-58. 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*, 1389-1401. https://doi.org/10.1007/s10802-015-0003-1

Torppa, M., Poikkeus, A. M., Laakso, M. L., Tolvanen, A., Leskinen, E., Leppänen, P. H. T., Puolakanaho, A., & Lyytinen, H. (2007). Modeling the early paths of

phonological awareness and factors supporting its development in children with and without familial risk of dyslexia. *Scientific Studies of Reading*, *11*(2), 73-103. https://doi.org/10.1080/10888430709336554

Tran, L., Sanchez, T., Arellano, B., & Swanson, H. L. (2011). A meta-analysis of the RTI literature for children at risk for reading disabilities. *Journal of Learning Disabilities*, *44*(3), 283–295. https://doi.org/10.1177/0022219410378447

Tremolada, M., Taverna, L., & Bonichini, S. (2019). Which factors influence attentional functions? Attention assessed by KITAP in 105 6-to-10-year-old children. Behavioral Sciences, 9(1), 7. https://doi.org/10.3390/bs9010007

Vadasy, P. F., & Sanders, E. A. (2009). Supplemental fluency intervention and determinants of reading outcomes. *Scientific Studies of Reading*, *13*(5), 383-425. https://doi.org/10.1080/10888430903162894

van Bergen, E., Bishop, D., van Zuijen, T., & de Jong, P. F. (2015). How does parental reading influence children's reading? A study of cognitive mediation. *Scientific Studies of Reading*, *19*(5), 325–339. https://doi.org/10.1080/10888438.2015.1050103

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 der Leij, A., & van Daal, V. H. P. (1999). Automatization aspects of dyslexia: Speed limitations in word identification, sensitivity to increasing task demands, and orthographic compensation. *Journal of Learning Disabilities*, *32*(5), 417-428. https://doi.org/10.1177/002221949903200507

van der Weijden, F. A., van den Boer, M., Zijlstra, A. H., & de Jong, P. F. (2024a). A school-based implementation of an early-literacy intervention: Relations among dosage, familial risk, parental education, and reading acquisition [Manuscript submitted for publication]. Department of Child Development and Education, University of Amsterdam.

van der Weijden, F. A., van den Boer, M., Zijlstra, A. H., van der Leij, A., Zijlstra, B. J. H., & de Jong, P. F. (2024b). Dosage explains individual differences in the outcomes of a prevention program for literacy problems [Manuscript submitted for publication]. Department of Child Development and Education, University of Amsterdam.

van der Weijden, F. A., van den Boer, M., Zijlstra, B. J. H., & de Jong, P. F. (2024c). Implementation takes time: Reduction of literacy problems in schools

implementing an early-literacy intervention. *Journal of Research on Educational Effectiveness*, 1–33 . http://dx.doi.org/10.1080/19345747.2024.2384365

van Dijk, W., Lane, H. B., & Gage, N. A. (2023). How do intervention studies measure the relation between implementation fidelity and students' reading outcomes? A systematic review. *The Elementary School Journal*, *124*(1), 56-84. https://doi.org/10.1086/725672

van Druenen, M., Scheltinga, F., Wentink, H., & Verhoeven, L. (2019). *Protocol preventie van leesproblemen groep 1 en 2 [Protocol prevention of reading problems kindergarten]* (2nd ed.). Expertise Centrum Nederlands.

van Ginkel, J. R., van der Ark, A. L., Sijtsma, K., & Vermunt, J. K. (2007). Two-way imputation: A bayesian method for estimating missing scores in tests and questionnaires, and an accurate approximation. *Computational Statistics and Data Analysis, 51*(8), 4013-4027. https://doi.org/10.1016/j.csda.2006.12.022

van Otterloo, S. G., van der Leij, A. Der, & Henrichs, L. F. (2008). Early home-based intervention in the Netherlands for children at familial risk of dyslexia. *Dyslexia*, *15*(3), 187-217. https://doi.org/10.1002/dys.376

van Til, A., Kamphuis, F., Keuning, J., Gijsel, M., Vloedraven, J., & de Wijs, A. (2018). Wetenschappelijke verantwoording LVS-toetsen DMT [Scientific justification LVS Three-Minute-Tests]. Cito.

van Uittert, A., Verhoeven, L., & Segers, E. (2022). Responsiveness to a game-based intervention to enhance reading efficiency in first graders. *Journal of Computer Assisted Learning*, *38*(1), 178-191. https://doi.org/10.1111/jcal.12599

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

Verhoeven, L. (1993). *Grafementoets. Toets voor auditieve synthese. Handleiding.* [Grafemetest. Test for auditory synthesis. Manual.] Cito.

Verhoeven, L. (2000). Components in early second language reading and spelling. *Scientific Studies of Reading*, *4*(4), 313-330. https://doi.org/10.1207/s1532799xssr0404_4

Verhoeven, L., & van Leeuwe, J. (2003). Ontwikkeling van decodeervaardigheid in het basisonderwijs [Development of recoding skill in primary education]. *Pedagogische Studiën*, *80*(4), 257–271. https://pedagogischestudien.nl/download?type=document&identifier=616546

Verhoeven, L., Voeten, M., van Setten, E., & Segers, E. (2020). Computer-supported early literacy intervention effects in preschool and kindergarten: A meta-analysis. *Educational Research Review*, *30*, 100325. https://doi.org/10.1016/j.edurev.2020.100325

Vernon-Feagans, L., Kainz, K., Amendum, S., Ginsberg, M., Wood, T., & Bock, A. (2012). Targeted reading intervention: A coaching model to help classroom teachers with struggling readers. *Learning Disability Quarterly*, *35*(2), 102-114. https://doi.org/10.1177/0731948711434048

Wagner, R. K., Torgesen, J. K., Laughon, P., Simmons, K., & Rashotte, C. A. (1993). Development of young readers' phonological processing abilities. *Journal of Educational Psychology, 85*(1), 83–103. https://doi.org/10.1037/0022-0663.85.1.83

Wang, Y., Deng, C., & Yang, X. (2016). Family economic status and parental involvement: Influences of parental expectation and perceived barriers. *School Psychology International*, *37*(5), 536-553. https://doi.org/10.1177/0143034316667646

Wang, Y., Huebner, E. S., & Tian, L. (2021). Parent-child cohesion, self-esteem, and academic achievement: The longitudinal relations among elementary school students. *Learning and Instruction*, *73*, 101467. https://doi.org/10.1016/j.learninstruc.2021.101467

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. (2008). Response to varying amounts of time in reading intervention for students with low response to intervention. *Journal of Learning Disabilities*, *41*(2), 126–142. https://doi.org/10.1177/0022219407313426

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

Wanzek, J., Vaughn, S., Scammacca, N., Gatlin, B., Walker, M. A., & Capin, P. (2016). Meta-analyses of the effects of tier 2 type reading interventions in grades K-3. *Educational Psychology Review*, *28*, 551-576. https://doi.org/10.1007/s10648-015-9321-7

Wegener, S., Wang, H. C., Beyersmann, E., Nation, K., Colenbrander, D., & Castles, A. (2022). The effects of spacing and massing on children's orthographic

learning. *Journal of Experimental Child Psychology*, *214*, 105309. https://doi.org/10.1016/j.jecp.2021.105309

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*, 453-469. https://doi.org/10.1146/annurev-publhealth-040617-013507

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, *10*, Article 3087. https://doi.org/10.3389/fpsyg.2019.03087

Wolgemuth, J. R., Abrami, P. C., Helmer, J., Savage, R., Harper, H., & Lea, T. (2014). Examining the impact of ABRACADABRA on early literacy in Northern Australia: An implementation fidelity analysis. *Journal of Educational Research*, *107*(4), 229-311. https://doi.org/10.1080/00220671.2013.823369

Wolgemuth, J. R., Savage, R., Helmer, J., Harper, H., Lea, T., Abrami, P. C., Kirby, A., Chalkiti, K, Morris, K., Carapentis, J., & Louden, W. (2013). ABRACADABRA aids Indigenous and non-Indigenous early literacy in Australia: Evidence from a multisite randomized controlled trial. *Computers and Education*, *67*, 250-264. https://doi.org/10.1016/j.compedu.2013.04.002

Wood, F. B., Hill, D. F., Meyer, M. S., & Lynn Flowers, D. (2005). Predictive assessment of reading. *Annals of Dyslexia*, *55*, 193-216. https://doi.org/10.1007/s11881-005-0011-x

Yuan, K. H., & Bentler, P. M. (2000). 5. Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, *30*(1), 165-200. https://doi.org/10.1111/0081-1750.00078

Zeleke, 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]. UvA-DARE (Digital Academic Repository). 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

Summary

The Prevention of Reading Problems

Reading problems can lead to academic failure and lowered academic self-esteem. Reading problems in first grade tend to persist into adolescence, even though most schools provide support when children lag behind. Research showed that even with intensive remedial instruction children have difficulties to overcome their reading difficulties, especially in reading fluency. Therefore, it seems better to prevent reading problems and provide extra support in kindergarten or even preschool to children at risk of reading problems. Early-literacy interventions targeting the main precursors of reading (letter knowledge and phonological awareness) generally show promising effects on later reading skills, and most studies show larger effects than remedial reading interventions. When such prevention programs are continued for two or three years, long-term effects that are still noticeable beyond second grade have also been established.

An example of an extensive program for the prevention of reading problems is the Dutch early-literacy intervention *Build!* (in Dutch: *Bouw!*). The intervention starts in kindergarten, before formal reading instruction begins. It is provided to children who are considered at risk of reading problems. The intervention continues for two years, until the middle of second grade. It is computer-based and covers the precursors of reading, letter knowledge and phonological awareness, as well as word reading accuracy and fluency. The effectiveness of *Build!* has been established in three RCTs. The intervention is currently being implemented in 80% of Dutch primary schools.

The effects of early-literacy interventions like *Build!* have mostly been evaluated in randomized controlled trials (RCTs). The implementation of the intervention is often guided by researchers. Relatively few interventions have subsequently been evaluated when the implementation of the intervention is the schools' own responsibility.

This Dissertation

In this dissertation, I evaluated the effects of the early-literacy intervention *Build!*, which was implemented on a large scale without researcher control. I investigated several factors and their relation with intervention outcomes. As mentioned in the title of this dissertation, these factors were: 1) treatment integrity, that is the extent

to which the intervention is implemented as intended (referred to as 'weeks of practice') and 2) the number of years schools' implemented the intervention (referred to as 'years of experience'). Additionally, I related two other factors to intervention outcomes: 3) family characteristics, specifically familial risk for dyslexia[3] and parental education and 4) reach, that is the proportion of the target population that participates in the intervention. This was done in three studies and in additional analyses presented in the general discussion.

Treatment Integrity and Family Characteristics

The first study (Chapter 2) was focused on treatment integrity, i.e. the extent to which the intervention is implemented as intended. Treatment integrity might be key to the success of interventions when they are implemented by schools, but only a few studies have examined it. In this study, it was investigated whether variations in treatment integrity were related to outcomes of the early-literacy intervention *Build!* in kindergarten. In particular, the study focused on one aspect of treatment integrity: dosage, i.e. the amount of practice during the intervention. Three aspects of dosage were distinguished: frequency of intervention sessions, session length, and the number of intervention weeks. The first goal of the study was to investigate–on a weekly basis–the effect of two aspects of dosage, frequency and length of intervention sessions, on the progress within the intervention. The effect of dosage on progress was examined with multilevel modeling. The second goal was to examine–at the child level–the relation between dosage and intervention outcomes in kindergarten and the potential mediating role of progress within the intervention. This was examined with a path model. In addition to frequency and length of intervention sessions, this model also included duration, i.e. the number of intervention weeks, as an indicator of treatment integrity.

The results of the first study showed that, on a weekly basis, the child's progress within the intervention varied strongly. About half of the variation in progress was explained by the number of sessions and the session length. Each extra intervention session and every five minutes of additional practice led to more progress within the intervention. The effect of the number of sessions was stronger than of session length. At the child level, I found that the three aspects of dosage (i.e. frequency, length, and duration) were related to intervention outcomes, after controlling for pre-test measures, and that these relations were fully mediated by the progress of children within the intervention. Similar to the week level, the number of sessions

[3] Familial risk for dyslexia refers to the existence of reading problems in the family.

had the largest effect on progress, followed by the number of intervention weeks and the session length. Dosage had stronger associations with gains in letter knowledge than in phonological awareness. The results of this study support the importance of dosage when interventions are implemented at a larger scale. It can explain why some children benefit more from the intervention than others. The study adds to current knowledge by showing the effects of different aspects of dosage, of which frequency of intervention sessions was found to matter most.

The second study (Chapter 3) followed up the first study by examining the progress of children for a longer period of time. While in the first study I studied in detail the association between dosage and the development of preliteracy skills during kindergarten, the second study followed the children until the middle of Grade 2, and included a larger sample (369 instead of 226 children). Dosage was measured as a single factor (i.e. the number of hours spent on the intervention). The first goal of this study was to examine whether dosage and progress were associated with preliteracy skills (end of kindergarten), word reading accuracy (mid- and end-Grade 1), and word reading fluency (mid- and end-Grade1). Three intervention periods were distinguished: 1) the second half of the second kindergarten year (similar to the first study), 2) the first half of first grade, and 3) the second half of first grade. In each period, I examined whether dosage and progress explained additional variance in intervention outcomes, after the initial level of these skills was controlled. The second aim was to investigate whether the intervention was equally effective for all children, by examining whether two family characteristics, parental education and familial risk for dyslexia[4], were related to children's outcomes directly and indirectly through dosage and progress within the intervention. Both research questions were analyzed with path models.

The results of the second study showed that a higher dose was associated with more progress within the intervention, and, in turn with more letter knowledge and better phonological skills at the end of kindergarten, better word reading accuracy at the beginning of first grade, and better word reading fluency, but not better word reading accuracy in the middle of first grade. After the middle of first grade, there were no additional effects of dosage on reading development. *T*he educational level of the parents was not related to dosage, nor to children's progress within the intervention or literacy outcomes. Familial risk was not related to dosage, but it was negatively related to progress within the intervention, preliteracy skills and also to later reading

[4] Children were considered to have a familial risk for dyslexia when at least one of the parents reported to have (a diagnosis of) dyslexia.

skills, even when the effects on progress and pre-literacy skills were taken into account. In line with the first study, these results emphasize the importance of dosage in school-based implementations and suggest that the relation between dosage en children's outcomes in computer-based interventions is mediated by progress within the intervention. Consequently, progress through the intervention could be used to adjust the dose to an individual child's needs to optimize intervention outcomes. The findings also showed that children with familial risk for dyslexia reach lower literacy levels, partly because they progressed more slowly through the intervention. In principle, they need more intervention sessions to reach the same level as the children without family risk. However, in the current school-based implementation of *Build!* they were not provided with additional practice. Schools might not be aware of this special need yet, and could improve intervention outcomes by providing children with familial risk with more practice than their classmates without such risk.

Intervention Effectiveness and Schools' Experience With the Intervention

The third study (Chapter 4) focused on the effectiveness of the large-scale implementation of *Build!*. Effects at the school level were evaluated, as schools and policy makers might want to know whether the implementation of the intervention leads to a reduction of reading problems in schools. In large-scale studies, effects might not be visible right after the intervention is implemented, but only after a few years. It is no small feat to implement an early-literacy intervention at school. Implementing an intervention takes time, resources, and leadership, and it might take more than one year to implement the intervention as intended. In this quasi-experimental study with a sample of 207 schools, reading outcomes of schools that had implemented the intervention were compared to those that had not implemented the intervention during a period of six school years (2014-2015 to 2019-2020). I evaluated whether there was a reduction of reading problems and/or an increase in the average reading ability in schools from the moment that they had implemented the intervention. Moreover, it was investigated whether these effects possibly transferred to spelling and reading comprehension. I also included mathematics, a skill unrelated to the intervention, on which I did not expect an effect. I determined effects during the intervention (mid- and end-Grade 1), effects at post-test (mid-Grade 2), and effects at follow-up (end-Grade 2 and mid-Grade 3, i.e. 0.5 to 1 year after the intervention should be finished). Furthermore, it was examined whether the effects became stronger when schools used the intervention for a longer time. These questions were

answered with difference-in-difference models, which is a pretest-posttest quasi-experimental design with an intervention and a control group, which we extended by the addition of multiple pre- and posttests.

The results of the third study showed that there was no change in literacy skills after the first year that schools had been using *Build!*. However, after schools implemented the intervention for two years, there were small overall decreases in the percentage of children with difficulties in reading fluency, spelling, and reading comprehension, while this was not shown for math. Moreover, the average reading fluency and spelling ability of children begun to increase after the implementation of *Build!*. Changes were small (only a few percent or scale points), but only small effects could be expected on a large scale, due to heterogeneous populations, a large variety in treatment integrity, the small target group of *Build!* (the children with the 25-30% lowest scores on preliteracy skills), and challenges regarding the selection of children at risk of reading problems. Findings suggest that it takes time to make an intervention work in schools. This might be easily overlooked by researchers. It has been suggested that it is very hard to find effects in large-scale implementation studies. Our findings suggest that effects might not be found immediately after implementation, but after schools have had several years of experience with the intervention.

The Reach of the Intervention

In the *General Discussion* (Chapter 5) I focused on reach, that is the proportion of the target population that participates in the intervention. In particular, three challenges that schools face when selecting children for the intervention were addressed. First, the selection of children who need the intervention, that is the children who will develop reading problems, can only be based on preliteracy skills and is thereby never perfect. Inevitably, some children are identified as 'not at risk' while they will develop reading problems (false negatives) and some children are identified as 'at risk' while they will develop sufficient reading skills (false positives). Second, there might be more children in need of the intervention than the number of children that can be provided with the intervention within the time and recourses that schools have available. Third, schools may want to identify false positives during the intervention and to know when they can stop the intervention, so that schools can spent time on those children who really need the intervention. To these ends I evaluated a selection procedure that was used by part of the schools in the first and second study, which was developed by researchers and schools to be brief and teacher-friendly. The procedure included two screening waves, one in October and one in January of the

second kindergarten year, in which letter knowledge and phonological awareness were assessed.

Findings presented in the general discussion showed that the selection procedure used by part of the schools in the first and second study of this dissertation, produced a similar percentage of false negatives as found in other screening procedures for the identification of kindergartners at risk of reading problems. Thereby, the proposed selection procedure is a good starting point for Dutch schools to identify children at risk of reading problems. The screening procedure could be improved with an additional screening wave after a few months of formal reading instruction, so that schools can identify false negatives, that are children who were not selected in kindergarten but do show reading difficulties after a few months.

Results also showed that a fair number of children did not receive the intervention although they were eligible. This decision seemed to be based in part on age and difficulty with rapid naming. The Eligible-NoBuild! group was less at risk for reading problems than the Eligible-Build! group. A smaller percentage of this group developed reading problems, even when they did not receive the intervention. This decision had both positive and negative consequences for reach: some children were prevented from unnecessary intervention, while others were not provided with the intervention, while needing it. There was also a fair number of children who received the intervention although they were not eligible. This decision seemed to be based in part on difficulty with rapid naming, several child characteristics (being a boy or having Dutch as a second language) and/or several family characteristics (home literacy environment, parental educational level, or family risk for dyslexia). The NotEligible-Build! group was clearly more at risk for reading problems than the NotEligible-No-Build! group. A larger percentage of this group developed reading problems even though they had received the intervention. This decisions also had positive and negative consequences for reach: some not eligible children were prevented from (more) severe reading problems by providing them with the intervention, while others were provided with unnecessary intervention. In preventing reading problems, schools may want to play safe. Therefore, they could best provide all eligible children with the intervention and additionally select some not eligible children who they consider to be in need of the intervention. Thereby, they prevent reading problems in as many children as possible.

However, they also provide children with unnecessary intervention, which constrains schools' time and resources. Therefore, schools may want to identify false positives during the intervention, that are children who do not need the intervention but are provided with it. In our sample, schools measured reading skills mid-Grade 1

and stopped the intervention for part of the children, which were mostly good and above-average readers[5]. Our findings suggest that only one out of 46 good and above-average readers developed reading problems afterwards[6]. I matched the children who stopped to children who continued and compared their reading development. Findings showed that stopping or continuing the intervention after mid-Grade 1 did not make a difference. Note that it were mainly better readers who stopped, so the analysis also included relatively better readers who continued. Poorer readers may thus still benefit from continuing the intervention. To create more time for children who really need the intervention, our findings suggest that schools can identify false positives by measuring reading skills mid-Grade 1 and that children with sufficient[3] reading skills who stop the intervention do not develop reading problems afterwards.

Conclusion

This dissertation emphasized the importance of studying interventions not only in researcher-led RCTs, but also when they are implemented on a large scale, under schools' own responsibility. Findings show that effects of evidence-based interventions at the school level might only show up after schools have implemented the intervention for several years, that these effects may be small, and that intervention outcomes can be affected by many factors, such as the amount of intervention practice, the extent to which schools succeed in selecting the right children for the intervention, and whether or not there are learning problems in the family. In educational research, few interventions have been studied on a large scale, while in health research this is common practice. This dissertation presents both methods and initial results on how to study the effectiveness of an intervention that is implemented on a large scale without researcher contol, and factors influencing this effectiveness. More large-scale studies on the effectiveness of interventions in natural school settings are needed to provide schools with information on how to make a success of the implementation of evidence-based interventions. This dissertation has revealed two factors that were related and may contribute to this success: providing children with *weeks of practice* and having multiple *years of experience* with the intervention.

---

[5] That are children with A or B scores on the Three Minute Test mid-Grade 1 (i.e. scoring in or above the 51st percentile on Word Reading Fluency)

[6] Reading problems was defined as having a D or E score on the Three Minute Test mid-Grade 1 (i.e. scoring in or below the 25th percentile on Word Reading Fluency)


acer

Samenvatting

Preventie van Leesproblemen

Als kinderen moeite hebben met technisch lezen, kan dit leiden tot lage schoolprestaties en een verminderd zelfvertrouwen in hun schoolse vaardigheden. Leesproblemen in groep 3 blijven vaak voortbestaan tot in de puberteit, ondanks dat scholen hulp bieden aan kinderen die moeite hebben met technisch lezen. Onderzoek toont aan dat veel kinderen met leesproblemen, zelfs na intensieve remediëring, moeite blijven houden met lezen, met name met vloeiend lezen. In plaats van hulp te bieden zodra er leesproblemen zijn, lijkt het daarom beter om te proberen leesproblemen te voorkomen door kinderen met een risico op leesproblemen op te sporen in groep 2 (of zelfs eerder) en dan al hulp te bieden. Vroege leesinterventies die zich richten op de voorlopers van het leren lezen (letterkennis[1] en klankbewustzijn[2]) laten over het algemeen veelbelovende effecten zien op latere leesvaardigheden, grotere effecten dan remediërende interventies. Wanneer dergelijke preventieve programma's niet stoppen na groep 2, maar twee of drie jaar worden voortgezet, worden er ook langetermijneffecten gevonden die nog zichtbaar zijn na groep 4.

Een voorbeeld van een meerjarig programma dat zich richt op de preventie van leesproblemen is de vroege leesinterventie *Bouw!*. De interventie begint in groep 2, voordat kinderen leren lezen in de klas. De interventie is bedoeld voor kinderen met een risico op leesproblemen, dat wil zeggen voor kinderen die moeite hebben met de voorlopers van het leren lezen, zoals letterkennis en klankbewustzijn. De interventie duurt twee jaar en eindigt halverwege groep 4. Het is een computerprogramma dat zich niet alleen richt op de voorlopers van het leren lezen, maar ook op leesaccuratesse en leesvloeiendheid. Drie gerandomiseerde studies met controlegroep (randomized controlled trials, RCTs) hebben de effectiviteit van *Bouw!* aangetoond. De interventie wordt momenteel ingezet op 80% van de basisscholen in Nederland.

[1] Letterkennis is het aantal geschreven letters dat een kind aan de bijbehorende klank kan koppelen.

[2] Klankbewustzijn is het vermogen klanken in gesproken taal te herkennen en te manipuleren, zoals het samenvoegen van klanken tot een woord (/k/ /i/ /p/ - kip).

Dit Proefschrift

De effecten van vroege leesinterventies, zoals *Bouw!,* zijn tot nu toe vaak onderzocht met RCTs. Hierbij wordt de implementatie van de interventie vaak begeleid door onderzoekers. Relatief weinig interventies zijn vervolgens onderzocht in een natuurlijke situatie, waarin scholen zelf verantwoordelijk zijn voor de implementatie van de interventie. In dit proefschrift heb ik de effecten van de vroege leesinterventie *Bouw!* onderzocht, die op grote schaal wordt geïmplementeerd op Nederlandse basisscholen. De hoofdvragen van dit proefschrift zijn: 1) Hangt oefentijd, een dimensie van interventietrouw, samen met de interventie-uitkomsten, wanneer een interventie op grote schaal door scholen wordt geïmplementeerd? 2) Zijn een familiair risico op dyslexie[3] en het opleidingsniveau van de ouders gerelateerd aan interventietrouw en de uitkomsten van vroege leesinterventies? 3) Wat zijn de effecten van vroege leesinterventies op schoolniveau en worden de effecten groter naarmate scholen meer ervaring hebben met de interventie? en 4) Wat is het bereik van de interventie, met andere woorden in welke mate krijgt de groep voor wie de interventie is bedoeld daadwerkelijk de interventie?

Interventietrouw en Gezinskenmerken

De eerste studie (Hoofdstuk 2) richtte zich op interventietrouw, dat is de mate waarin de interventie wordt ingezet zoals bedoeld. Interventietrouw is mogelijk de sleutel tot het succes van de interventie in de praktijk, maar slechts een paar studies hebben dit onderzocht. In deze studie werd onderzocht of natuurlijke variatie in interventietrouw samenhing met de uitkomsten van de vroege leesinterventie *Bouw!* in groep 2. De studie richtte zich op één aspect van interventietrouw: oefentijd. Er werden drie aspecten van oefentijd onderscheiden: het aantal sessies per week, de sessieduur en het aantal interventieweken. De samenhang tussen de oefentijd en de voortgang binnen de interventie werd onderzocht op twee niveaus: op weekniveau en op kindniveau. Op weekniveau werden twee aspecten van oefentijd onderscheiden: het aantal sessies per week en de duur van een sessie. Er werd onderzocht hoeveel voortgang kinderen maakten binnen de interventie in weken dat ze één keer oefenden, twee keer, drie keer etc. en in weken dat ze gemiddeld tien tot vijftien minuten oefenden per sessie, vijftien tot twintig minuten, meer dan twintig minuten etc. Voortgang binnen de interventie werd gemeten met het aantal nieuwe lessen dat een kind binnen de interventie voltooide tussen midden en eind groep 2. Op kindniveau werd

[3] Een familiair risico op dyslexie verwijst naar leesproblemen in het gezin.

onderzocht of kinderen meer groei lieten zien in de voorlopers van het leren lezen (letterkennis en klankbewustzijn) tussen midden en eind groep 2 naarmate ze meer oefenden en in hoeverre dit verband kon worden toegeschreven aan de grotere voortgang binnen de interventie. Hierbij werden alle drie de aspecten van oefentijd meegenomen: het gemiddeld aantal sessies per week, de gemiddelde duur per sessie en het aantal interventieweken. Voor de analyses op weekniveau werd *multilevel modeling* gebruikt, voor de analyses op kindniveau werd *een* padmodel gebruikt.

De resultaten op weekniveau tonen aan dat de voortgang die kinderen maakten binnen de interventie sterk varieerde per week, doordat het gemiddeld aantal sessies per week en de gemiddelde sessieduur per week verschilde. De frequentie van de sessies had een sterkere samenhang met de voortgang binnen de interventie dan de sessieduur. Dit betekent dat veel korte sessies efficiënter waren dan weinig lange sessies. Het bleek dat kinderen in groep 2 het efficiëntst oefenden in weken met drie sessies van elk tien tot vijftien minuten en kinderen in groep 3 in weken met vier sessies van elk tien tot vijftien minuten. Meer oefenen loonde, maar de meeropbrengst van extra oefenen nam wel af naarmate de hoeveelheid oefening toenam.

Op kindniveau vond ik dat kinderen die meer hadden geoefend (meer sessies per week, langere sessies *of* meer interventieweken) meer waren gegroeid in letterkennis en klankbewustzijn tussen midden en eind groep 2. Dit verband was volledig toe te schrijven aan het aantal nieuwe lessen dat ze binnen de interventie hadden voltooid. Net als op weekniveau, hing het aantal sessies per week sterker samen met de voortgang binnen de interventie dan de andere aspecten van oefentijd, wat wederom laat zien dat veel korte sessies per week efficiënter was dan weinig lange sessies. Voortgang binnen de interventie was niet alleen gerelateerd aan de oefentijd, maar ook aan het aantal herhaallessen: kinderen die meer moesten herhalen binnen de interventie, maakten minder voortgang, terwijl ze evenveel oefenden. Verder lieten de resultaten zien dat oefentijd en voortgang sterker samenhingen met de vooruitgang in letterkennis dan met de vooruitgang in klankbewustzijn. De interventie leek dus een groter effect te hebben op letterkennis dan op klankbewustzijn. De bevindingen van deze studie ondersteunen dat oefentijd belangrijk is, wanneer de implementatie van de interventie de eigen verantwoordelijkheid van de scholen is. Omdat voortgang binnen de interventie niet voor elk kind gelijk is, kunnen scholen de oefentijd van een individueel kind het best afstemmen op de voortgang die het kind binnen de interventie maakt. De studie heeft tevens nieuwe wetenschappelijke kennis opgeleverd door aan te tonen dat er een verschil is tussen oefentijd en voortgang binnen de interventie, namelijk dat voortgang negatief samenhing met het aantal herhaallessen en daarmee mogelijk met het leervermogen van de leerling, en oefentijd niet. Ook

laten de bevindingen zien dat het belangrijkste aspect van oefentijd het aantal sessies per week lijkt te zijn.

De tweede studie (Hoofdstuk 3) was een vervolg op de eerste studie: opnieuw werd de relatie tussen oefentijd, voortgang binnen de interventie en interventie-uitkomsten onderzocht, maar dan gedurende een langere periode. In de eerste studie werden kinderen gevolgd van midden tot eind groep 2 met letterkennis en klankbewustzijn als uitkomstmaten en werd onderscheid gemaakt tussen drie aspecten van oefentijd. In de tweede studie werden kinderen langer gevolgd, van midden groep 2 tot en met midden groep 4, met leesvaardigheden als uitkomstmaat in groep 3 en 4. De tweede studie had ook een grotere steekproef dan de eerste studie (369 in plaats van 226 kinderen). Verder werden de drie aspecten van oefentijd samengenomen, zijnde het aantal uur dat een kind had geoefend met *Bouw!* gedurende een interventieperiode. Daarbij werden drie interventieperioden onderscheiden: 1) de tweede helft van groep 2 (net als in de eerste studie), 2) de eerste helft van groep 3 en 3) de tweede helft van groep 3. Het eerste doel van de tweede studie was om in elke periode te onderzoeken of oefentijd samenhing met voortgang binnen de interventie (het aantal nieuwe *Bouw!*-lessen dat een kind voltooide binnen een periode) en of de voortgang binnen de interventie vervolgens samenhing met letterkennis en klankbewustzijn (eind groep 2), woordleesaccuratesse (midden en eind groep 3) en woordleesvloeiendheid (midden en eind groep 3). Hierbij werden de (voorbereidende) leesvaardigheden aan het begin van een periode steeds meegenomen, zodat kon worden bekeken of kinderen die meer oefenden ook sterker vooruitgingen gedurende een bepaalde interventieperiode. Het tweede doel was om in kaart te brengen of de interventie even effectief was voor alle kinderen. Dit werd gedaan door te onderzoeken of twee gezinskenmerken, het type opleiding van de ouders en een familiair risico op dyslexie, verband hielden met de oefentijd, de voortgang binnen de interventie en de (voorbereidende) leesvaardigheden van de kinderen. Beide onderzoeksvragen werden geanalyseerd met *padmodellen*.

De resultaten van de tweede studie laten zien dat kinderen die meer oefenden ook meer vooruitgingen binnen de interventie en daardoor meer groeiden in, respectievelijk letterkennis en klankbewustzijn in groep 2, woordleesaccuratesse begin groep 3 en midden groep 3 en in woordleesvloeiendheid begin groep 3 en midden groep 3. Tussen midden en eind groep 3 groeiden de kinderen die meer oefenden, niet sneller in woordleesaccuratesse en woordleesvloeiendheid.

Wat betreft de gezinskenmerken, werd gevonden dat kinderen van ouders met een meer theoretische opleiding evenveel oefenden, even snel vooruitgingen binnen de interventie en vergelijkbare (voorbereidende) leesvaardigheden hadden als

kinderen van ouders met een meer praktische opleiding. Hierbij is het belangrijk om op te merken dat de steekproef in verhouding minder theoretisch opgeleide ouders bevatte dan de nationale populatie. Kinderen met een familiair risico op dyslexie oefenden evenveel als kinderen zonder een familiair risico. Ook maakten ze minder voortgang binnen de interventie in de tweede helft van groep 2 en de eerste helft van groep 3. Daarnaast hadden ze een kleinere letterkennis eind groep 2, wat een negatief effect had op hun latere leesvaardigheden. Echter kon dit hun zwakkere leesaccuratesse en leesvloeiendheid midden groep 3 niet volledig verklaren. Ze hadden ook moeite met het leren lezen zelf. Verder konden de lagere leesvaardigheden midden in groep 3 ook verklaard worden, doordat ze minder voortgang maakten binnen de interventie, ondanks dat ze evenveel oefenden. Dit laatste zou voorkomen kunnen worden door kinderen met een familiair risico meer oefening aan te bieden.

Net als in de eerste studie, laten de resultaten van de tweede studie zien dat oefentijd belangrijk is wanneer interventies in de praktijk worden ingezet. Oefentijd hangt niet alleen samen met interventie-uitkomsten in groep 2, maar ook in de eerste helft van groep 3. Opnieuw was dit verband volledig toe te schrijven aan de voortgang die kinderen maakten binnen de interventie. De tweede studie laat wederom zien dat voortgang binnen de interventie verschilt van oefentijd, namelijk dat voortgang gerelateerd was aan een familiair risico op dyslexie en aan de voorlopers van het leren lezen bij aanvang van de interventie, en oefentijd niet. Dit impliceert dat alleen voortgang gerelateerd lijkt te zijn aan het leervermogen van een leerling. Niet alle kinderen gaan dus even snel vooruit binnen de interventie. De oefentijd van kinderen kan dus het best afgestemd worden op hun voortgang binnen de interventie.

Effecten van de Interventie en Ervaring van Scholen met de Interventie

De derde studie (Hoofdstuk 4) was gericht op de effectiviteit van *Bouw!* in de praktijk. Op schoolniveau werden de effecten van *Bouw!* onderzocht, zodat in kaart kon worden gebracht of de implementatie van *Bouw!* leidt tot de vermindering van leesproblemen op scholen. Dit kan waardevolle informatie zijn voor scholen, schoolbestuurders en beleidsmakers. Hierbij werd verondersteld dat de effecten van de implementatie van *Bouw!* wellicht niet meteen zichtbaar zijn vanaf het moment dat het interventie wordt geïmplementeerd, maar pas als scholen een aantal jaren ervaring hebben met de interventie. Het is immers geen kleinigheid om een vroege leesinterventie te implementeren op school. Implementatie kost tijd en geld en er is leiderschap voor nodig. Mogelijk duurt het langer dan één jaar om een interventie te implementeren zoals deze is bedoeld.

In deze quasi-experimentele studie met een steekproef van 207 scholen werden leesuitkomsten vergeleken van scholen die *Bouw!* hadden geïmplementeerd tijdens een periode van zes jaar (2014-2015 t/m 2019-2020) en scholen die dat niet hadden gedaan. Er werd onderzocht of binnen de school het percentage kinderen met leesproblemen afnam en de gemiddelde scores op technisch lezen toenam in de jaren nadat de scholen met *Bouw!* waren gestart. Daarnaast werd onderzocht of er transfereffecten waren op spelling en begrijpend lezen. Ook rekenen werd meegenomen in het onderzoek, een vaardigheid die ongerelateerd is aan de interventie en waarop er geen effect werd verwacht. Wanneer er geen verschil zou zijn tussen de interventiescholen en controlescholen in rekenen, maar wel in de andere vaardigheden, zou dit de waarschijnlijkheid vergroten dat een vermindering van leesproblemen toe te schrijven is aan de implementatie van Bouw! en niet aan een schoolbrede of landelijke verandering die leidde tot een algehele verbetering in schoolprestaties. Alle vaardigheden werden gemeten tijdens de interventie (midden en eind groep 3), na afloop van de interventie (midden groep 4) en een half en één jaar nadat de interventie was afgerond (eind groep 4 en midden groep 5). Daarnaast werd onderzocht of de effecten sterker werden met elk jaar dat scholen met de interventie werkten. Deze vragen werden beantwoord met een *difference-in-difference model*, een quasi-experimenteel design met een voor- en nameting en een interventie- en controlegroep. In deze studie was sprake van meerdere voor- en nametingen, zodat gekeken kon worden naar al dan niet veranderende trends in scores.

De resultaten tonen aan dat, voorafgaand aan de implementatie van *Bouw!*, het percentage kinderen met problemen in technisch lezen en begrijpend lezen op interventie- en controlescholen elk schooljaar toenam met 1 tot 2%. Het percentage kinderen met spellingsproblemen bleef gelijk. Dit veranderde niet, nadat scholen één jaar met *Bouw!* hadden gewerkt. Maar nadat scholen twee jaar met *Bouw!* hadden gewerkt, stabiliseerde het percentage kinderen met problemen in technisch lezen en begrijpend lezen op de interventie-scholen of begon het zelfs af te nemen met 1% per schooljaar. Het percentage kinderen met spellingsproblemen begon af te nemen met 1 tot 3% per schooljaar. Vergelijkbare resultaten werden gevonden voor de gemiddelde scores op technisch lezen en spelling binnen de school. Voorafgaand aan de implementatie van *Bouw!*, waren de gemiddelde scores op technisch lezen en spelling op de interventiescholen stabiel over de schooljaren heen. Dit veranderde niet, nadat scholen één jaar met *Bouw!* hadden gewerkt. Maar nadat scholen twee jaar met *Bouw!* hadden gewerkt, begonnen de gemiddelde scores op technisch lezen en spelling elk schooljaar toe te nemen. Zo'n verandering werd niet gevonden bij begrijpend lezen. Ook waren er geen veranderingen in rekenen, nadat scholen twee jaar met *Bouw!*

hadden gewerkt, noch in het percentage kinderen met *reken*problemen noch in de gemiddelde scores binnen de school. Dit betekent dat de veranderingen niet te wijten zijn aan een verandering die leidde tot een algehele verbetering van het onderwijs. Waarschijnlijker is dat de veranderingen komen door *Bouw!*, dan wel een verbetering van het taalonderwijs in het algemeen, die mogelijk in gang is gezet door *Bouw!*. Kortom, de gevonden verbeteringen in technisch lezen, spelling en begrijpend lezen binnen de school zijn niet met zekerheid toe te schrijven aan de implementatie van *Bouw!*, maar ze vonden tegelijkertijd plaats met de implementatie van *Bouw!*.

De gevonden veranderingen in technisch lezen, spelling en begrijpend lezen nadat scholen twee jaar ervaring hadden met *Bouw!*, waren klein, maar in de praktijk zijn alleen kleine effecten van *Bouw!* te verwachten om verschillende redenen, zoals verschillen tussen scholen in interventietrouw en omdat *Bouw!* wordt aangeboden aan slechts 25-30% van de kinderen. De resultaten van dit onderzoek suggereren dat het tijd kan kosten voordat een interventie, die in eerder kleinschalige RCTs effectief bleek, ook in de praktijk resultaat heeft. Dit wordt door onderzoekers mogelijk over het hoofd gezien. Er worden vaak kleine interventie-effecten gevonden met grootschalige implementatiestudies. Volgens de bevindingen van dit proefschrift zou een verklaring hiervoor kunnen zijn dat effecten mogelijk niet direct na implementatie optreden, maar pas nadat scholen enkele jaren ervaring hebben met de interventie.

Bereik van de Interventie

In de *Algemene Discussie* (Hoofdstuk 5) staat het bereik van de interventie centraal: de mate waarin de groep voor wie de interventie is bedoeld daadwerkelijk de interventie krijgt. Bij het identificeren van deze groep kinderen (bij *Bouw!*: kinderen die later leesproblemen zullen ontwikkelen) staan scholen voor verschillende uitdagingen. Drie van deze uitdagingen komen aan bod. De eerste is om in groep 2 te voorspellen welke kinderen later leesproblemen zullen ontwikkelen. Omdat kinderen in groep 2 over het algemeen nog niet kunnen lezen, kunnen eventuele leesproblemen alleen worden voorspeld op basis van voorbereidende leesvaardigheden, zoals letterkennis en klankbewustzijn. Hierbij is het onvermijdelijk dat sommige kinderen worden aangemerkt als 'heeft geen risico op leesproblemen' terwijl ze later wel leesproblemen ontwikkelen (vals negatieven), en dat andere kinderen worden aangemerkt als 'heeft een risico op leesproblemen' terwijl ze later voldoende leesvaardigheden ontwikkelen (vals positieven). De tweede uitdaging is dat het aantal kinderen dat in aanmerking komt voor de interventie op sommige scholen mogelijk groter is dan het aantal kinderen dat de interventie kan krijgen binnen de tijd en het budget van de school.

De derde uitdaging is de opsporing van vals positieven tijdens de interventie, dat zijn kinderen die voor de interventie in aanmerking komen, maar de interventie in werkelijkheid niet nodig hebben, en de daaraan gekoppelde beslissing of deze kinderen kunnen stoppen met de interventie om tijd te creëren voor de kinderen die de interventie echt nodig hebben.

In het kader van deze drie uitdagingen werd een selectieprocedure geëvalueerd die werd gebruikt door een deel van de scholen in de eerste en tweede studie. Deze was bedoeld als korte procedure en werd ontworpen door onderzoekers en scholen samen. De procedure bevatte twee screeningsmomenten waarbij letterkennis en klankbewustzijn werden getoetst: in oktober en in januari van groep 2. Tussen de twee screeningsmomenten in gaven de scholen de 30% zwakste leerlingen extra instructie in kleine groepjes. Kinderen die bij de tweede screening nog steeds zwak scoorden (de kinderen met de 25-30% zwakste scores) werden geselecteerd voor *Bouw!*.

De bevindingen laten zien dat de selectieprocedure die werd gebruikt in de eerste en tweede studie van dit proefschrift, leidde tot 15% vals negatieven (15% van de kinderen die niet in aanmerking kwamen voor de interventie, ontwikkelde toch leesproblemen eind groep 3), eenzelfde percentage als werd gevonden bij andere screeningsprocedures voor groep 2 die wetenschappelijk zijn onderzocht. Daarmee is de voorgestelde selectieprocedure een goed startpunt voor Nederlandse basisscholen om kinderen op te sporen met een risico op leesproblemen. De screeningsprocedure kan verbeterd worden door een derde screeningsmoment toe te voegen die plaatsvindt aan het begin van groep 3, nadat kinderen enkele maanden leesonderwijs hebben gekregen. Daarmee kunnen scholen in een relatief vroeg stadium nog een aantal zwakke lezers opsporen en die alsnog laten starten met de interventie. Voor deze kinderen is het wellicht belangrijk dat hun oefentijd wordt verhoogd, zodat zij voldoende voortgang maken binnen de interventie om het programma voor midden groep 4 af te ronden.

De resultaten laten ook zien dat een aanzienlijk deel van de kinderen die voor de interventie in aanmerking kwamen, niet met *Bouw!* zijn gestart. Deze beslissing leek gebaseerd te zijn op hun lage leeftijd en hun hoge benoemsnelheid[4]. Deze groep kinderen had een minder groot risico op leesproblemen dan de groep die in aanmerking kwam en *wel* met *Bouw!* startte: een kleiner percentage van deze groep ontwikkelde later leesproblemen. De beslissing om hen niet met *Bouw!* te laten starten, had zowel

[4] Benoemsnelheid is het vermogen om de namen van bekende symbolen (plaatjes, cijfers, letters) snel uit het langetermijngeheugen op te halen.

positieve als negatieve gevolgen voor het bereik van de interventie. Aan de ene kant werd voorkomen dat sommige kinderen *Bouw!* met Bouw! startten, terwijl ze het niet nodig hadden. Aan de andere kant startten sommige kinderen niet met *Bouw!*, terwijl ze het wel nodig hadden.

Er was ook een aanzienlijk deel van de kinderen die met *Bouw!* startte, terwijl die hier niet voor in aanmerking kwamen. Deze beslissing leek gebaseerd te zijn op hun lage benoemsnelheid, hun minder geletterde thuisomgeving (gerapporteerd door de leerkracht) en/of hun meer praktisch geschoolde ouders. Ook werden jongens, kinderen met een familiair risico op dyslexie en kinderen met Nederlands als tweede taal vaker alsnog geselecteerd voor *Bouw!* dan respectievelijk meisjes, kinderen zonder een familiair risico en kinderen met Nederlands als moedertaal. Deze groep had ook een groter risico op leesproblemen dan de groep kinderen die niet aanmerking kwam en niet met *Bouw!* startte: een groter percentage ontwikkelde leesproblemen. De beslissing om hen met *Bouw!* te laten starten, had zowel positieve als negatieve gevolgen voor het bereik van de interventie. Aan de ene kant werd voorkomen dat sommige kinderen (ernstige) leesproblemen kregen door hen te laten starten met *Bouw!*. Aan de andere kant startten sommige kinderen met *Bouw!*, terwijl ze het niet nodig hadden. Bij het selecteren van kinderen voor de interventie kunnen scholen wellicht het best het zekere voor het onzekere nemen. Dat betekent dat ze de interventie zouden kunnen aanbieden aan alle kinderen die hiervoor in aanmerking komen en aan kinderen van wie scholen denken dat ze ook baat hebben bij de interventie op basis van kenmerken die niet zijn opgenomen in de selectieprocedure. Op deze manier worden leesproblemen bij zoveel mogelijk kinderen voorkomen.

Dit heeft echter gevolgen voor het aantal vals positieven (kinderen die met *Bouw!* starten, terwijl ze het niet nodig hebben), wat tijd en geld kost. Daarom willen scholen mogelijk vals positieven opsporen tijdens de interventie. Scholen in onze steekproef toetsten kinderen midden groep 3 op woordleesvloeiendheid en lieten een deel van de kinderen stoppen met *Bouw!*, veelal goede en bovengemiddelde lezers (kinderen met A- en B-scores). De resultaten in de algemene discussie laten zien dat slechts één van de 46 goede en bovengemiddelde lezers die stopten met *Bouw!*, vervolgens leesproblemen ontwikkelde. Ook heb ik de leesontwikkeling van een groep kinderen die midden groep 3 stopte vergeleken met een groep kinderen die doorging en vergelijkbaar was qua niveau. De resultaten laten zien dat het geen verschil maakte of kinderen stopten of doorgingen met *Bouw!* na midden groep 3. Hierbij is het goed om op te merken dat met name de goede lezers stopten. Zwakke lezers hebben mogelijk wel baat bij het voortzetten van de interventie. De resultaten suggereren dat scholen vals positieven kunnen opsporen door midden groep 3

woordleesvloeiendheid te toetsen en dat de meeste kinderen die midden groep 3 goede of bovengemiddelde leesvaardigheden hebben en met *Bouw!* stoppen, vervolgens geen leesproblemen ontwikkelen, waardoor er meer tijd beschikbaar komt voor kinderen die de interventie echt nodig hebben.

Conclusie

Onderzoek naar de effectiviteit van een interventie stopt vaak nadat een interventie is geëvalueerd in een of meerdere kleinschalige studies waarbij de implementatie wordt begeleid door onderzoekers. Maar dan is nog niet bekend of de interventie ook effectief is in de praktijk. De resultaten van dit proefschrift benadrukken dat het belangrijk is om de effectiviteit van interventies ook te onderzoeken in een natuurlijke situatie, waarbij de implementatie de eigen verantwoordelijkheid is van de school. De resultaten impliceren dat de effecten van effectief bewezen interventies in de praktijk mogelijk pas zichtbaar zijn op schoolniveau, nadat scholen enkele jaren ervaring hebben met de interventie, dat de effecten klein zijn en dat ze kunnen worden beïnvloed door vele factoren, zoals hoeveel kinderen oefenen, de mate waarin scholen de juiste kinderen weten te selecteren voor de interventie en of er leerproblemen zijn in de familie van de kinderen die de interventie krijgen. Er zijn meer studies nodig die gericht zijn op de effecten van effectief bewezen interventies in een natuurlijke situatie om zo meer kennis te verkrijgen die scholen kunnen gebruiken om effectief bewezen interventies te laten slagen in de praktijk.

Dankwoord

Veel mensen hebben bijgedragen aan de totstandkoming van dit proefschrift. In dit dankwoord wil ik iedereen bedanken die me in de afgelopen jaren heeft geholpen en gesteund.

Allereerst wil ik alle **scholen**, **ouders**, **kinderen** en **schoolbesturen** bedanken die hebben deelgenomen aan dit onderzoek. Zonder jullie inzet en tijd had dit onderzoek nooit kunnen plaatsvinden! Ik wil alle schoolleiders bedanken voor het steunen van dit onderzoeksproject. In het bijzonder wil ik alle leerkrachten, intern begeleiders en andere betrokkenen bedanken die kinderen hebben getoetst voor dit onderzoek, ook tijdens de coronapandemie.

De grootste bijdrage aan dit proefschrift is geleverd door mijn promotor en copromotoren. **Peter de Jong**, je hebt de grote lijnen van dit proefschrift uitgezet en interessante invalshoeken aangedragen bij de analyses en het schrijven van dit proefschrift. Ook hebben jouw ideeën ervoor gezorgd dat we bruikbare variabelen hebben kunnen destilleren uit de chaotische logboeken van *Bouw!*. Daarnaast heb je een grote bijdrage geleverd aan de modellen in dit proefschrift, waarbij je oplossingen hebt bedacht voor allerlei problemen, zoals ontbrekende data en extreme oefentijden. Verder heb je me begeleid bij het schrijven, herschrijven en publiceren van de artikelen, wat voor mij het lastigste onderdeel was. Bedankt voor je begeleiding in elke fase: voor je feedback, bereikbaarheid, snelle reacties, betrokkenheid en geduld met mij. Het was ontzettend leuk dat je steeds zo nieuwsgierigheid was naar de resultaten van het onderzoek! Ik waardeer het enorm dat je altijd tijd voor me maakte en ik bij je binnen kon lopen. **Madelon van den Boer**, bedankt voor al jouw ideeën voor dit proefschrift, je kritische vragen bij de analyses en je hulp bij het stellen van prioriteiten, waarbij jouw relativeringsvermogen erg behulpzaam was. Bovenal bedankt voor je begeleiding bij schrijven en herschrijven van de artikelen! Zonder jou zou er weinig structuur in de artikelen zitten, waren de zinnen in dit proefschrift onnodig lang en wollig, en zouden alle tabellen 'Tabel 1' hebben geheten. **Haytske Zijlstra**, bedankt voor het ontwerpen van de selectieprocedure, het opzetten van de dataverzameling en voor de puntjes op de i bij het schrijven, waarbij jouw kennis van het programma en van de dagelijkse praktijk onmisbaar was! Bedankt dat je me hebt laten kennismaken met mensen die een belangrijke rol hadden in het onderzoeksproject. Bedankt voor de leuke foto's in dit proefschrift. Ook wil ik je bedanken voor jouw betrokkenheid, gezellige koffiemomentjes, vrolijke telefoontjes en berichtjes op

Whatsapp en de kansen die je me gaf om de resultaten van ons onderzoek naar buiten te brengen.

Daarnaast wil ik alle opponenten, **Elise de Bree**, **Frank Cornelissen**, **Pol Ghesquière**, **Judith Rispens** en **Eliane Segers**, bedanken voor de tijd die jullie hebben gestoken in het beoordelen van dit proefschrift en in het opponeren.

Twee personen hebben een belangrijke bijdrage geleverd aan het analyseren van de data. **Bonne Zijlstra**, bedankt voor je bijdrage aan de *multilevel models* en *difference-in-difference models*: voor je hulp bij het bouwen en trimmen van de modellen, bij het checken van de assumpties en bij het vormgeven van dummyvariabelen. Ook heb je oplossingen bedacht voor problemen met verschillende toetsversies, ontbrekende data en covid. **Arno Havermans**, bedankt dat je de anonieme logboeken van *Bouw!* hebt getransformeerd in bruikbare datasets, die we op les-, sessie-, week- en kindniveau konden bekijken en analyseren. We hebben de variabelen vaak moeten herzien en telkens stond je voor ons klaar! Ook heb ik veelvuldig gebruik gemaakt van jouw uitgebreide kennis van *Bouw!*, nauwkeurig vastgelegd in Excelbestanden. Ik zou niet weten hoe ik dit onderzoek had moeten doen zonder al jouw kennis en kunde! Ook bedankt voor de mooie foto's in dit proefschrift.

Verschillende mensen van Lexima hebben dit onderzoek mogelijk gemaakt. **Harry Kleintjes** en **Falk Beerten**, bedankt voor het aangaan, steunen en faciliteren van dit onderzoeksproject! **Daniëlle Vogelpoel**, bedankt voor de vriendelijke en duidelijke communicatie met de scholen. Ik vond het heel fijn om met je samen te werken. Ik kreeg van jou altijd een warm ontvangst bij Lexima. **Tanja Piera**, bedankt voor je hulp bij het bouwen van de online vragenlijsten en de begeleiding bij de gehele administratie rondom de dataverzameling! Ik vond het heel fijn om met je samen te werken, o.a. vanwege je organisatorisch vermogen en scherpe blik. **Rosalin van der Hoeven**, bedankt voor je hulp bij het werven van schoolbesturen, de communicatie met de scholen en het leiden van de projectgroep! Ik vond het fijn om met je samen te werken. **Suzan Ramakers**, bedankt voor het leiden van de projectgroep en je werk voor het onderzoek achter de schermen. **Tom Kleintjens**, bedankt voor alle logboekbestanden.

Ik wil alle studenten bedanken die hebben geholpen de data te verzamelen in District 2: Ellis Spiering, Femke Borst, Tessa Kuiper, Laurine Michel, Eline Roep, Kim Eilander, Heleen Boogaart, Jessica Schipper, Lisanne Krosse, Colinda van Harten, Aniek van den Hoogen, Lotte van der Waal, Liena Ghzawi, Esmée Westerink, Iza van der Maat, Lisanne de Krosse, Sabine Klinkhamer, Miriam Bouanane, Noortje Slagter, Zara Hage, Charlotte Verhoeve, Fraukje de Haan en Monique Oehlers.

Verder wil ik alle mensen bedanken die hebben geholpen schoolbesturen te werven voor dit onderzoek: Haytske Zijlstra, Rosalin van der Hoeven, Cindy Deijle, Niels de Ruig, Ellen van der Steene, Judith Kuipers, Heleen Reinds, Anne Velders, Monica Tolenaar en Madelon van den Boer.

Nicole Siers, bedankt voor het ontwerpen van de omslag. Bedankt voor je creativiteit, eigen stijl en voor de fijne samenwerking!

Ook wil ik alle collega's bedanken die mij hebben vergezeld, gesteund en bijgestaan tijdens mijn promotieonderzoek. **Mengdi Chen**, roommate, I spent most of my time with you. Thank you for taking me to the best Asian restaurants in Amsterdam. You taught me how to make dumplings and to cook green beans the Chinese way! Thank you for answering questions about statistics and APA. In particular, thank you for listening! **Rianne Bosman**, bedankt voor je hulp tijdens de eerste fase van mijn promotieonderzoek. Bedankt voor de goede boekentips, voor het organiseren van leuke uitjes (spelletjesavond, etentjes) en voor het verzorgen van verjaardagscadeautjes. Jij zorgde ervoor dat er niet alleen hard werd gewerkt, maar dat het ook gezellig was in de onderzoeksgroep! **Alexander Krepel**, bedankt voor de goede restaurants waar je ons mee naartoe nam, voor je gezellige borrelethos en voor je kookkunsten tijdens de schrijfweek. Het recept 'bloemkool uit de oven met harissa en kikkererwten' heb ik na de schrijfweek nog vaak gemaakt. **Janneke de Ruiter**, bedankt voor je gezelligheid en vrolijke verhalen op kantoor. Jij toverde altijd een lach op mijn gezicht! **Loes Bazen**, wat vond ik jouw gezelschap fijn, toen je veel op de UvA te vinden was! Bedankt voor je openheid, begrip, en leuke boekentips op kantoor en tijdens onze schrijfweek in Kasteel Slangenburg. **Niels de Ruig**, bedankt voor het samen optrekken tijdens hoogte- en dieptepunten, voor je telefoontjes en vrolijke verhalen! Ook wil ik je bedanken voor het werven van een schoolbestuur voor dit onderzoek. **Qingqing Du**, thank you for being always in the office, for your motivating work ethic, for sharing ups and downs during our PhD, and for having fun together! **Lotte Visser**, bedankt voor de gezelligheid die je op het kantoor bracht en voor je betrokkenheid in de laatste periode van mijn promotieonderzoek! **Emiel Schoneveld**, bedankt voor de energie en vrolijkheid die je op kantoor bracht en voor het meedenken over de omslag van mijn proefschrift. **Manon Toonen** en **Lara Luberti**, bedankt voor de leuke praatjes op kantoor en voor het meeleven met de laatste fase van mijn promotieonderzoek! **Bieke Schreurs**, bedankt dat je samen met mij de scholen hebt geworven uit District 2. Ik vond het leuk om samen met jou interviews te doen op de focusscholen en om bijeenkomsten te organiseren voor de scholen. Daarnaast wil ik je bedanken voor alles wat ik van je heb geleerd op inhoudelijk vlak en persoonlijk vlak. Dan denk ik aan jouw onderzoeksresultaten, aan interessante artikelen die je

met me hebt gedeeld en aan het omgaan met kritiek. **Frank Cornelissen,** bedankt voor je onmisbare bijdrage aan het onderzoeksproject. Ik vond het erg leuk om van je te leren over onderwijsinnovaties en wat daarbij komt kijken. **Helma Koomen**, bedankt dat je me de vacature voor deze promotieplek hebt doorgestuurd. Dit onderwerp past veel beter bij mij! Ook wil ik je bedanken voor je vrolijke verhalen aan de lunchtafel, je bracht sfeer in de onderzoeksgroep! **Debora Roorda**, bedankt voor alle tips voor het begeleiden van bachelorscripties! Daar heb ik veel aan gehad. Bedankt voor je grappige en vrolijke verhalen, eigenheid en aanwezigheid bij borrels en etentjes! Je was altijd van de partij! **Marjolein Zee**, bedankt je tips voor het begeleiden van bachelorscripties, de uitnodiging voor je bruiloft en voor je gezelligheid tijdens borrels en etentjes. **Elise de Bree**, bedankt voor je vrolijkheid, interesse en inlevingsvermogen op de zaak en tijdens de schrijfweek. Je bent het zonnetje in huis! **Aryan van der Leij**, jij stond aan de wieg van *Bouw!*. Je hebt *Bouw!* ontworpen, onderzocht en op de markt gebracht. Bedankt dat ik hier verder op mocht bouwen. Bedankt voor je bijdrage aan de projectgroep, je bevlogenheid en inspirerende lezingen bij het NDC.

Tenslotte wil ik alle mensen thuis bedanken die mij hebben ondersteund. Lieve **André**, bedankt voor je onvoorwaardelijke steun. Bedankt dat je altijd naar me luisterde en met me meeleefde. Bedankt dat je Guus in het weekend mee op pad nam, zodat ik het proefschrift af kon maken. Bedankt voor alle tijd die je hebt gestoken in het zorgvuldig vormgeven van het binnenwerk! Je hebt het met veel geduld, gevoel voor esthetiek en oog voor detail gedaan. Ik wil mijn ouders, **Lilia** en **Fridus van der Weijden**, bedanken voor jullie vertrouwen in mij, voor jullie luisterend oor en stimulans om door te zetten, ook wanneer het even tegenzat!

Publications

Peer-reviewed publications

van der Weijden, F. A., van den Boer, M., Zijlstra, B. J. H., & de Jong, P. F. (2024c). Implementation takes time: Reduction of literacy problems in schools implementing an early-literacy intervention. *Journal of Research on Educational Effectiveness*, 1–33. http://dx.doi.org/10.1080/19345747.2024.2384365

Papers in progress

van der Weijden, F. A., van den Boer, M., Zijlstra, A. H., & de Jong, P. F. (2024a). A school-based implementation of an early-literacy intervention: Relations among dosage, familial risk, parental education, and reading acquisition [Manuscript submitted for publication]. Department of Child Development and Education, University of Amsterdam.

van der Weijden, F. A., van den Boer, M., Zijlstra, A. H., van der Leij, A., Zijlstra, B. J. H., & de Jong, P. F. (2024b). Dosage explains individual differences in the outcomes of a prevention program for literacy problems [Manuscript submitted for publication]. Department of Child Development and Education, University of Amsterdam.

Vroege leesinterventies kunnen leesproblemen mogelijk voorkomen, ook op lange termijn, maar de effecten zijn zelden onderzocht in de praktijk. Dit proefschrift gaat over de vroege leesinterventie Bouw!, die op grote schaal wordt geïmplementeerd op Nederlandse basisscholen. De belangrijkste bevindingen laten zien dat de interventie-uitkomsten samenhingen met de tijd die kinderen besteedden aan de interventie, met name het aantal interventiesessies per week. Kinderen met een familiair risico op dyslexie gingen minder snel vooruit binnen de interventie, wat betekent dat ze meer sessies nodig hebben dan gemiddeld. Er was geen onmiddellijk effect van Bouw! op schoolniveau, maar nadat scholen twee of meer jaren ervaring hadden met Bouw!, was er een kleine afname te zien in het percentage kinderen met problemen in technisch lezen, spelling en begrijpend lezen, evenals een kleine toename in de gemiddelde scores op deze vakken binnen de school.