The effects of Success for All in the Netherlands on the reading achievement of first-grade students at risk of reading problems

Hingstman, Mariëtte; Warrens, Matthijs J.; Doolaard, Simone; Bosker, Roel J.

Published in:
Studies in Educational Evaluation

DOI:
10.1016/j.stueduc.2023.101257

Publication date:
2023

Studies in Educational Evaluation 77 (2023) 101257

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/locate/stueduc

The effects of Success for All in the Netherlands on the reading achievement of first-grade students at risk of reading problems

Mariëtte Hingstman [*,1], Matthijs J. Warrens [2], Simone Doolaard [3], Roel J. Bosker [4]

University of Groningen, Faculty of Behavioral and Social Sciences, GION Education/Research, Grote Rozenstraat 3, 9712 TG Groningen, the Netherlands

ARTICLE INFO

Keywords:
Students at risk
Success for All
School reform
Struggling readers
Response to Intervention
Multilevel analysis

ABSTRACT

This article reports outcomes of a quasi-experimental evaluation of Success for All (SfA), a comprehensive school reform program that has recently been introduced in the Netherlands. The Response to Intervention framework is used to describe how SfA supports students at different Tiers. The effects of SfA on five reading subskills were investigated for first-grade students at risk of reading problems. 299 students from two different cohorts were involved. Multilevel analyses demonstrated a significant effect of SfA on reading comprehension (ES = +0.26) in the first cohort. For the second cohort and the other reading subskills, mostly small positive effects of SfA were found, though these effects were not statistically significant. Furthermore, the relationship between tutoring intensity and reading achievement was examined. In the second cohort, a significant negative association of tutoring intensity with word and text reading skills was found. Implementation issues that may have impacted the outcomes are discussed.

1. Introduction

In research regarding students at risk of academic failure, much emphasis is placed on good reading skills. Being a proficient reader is a significant predictor for success in school and life (O'Connor & Vadasy, 2011; Snow et al., 1998). Therefore, it is alarming that around 23% of 15-year-olds in OECD countries perform below the baseline level of reading proficiency (OECD, 2019). Due to the covid-19 pandemic, reading difficulties have worsened, especially among younger students and students from disadvantaged backgrounds (Engzell et al., 2021; Hammerstein et al., 2021). In the Netherlands, where the current study is conducted, a downward trend in reading achievement was observed in both PIRLS (concerning 10-year-olds) and PISA (concerning 15-year-olds) in the past decade (Mullis et al., 2017; OECD, 2019). Furthermore, the most recent PIRLS and PISA studies showed that the Dutch students are among the least motivated for reading, which increases the risk of ending up in a vicious circle of poor performance and demotivation (Hebbecker et al., 2019; Mullis et al., 2017; OECD, 2019). To ensure that more students develop sufficient reading skills, investment in both prevention and intervention of reading problems is needed. There is compelling evidence that high-quality instruction can reduce the gap between students with high and low (pre-)literacy skills (Dietrichson et al., 2017; Fletcher et al., 2018; Hagans & Good, 2013; Neitzel et al., 2021; Reschly, 2010).

1.1. Response to intervention

A widely used framework for the identification and early intervention of reading problems is the Response to Intervention (RTI) model. While RTI has its origins in the United States, it has become a blueprint for support systems in other countries as well (for a well-known Dutch example of a RTI model targeted at reading difficulties, see Gijsel et al. (2011)). The RTI model is multi-tiered, with whole-class instruction at Tier 1 (provided to all students), additional instruction at Tier 2 (provided to small groups), and intensified one-to-one instruction at Tier 3, for the lowest achievers (Fuchs & Fuchs, 2006). Monitoring student progress is crucial, so struggling students can be moved to a higher tier as soon as possible (Fuchs & Fuchs, 2006; Houtveen & van de Grift,

* Corresponding author.
E-mail address: m.hingstman@rug.nl (M. Hingstman).
[1] ORCID: 0000–0002-4199–2427
[2] ORCID: 0000–0002-7302–640X
[3] ORCID: 0000–0001-6977–2904
[4] ORCID: 0000–0002-1495–7298

https://doi.org/10.1016/j.stueduc.2023.101257
Received 12 August 2022; Received in revised form 15 February 2023; Accepted 13 March 2023
Available online 21 March 2023

2007). In RTI, learning disabilities are redefined as 'inadequate response to intervention'. The so-called nonresponders face enduring problems in their development and therefore might qualify for special education services, which could be considered Tier 4 (Fuchs & Fuchs, 2006). Nevertheless, for a substantial proportion of young struggling learners, it is assumed that appropriate instruction can prevent learning disabilities (Burns et al., 2005; Reschly, 2010; Snow et al., 1998). It is important to note that the potential of RTI depends on the quality of the program(s) used and of the implementation (Fuchs & Fuchs, 2009; Hoover, 2010).

1.2. Evidence-based multi-tiered reading instruction

How to effectively teach children how to read is described in several studies, of which the reviews of Snow et al. (1998) and the National Reading Panel (2000) have been two of the most influential ones. Twenty years later, key elements described in these reports (phonemic awareness, phonics, fluency, vocabulary, and text comprehension) are still seen as the foundations of literacy instruction, as later reviews also demonstrated positive effects of focusing on these elements (Education Endowment Foundation, 2017; Gersten et al., 2017). Some recommendations from these reports are that beginning readers need explicit instruction to become familiar with the alphabetic principle, and that students should be exposed to a variety of texts to develop fluency in reading. Teachers play an important role in promoting comprehension by building background knowledge and vocabulary, and by teaching text comprehension strategies. Furthermore, speaking and listening activities stimulate students' wider understanding of language. Integrating cooperative learning activities in reading lessons can be helpful in this regard (Dietrichson et al., 2017). Because of the reciprocal relationship between intrinsic reading motivation and reading achievement, it is crucial that students experience feelings of competence when learning to read (Hebbecker et al., 2019). To help students acquire adequate reading skills, it is also important that teachers closely monitor students' progress and provide tailored feedback (Dietrichson et al., 2017; Wisniewski et al., 2020). As teachers play a pivotal rule in students' reading development, attention should be paid to ongoing professional development and administrative support (Desimone, 2002; Durlak & DuPre, 2008; Fuchs & Vaughn, 2012; Hughes & Dexter, 2011).

Most children who experience difficulties in learning to read do not need radically different types of instruction (Snow et al., 1998). They particularly benefit from intensified instruction, which is often provided in the form of tutoring (Fuchs & Vaughn, 2012). A recent synthesis of research on programs for struggling readers in primary education (Neitzel et al., 2021) reported positive effects of both one-to-one and small-group tutoring programs, especially when integrated in a whole-class or whole-school program. Substantial positive effects of tutoring were also found in the meta-analyses of Dietrichson et al. (2017) and Gersten et al. (2020).

The current study focuses on a program that combines the effective elements described above and that has demonstrated positive effects in several (quasi-)experimental studies (Cheung et al., 2021), in particular for initially low-achieving students: Success for All (SfA).

1.3. Success for All

In the 1980's the SfA program was developed by researchers from Johns Hopkins University in Baltimore (United States), primarily targeting schools that serve large numbers of students from disadvantaged backgrounds. These students are more vulnerable to reading failure than others, because they often have limited financial, social and cultural resources (Cortiella & Horowitz, 2014; OECD, 2016). Although schools serving many disadvantaged students often receive additional funding (e.g. Title 1), mitigating the effect of background remains a challenge (Dietrichson et al., 2017). SfA attempts to ensure that every child achieves his or her fullest potential by incorporating various evidence-based components: daily 90-minute reading lessons that contain explicit instruction and cooperative learning; frequent assessment; flexible regrouping; tutoring for struggling readers; a kindergarten program; and Solutions Teams that cover social-emotional learning, parental involvement and attendance. A full-time facilitator oversees implementation and supports teachers, and school teams receive training and coaching from an external SfA coach (Slavin et al., 2009). Although the majority of staff members working at SfA schools agree that SfA benefits their school, they often struggle to implement all key features with high fidelity. In particular the first year of implementation can be demanding because of the comprehensiveness and complexity of the program (Quint et al., 2015).

The SfA program can be characterized in multiple ways. SfA is often described as a comprehensive school reform (CSR) program (e.g. Cheung et al., 2021; CSRQ, 2006; Van Kuijk et al., 2021), because of its broad scope, with components addressing reading instruction, social-emotional learning, parental involvement and professional development. Overall, CSR programs are known to be effective, although large differences in implementation quality and outcomes are observed (Borman et al., 2003; CSRQ, 2006). SfA is a CSR program that has demonstrated robust positive effects across different contexts and study designs (Borman et al., 2003; CSRQ, 2006). As the current study focuses on the reading instruction component of SfA, consisting of a classroom program and a tutoring intervention, we will additionally use the RTI framework to describe how SfA aims to differentiate instruction for students at risk of reading problems at Tier 1, 2 and 3. A meta-analysis of Deunk et al. (2018) on effective differentiation practices found the greatest effects for practices that were embedded in a broader (CSR) program, as is the case in SfA. Fuchs and Fuchs (2009) also argue that RTI models have the largest potential to improve education when interventions at multiple Tiers are unified within a school.

Schools in the Netherlands face challenges in combating educational disadvantages as well (Van Huizen, 2018). The achievement gap between children from disadvantaged and non-disadvantaged families has even increased in recent years (Inspectie van het Onderwijs, 2018). Therefore, it was decided to set up a Dutch version of SfA through a Research & Development (R&D) project, in collaboration with primary schools that wanted to improve their students' reading outcomes.

1.4. Present study

In this study, for the first time the effects of the Dutch SfA program on the achievement of students who are at risk of reading problems were investigated. As American studies on the effects of SfA for this subgroup have demonstrated positive outcomes, it is important to investigate whether SfA could produce similar benefits in a different context. The scope of the current study is broad: where the previous studies on SfA only reported one, two or three types of reading outcomes (Cheung et al., 2021), we decided to include five different outcome measures, in order to do justice to the multifaceted character of the early reading process. We focus on students in first grade, because of the importance of intervening early in the lives of students at risk. The main research question to be answered is: *"What are the effects of Success for All on the reading achievement of students at risk of reading problems?"* Based on previous research on SfA and on evidence-based multi-tiered reading instruction in general (e.g. Cheung et al., 2021; Bursuck & Blanks, 2010; Neitzel et al., 2021), we hypothesize that students at risk at SfA schools will perform better in reading than students at risk at control schools.

Within the group of first-grade students at SfA schools, we will investigate a second research question: *"Is there a relationship between the intensity of the Success for All tutoring intervention and students' reading achievement?"* In general, spending extra time on reading is associated with better outcomes (Neitzel et al., 2021), which leads to the expectation that students who receive the most intensive forms of tutoring will show the most progress during first grade. However, the relationship might be weak, because students receiving highly intensive interventions are often the ones struggling most (according to the idea of

RTI). Students' pre-literacy skills (measured at the end of kindergarten) will be taken into account in the analyses.

2. Method

2.1. Design

A quasi-experimental design was used in this study. Ten schools from three school boards were involved. Schools were assigned to conditions in accordance with agreements between the school boards and the R&D team, resulting in six schools that implemented SfA and four schools that served as control schools. Control schools continued their 'business as usual', with *Veilig Leren Lezen (VLL), 2e maanversie* (Zwijsen, 2003), a widely used phonics curriculum for first graders in the Netherlands, being used during the whole-class reading lessons. The *VLL* manual does not prescribe an exact lesson duration, but the activities are considered to fit within a 90-minute time frame (Zwijsen, personal communication, April 22, 2020). For struggling readers, most control schools provided additional instruction in small groups, using either activities from *VLL* or programs like *Connect* (Smits & Braams, 2006). Furthermore, the computer program *Lezen met Zoem* (part of *VLL*) was used to let struggling readers have extra practice (independently).

We evaluated the progress of students at risk in two cohorts: one consisting of students who were in first grade in 2016–2017 (cohort 1); the other consisting of students who were in first grade in 2017–2018 (cohort 2). Students were pretested at the end of kindergarten and were posttested at the end of Grade 1. By then, students in the SfA condition were exposed to the SfA program for one school year.

2.2. Intervention

Since 2015, SfA is being implemented in Dutch primary schools. In the daily 90-minute reading lessons (Tier 1), that are described in detail in teacher manuals, much emphasis is placed on cooperative learning. SfA's Grade 1 program, *Reading Roots,* was integrated with *Veilig Leren Lezen*. As SfA was introduced step-wise in the schools (one grade at a time), flexible regrouping was not yet applied. In the meantime, teachers could differentiate their instruction for low achievers by using 'light' (shortened and simplified) versions of books and by providing extended instruction during other times of the day and/or at the end of each seven- or eight-week instructional cycle. Furthermore, teachers could assign struggling readers to tutoring by a (para)professional, preferably provided one-to-one or in pairs (Tier 2 or 3). Schools were advised to schedule 20-minute tutoring sessions for at least four days a week. To assist tutors, a tutoring manual was developed, containing assessments and activities that were aligned with the Tier 1 SfA program. For each student, a personalized tutoring plan had to be written, in order to tailor the activities to the student's needs. After each eight- or nine-week tutoring cycle, tutors had to evaluate the student's progress with the teacher and the SfA facilitator, and the need for further tutoring had to be determined. Regarding Solutions Teams, only the parental involvement module was used in the first years of SfA implementation in the Netherlands.

2.3. Treatment fidelity

Various efforts have been made to increase fidelity of implementation of the SfA program. Teachers who started teaching SfA followed a one-day training course. Additionally, members of the R&D team visited classrooms, provided teachers with feedback and arranged meetings with teachers and tutors from different schools, aimed at sharing experiences. Teachers and tutors were also supported by the SfA facilitator at their school, who was appointed for four hours a week and therefore often combined facilitating tasks with a job as instructional coach or teacher. The amount of coaching and assistance SfA teachers in the Netherlands received was less compared to what is common in SfA schools in the U.S. Taking into account that achieving a high level of fidelity is already difficult for U.S. schools (e.g. Quint et al., 2015), this might explain why some implementation issues were faced at the participating Dutch schools. From observations, meetings and a questionnaire, we deduced that there was variation between teachers in how they managed to teach SfA lessons. Things several teachers struggled with were guiding cooperative learning activities and incorporating all SfA's lesson elements into the daily classroom schedule (which sometimes resulted in a shortening of the 90-minute lessons). Nevertheless, SfA seemed to help teachers in improving their teaching skills. The participating schools' fidelity to the SfA program as a whole is described in more detail (in Dutch) by Mullender-Wijnsma et al. (2020).

As our study concerns students at risk, we will mainly focus on the schools' fidelity to the tutoring component of SfA. Interviews with tutors, school records and observations of tutoring sessions (which were conducted in both school years) demonstrated some differences between schools and between tutors. Although tutoring was provided at all six schools and was widely valued as an effective tool to support struggling readers, some schools had difficulty organizing the tutoring sessions as prescribed. In both school years, because of staffing issues one school started late (in February/March) and in 2017–2018 another school stopped early (in May). Instead of being tutored by a (para)professional, some students (10% in 2016–2017, 26% in 2017–2018) were tutored by a volunteer or intern. Although most schools managed to provide tutoring one-to-one or in pairs, tutoring in small groups also took place (concerning 17% of tutored students in 2016–2017, and 23% in 2017–2018). Meeting the expectations with regard to intensity turned out to be difficult as well: in 2016–2017, only 60% received tutoring at least four times a week. Most sessions lasted 20 min, as prescribed. In 2017–2018, the pattern was different: now, 82% received tutoring at least four times a week, but the sessions were shortened to 15 min at most schools. Moreover, sessions were sometimes canceled because tutors had other duties, such as substituting absent teachers. At two schools, the SfA facilitator coordinated tutoring (i.e. oversees students' progress, supports tutors). Two schools assigned a teacher for coordinating tasks, and at the other two schools the tutors themselves were mainly responsible for the entire tutoring process. We observed that tutors who received little support had difficulty setting and evaluating goals and engaging students.

Although tutors started to use the activities suggested in the manual less frequently after the initial implementation phase, the skills that were mainly practiced during the sessions (phonemic awareness, letter knowledge, word knowledge and fluency) remained the same. At one school, in 2016–2017 activities from *VLL/Reading Roots* were alternated with activities from the external program *BLOON* (Lexima, 2020). In 2017–2018, another school started using the external program *Bouw!* (Lexima, 2019) for tutoring.

2.4. Sample

We evaluated the effects of SfA on a subgroup. The target population of this study consists of Dutch first-grade students at risk of reading problems. In this study, 'at risk' is defined as: students entering Grade 1 with low pre-literacy skills. We included students whose scores fell into category IV or V (the lowest 40%) on one or both subtests of the Screeningsinstrument Beginnende Geletterdheid (SBG), and/or on the Taal voor Kleuters (TVK) pretest. We made use of two kindergarten pretests because they are complementary with regard to measured skills, according to its developers (Lansink & Hemker, 2010), and thus together provide a more complete picture of kindergartners' pre-literacy skills. The total cohort 1 sample consisted of 148 students at risk of reading problems (89 SfA, 59 control) from 19 classes (9 SfA, 10 control) at 10 schools (6 SfA, 4 control). The total cohort 2 sample consisted of 151 students at risk of reading problems (75 SfA, 76 control) from 19 classes (9 SfA, 10 control) at 10 schools (6 SfA, 4 control).

The participating schools were located in the north of the

Table 1
School and student sample characteristics at baseline.

	SfA	Control
Cohort 1 schools (2016–2017)	Mean (*SD*) (n = 6)	Mean (*SD*) (n = 4)
School size (number of enrolled students)	267 (50.9)	412 (157.7)
Students from disadvantaged backgrounds (%)	14.8 (4.9)	14.8 (6.2)
Cohort 1 students (2016–2017)	Mean (*SD*) (n = 89)	Mean (*SD*) (n = 59)
Age in months at beginning Grade 1	75.9 (4.9)	75.7 (4.0)
Score on SBG phonological awareness pretest	203.4 (21.2)	199.1 (19.6)
Score on SBG letter knowledge pretest	20.6 (7.0)	16.8 (7.0)
Score on TVK pretest	63.8 (8.1)	65.1 (9.8)
	No (%) (n = 89)	No (%) (n = 59)
Gender - male	48 (53.9%)	38 (64.4%)
Students from disadvantaged backgrounds	13 (14.8%)	13 (22.0%)
Cohort 2 schools (2017–2018)	Mean (*SD*) (n = 6)	Mean (*SD*) (n = 4)
School size (number of enrolled students)	258 (55.6)	383 (153.3)
Students from disadvantaged backgrounds (%)	16.8 (6.7)	13.5 (5.2)
Cohort 2 students (2017–2018)	Mean (*SD*) (n = 75)	Mean (*SD*) (n = 76)
Age in months at beginning Grade 1	76.8 (4.8)	76.6 (4.6)
Score on SBG phonological awareness pretest	194.3 (20.5)	197.1 (18.1)
Score on SBG letter knowledge pretest	19.9 (6.6)	18.3 (6.6)
Score on TVK pretest	66.6 (8.1)	65.7 (8.2)
	No (%) (n = 75)	No (%) (n = 76)
Gender - male	46 (61.3%)	51 (67.1%)
Students from disadvantaged backgrounds	14 (18.7%)	10 (13.2%)

Netherlands, in neighborhoods with relatively many low-income households (Centraal Bureau voor de Statistiek, 2016). Via the weighted student funding system in the Netherlands, schools received additional resources for students from disadvantaged backgrounds. In this system, a distinction is made between students with a 0.3[5] disadvantaged background (both parents/legal guardians completed less than two years of secondary school) and a 1.2[5] disadvantaged background (one or both parents/legal guardians did not participate in secondary education at all). In both school years, intervention schools and control schools did not differ significantly in terms of school size (cohort 1: $t(8) = -2.141$, $p = .07$; cohort 2: $t(8) = -1.876$, $p = .10$) or percentage of students from disadvantaged backgrounds (cohort 1: $t(8) = 0.005$, $p = .996$; cohort 2: $t(8) = 0.846$, $p = .422$) (Dienst Uitvoering Onderwijs, 2016, 2017).

An overview of the school and student sample characteristics at baseline is presented in Table 1. Because some initial differences were observed between conditions in terms of pretest scores, gender and background, for the analyses propensity score weighting was applied to better balance the conditions. Separate analyses were conducted for the two cohorts, because we expected the level of implementation of SfA to be higher in the second year, as teachers became more familiar with the program. This might result in greater effects of the program in the second cohort.

[5] To illustrate how these weights are used: For example, in a class of 24 students, 8 have no weight, 8 have a 0.3 wt, and 8 have a 1.2 wt. Therefore the school receives funding for 8 * (1) + 8 * (1 +0.3) + 8 * (1 +1.2) = 36 students, instead of 24.

2.5. Measures

All tests used in this study are standardized reading tests developed by the Dutch Central Institute for Test Development (CITO). The quality of the tests is assessed by the Dutch Committee on Tests and Testing (COTAN), which is an independent committee of test construction experts who review psychological and educational tests. Each test is rated on six criteria: theoretical basis of the test construction, quality of the test materials, quality of the test manual, norms, reliability, and construct validity. Furthermore, criterion validity is assessed for tests with a predictive function. The review system used by COTAN is described in detail by Evers et al. (2015).

The first pretest, Screeningsinstrument Beginnende Geletterdheid (SBG), is a digital test that consists of two subtests. One (called SBG-1 in this study) measures phonological awareness, the other (SBG-2) measures letter knowledge. Both constructs are often used to identify students at risk (e.g. Quint et al., 2015; Zijlstra, 2015), as they are important predictors of future reading performance (Ehri et al., 2001; Van Bergen et al., 2014). The SBG is rated as sufficient/good on all criteria (COTAN, 2011a, 2011b). Reliability for SBG-1 was estimated at .90 using a ρ coefficient. For SBG-2, a Cronbach's alpha of .91 is reported (Vloedgraven et al., 2011). The second pretest, Taal voor Kleuters (TVK), measures vocabulary, critical listening skills, auditory processing skills, and knowledge of written text. TVK was also rated as sufficient/good on all criteria that are assessed (COTAN, 2011a, 2011b). The estimated MAcc reliability coefficient is .87 (Lansink & Hemker, 2010). Criterion validity is not evaluated for the TVK (COTAN, 2011a, 2011b).

We made use of several posttests that measure different subskills of reading. The SBG was administered again at the end of Grade 1, though there is a small difference with the kindergarten version with regard to the second subtest: students in Grade 1 have to sound out letters to a test leader ('productive letter knowledge') instead of recognizing them ('receptive letter knowledge'). Reliability of the SBG-2 posttest was estimated at .70, using the test-retest coefficient. The ρ coefficient of the SBG-1 posttest is .79 (Vloedgraven et al., 2011). The other tests that were administered at the end of Grade 1 were Analyse van Individualiseringsvormen (AVI), measuring text reading skills; Drie-Minuten-Toets (DMT), measuring word reading skills; and Begrijpend Lezen (BL), measuring reading comprehension. The AVI and DMT were both rated as good on all criteria that were assessed (i.e. everything but criterion validity) (COTAN, 2010a, 2010b). Reported Cronbach's alpha coefficients were above .94 for the used forms of AVI, and above .92 for the used forms of DMT (Krom et al., 2010). The newest version (3.0) of the BL test that was used by the participating schools is not yet evaluated by COTAN. However, other reports demonstrate good reliability and validity of the BL test. The estimated MAcc reliability coefficient is .92 (Expertgroep Toetsen PO, 2016; Jolink et al., 2015).

School records about tutoring practices at the SfA schools were also collected. At the end of the school year, SfA facilitators were asked to provide us with an overview of the students who received tutoring. Furthermore, to be able to answer the second research question about the relationship between the intensity of the tutoring intervention and reading achievement, we asked SfA facilitators to share (per tutored student) the duration of the tutoring intervention, the number of sessions per week, and the average duration of the sessions. For the analyses, this was converted into one 'dosage' variable: the total number of hours the student was tutored. Although it might be possible that tutoring in pairs or small groups reduces effective learning time because tutors have to divide their attention over multiple students, we were not able to take this into account in the analyses.

2.6. Procedure

Active informed consent for participation in the research project was obtained from the parents or legal guardians of the students. The procedures were approved by the ethics committee of the Department of

Table 2
Cohort 1 ($n = 148$): Multilevel models predicting reading achievement at the end of Grade 1.

	Reading outcome									
	SBG-1 (phonological awareness)		SBG-2 (letter knowledge)		BL (reading comprehension)		DMT (word reading)		AVI (text reading)	
Fixed Part	B	S.E.	B	S.E.	B	S.E.	B	S.E.	B	S.E.
Intercept	-1.327	1.688	1.078	1.270	-1.927	1.438	-0.498	1.466	-0.969	4.075
Score on SBG-1 pretest	0.165	0.092	-0.016	0.101	0.294*	0.096	0.238 *	0.112	0.216	0.330
Score on SBG-2 pretest	0.247*	0.087	0.318*	0.083	0.343*	0.079	0.475*	0.091	1.321*	0.370
Score on TVK pretest	0.180	0.114	-0.193	0.114	0.040	0.165	-0.056	0.115	-0.496	0.588
Scoring low on only SBG pretest[a]	-0.081	0.219	0.193	0.342	0.349	0.302	0.104	0.243	1.183	1.042
Scoring low on only TVK pretest[a]	-0.541	0.351	0.022	0.422	-0.797*	0.230	-0.787*	0.295	-1.088	1.046
Gender (girl)[a]	0.108	0.149	-0.302	0.188	-0.066	0.176	-0.528*	0.147	-1.882*	0.622
Age	0.018	0.023	-0.015	0.016	0.022	0.018	0.009	0.019	0.014	0.052
Disadvantaged background 0.3[a]	-0.054	0.224	-0.158	0.181	-0.183	0.204	-0.389	0.202	-0.161	0.815
Disadvantaged background 1.2[a]	0.504	0.293	0.065	0.333	0.204	0.333	0.092	0.317	-0.325	0.972
Success for All[a]	0.139	0.267	0.076	0.240	0.296*	0.139	0.124	0.176	1.312	0.894

Random Part	Variance	S.E.	Variance	S.E.	Variance	S.E.	Variance	S.E.	Variance	S.E.
Level-two (classes)	0.230	0.085	0.046	0.056	0.002	0.017	0.074	0.055	1.365	1.521
Level-one (students)	1.244	0.158	1.425	0.159	1.333	0.237	1.165	0.178	n.a.	n.a.

*$p < .05$
aRespectively 'scoring low on both pretests', 'boy', 'no disadvantaged background', and 'control' were the reference categories.

Pedagogical and Educational Sciences (University of Groningen). All tests were taken at the end of the school year, between late May and early July. The SBG was administered by trained test leaders (undergraduate students), outside of the classroom. During the digital subtests, one test leader supervised up to three students at a time (depending on the number of available computers). The productive letter knowledge subtest was administered one-to-one. The administration of the other (nondigital) tests was the responsibility of the schools, as it was part of their regular testing schedule. The BL and TVK tests were taken in the regular whole-class setting. The assignments of the TVK (kindergarten) test were read out loud by the classroom teacher, after which children had to underline the correct answer in their booklets. AVI and DMT were administered in a one-to-one setting. Students were tested by either their classroom teacher or another staff member from their school.

2.7. Analysis

At first, missing data was imputed using the *mice* package (Van Buuren & Groothuis-Oudshoorn, 2011), version 3.9.0, in R statistical software (R Core Team, 2020). For each cohort, five imputed datasets were created (multiple imputation). Percentages of imputed data per variable are presented in Appendix B. The score of one student from cohort 2 on the SBG-2 posttest was extremely high and probably caused by a processing error. This outlier considerably impacted the models (influential case) and therefore the score was removed from the original dataset and thereafter imputed.

Secondly, propensity score weighting was applied in both cohorts to reduce the chance that differences in achievements between conditions were due to initial differences between students. The pretest scores, age, gender and having a disadvantaged background were used to estimate a propensity score for each student. Propensity score weighting was done in R statistical software as well, using the *twang* package (Griffin et al., 2014), version 1.6.

For answering the first research question about the effects of SfA on the reading achievement of students at risk of reading problems, a two-level univariate multilevel model was estimated for each outcome measure. To facilitate comparisons between models for different outcome measures, test scores were standardized. Multilevel analyses (Snijders & Bosker, 2012) were performed using MLwiN software (Rasbash et al., 2000), version 3.04. The analyses were performed separately for each dataset with imputed data, and thereafter the results were combined using formulas of Rubin and Schenker (1986). These combined models are presented in the Results section. Students (level 1) were nested in classes (level 2). The number of schools ($n = 10$) was too small to take the school level into account in the analyses. Because of the relatively small number of classes ($n = 19$), RIGLS estimation was used. For each outcome measure, the same model was fitted, containing the condition variable and covariates. Because initial differences between conditions were still present to some extent after propensity score weighting, we decided to apply a doubly robust correction for initial differences by including the pretest scores and student characteristics (age, gender and disadvantaged background) as covariates in the models. Furthermore, we included a covariate indicating if the student was initially scoring low on both pretests (SBG and TVK), or on only one of them.

The outcome variables SBG-1 (phonological awareness), SBG-2 (letter knowledge), DMT (word reading) and BL (reading comprehension) were continuous. Because the outcome variable AVI (text reading) was dichotomous (0 = not reading on grade level, 1 = reading on grade level), we fitted multilevel logistic regression models for this variable. Propensity score weighting was not applied in the AVI analyses, as this is not recommended for discrete response models (Pillinger, 2011). For estimating the AVI models, we used Markov Chain Monte Carlo (MCMC) estimation with a burn-in length of 20,000 and a monitoring chain length of 50,000.

Effect sizes (ES) (Cohen's *d*) for the continuous outcomes were calculated by dividing regression coefficients of condition by the square root of the sum of level-two (classes) and level-one (students) variances of the final models. For the dichotomous outcomes, the procedure described by Lipsey and Wilson (2001) was followed (p. 187–188), resulting in estimated effect sizes that are equivalent to Cohen's *d*. Because we used five imputed datasets, the proportions of students reading at grade level in both conditions were calculated for each dataset, and the means of the five proportions were used for calculating the effect sizes. For indicating the magnitude of effect sizes, the interpretation of Cohen (1988) was followed. We used a significance level of .05 for all statistical tests. All effects were tested two-sided, except for the effect of condition. This effect was tested one-sided as it was hypothesized that students in the SfA condition would outperform students in the control condition.

For answering the second research question about the relationship between the intensity of the tutoring intervention and reading achievement, we analyzed the data of the group of students at SfA schools who received tutoring (cohort 1: $n = 50$, cohort 2: $n = 44$). For the multilevel analyses, the same procedure as for the first research question was followed, with two exceptions: propensity score weighting was not applied, as all students were in the same condition (SfA), and the 'Tutoring: dosage' (total number of hours the student was tutored)

Table 3
Cohort 2 ($n = 151$): Multilevel models predicting reading achievement at the end of Grade 1.

	Reading outcome									
	SBG-1 (phonological awareness)		SBG-2 (letter knowledge)		BL (reading comprehension)		DMT (word reading)		AVI (text reading)	
Fixed Part	B	S.E.	B	S.E.	B	S.E.	B	S.E.	B	S.E.
Intercept	4.705	1.186	1.265	1.586	2.127	1.403	2.382	1.185	9.363	4.802
Score on SBG-1 pretest	0.269 *	0.097	-0.009	0.099	0.101	0.095	0.213	0.109	0.204	0.323
Score on SBG-2 pretest	0.331*	0.097	0.299*	0.073	0.438*	0.092	0.461*	0.079	0.877*	0.395
Score on TVK pretest	0.251*	0.117	0.093	0.155	0.095	0.155	-0.047	0.114	-0.097	0.352
Scoring low on only SBG pretest[a]	-0.142	0.192	-0.105	0.218	0.380	0.247	0.336	0.230	1.087	0.869
Scoring low on only TVK pretest[a]	-0.307	0.455	-0.103	0.337	-0.008	0.417	-0.330	0.246	-0.232	0.964
Gender (girl)[a]	-0.282*	0.130	-0.135	0.166	0.008	0.149	-0.294*	0.109	-0.061	0.486
Age	-0.059*	0.015	-0.019	0.018	-0.031	0.017	-0.032*	0.016	-0.135*	0.065
Disadvantaged background 0.3[a]	0.248	0.270	-0.023	0.275	0.013	0.295	0.018	0.159	0.004	1.148
Disadvantaged background 1.2[a]	-0.137	0.240	-0.343	0.265	-0.035	0.203	0.048	0.227	0.835	0.835
Success for All[a]	0.173	0.134	0.018	0.186	-0.084	0.223	-0.096	0.225	0.235	0.739
Random Part	Variance	S.E.	Variance	S.E.	Variance	S.E.	Variance	S.E.	Variance	S.E.
Level-two (classes)	0.003	0.014	0.047	0.058	0.112	0.060	0.146	0.068	1.425	1.418
Level-one (students)	1.236	0.151	1.451	0.153	1.183	0.200	1.145	0.212	n.a.	n.a.

*$p < .05$
aRespectively 'scoring low on both pretests', 'boy', 'no disadvantaged background', and 'control' were the reference categories.

variable was included as a predictor in the models instead of condition.

3. Results

3.1. Effects of SfA on reading achievement

For cohort 1, the multilevel models predicting reading achievement at the end of Grade 1 are presented in Table 2. Regarding the covariates, some significant effects were found. The SBG-1 (phonological awareness) pretest was a significant predictor of BL (reading comprehension) ($p < .01$) and DMT (word reading) ($p = .03$). The SBG-2 (letter knowledge) pretest turned out to be a significant predictor in all models, all with p-values $< .01$. Students who only scored low on the TVK pretest (i.e., not on the SBG pretests) scored significantly lower on BL ($p < .01$) and DMT ($p < .01$) than students who scored low on both pretests. Furthermore, girls scored significantly lower on DMT ($p < .01$) and AVI ($p < .01$) than boys. A small and significant positive effect of SfA (ES $= +0.26$, $p = .02$) on the BL outcomes at the end of Grade 1 was found. For the other reading outcomes, no significant differences between the SfA condition and the control condition were found. The effects of SfA on SBG-1, SBG-2 and DMT can be considered small and positive (SBG-1: ES $= +0.11$, $p = .30$; SBG-2: ES $= +0.06$, $p = .38$; DMT: ES $= +0.11$, $p = .24$). SfA has a medium effect on AVI (ES $= +0.51$, $p = .07$). The effect is significant at the $p < .10$ level, not at the $p < .05$ level.

For cohort 2, Table 3 displays the multilevel models predicting reading achievement at the end of Grade 1. The SBG-1 pretest was a significant predictor of SBG-1 posttest outcomes ($p < .01$). The SBG-2 pretest was again a significant predictor in all models ($p = .05$ for AVI, $p < .01$ for the other models). The TVK pretest was a significant predictor in the SBG-1 model ($p = .03$). Girls had significantly lower scores than boys on SBG-1 ($p = .03$) and DMT ($p < .01$) outcomes. Age was a significant predictor in the SBG-1 ($p < .01$), DMT ($p = .05$) and AVI ($p = .04$) models. We did not find a significant effect of the SfA intervention on any of the five reading outcomes in cohort 2. For SBG-1, there was a small positive (ES $= +0.16$, $p = .10$) effect. For SBG-2, we found an ES of $+ 0.02$ ($p = .46$). For BL, the effect of SfA was negative, but very small (ES $= -0.07$, $p = .35$) The same holds for DMT (ES $= -0.08$, $p = .33$). Finally, the effect of SfA on AVI was positive and small (ES $= +0.22$, $p = .38$).

3.2. Relationship between the intensity of the SfA tutoring intervention and reading achievement

Of the 89 students at risk of reading problems in the SfA condition in cohort 1, 50 students received tutoring (56%). On average, these students were tutored for 20.1 h in total ($SD = 11.6$). In Table 4, the multilevel models predicting reading achievement of tutored students in cohort 1 are displayed. Neither the covariates nor the intensity of the

Table 4
Cohort 1: Multilevel models predicting reading achievement of tutored students ($n = 50$) at the end of Grade 1.

	Reading outcome									
	SBG-1 (phonological awareness)		SBG-2 (letter knowledge)		BL (reading comprehension)		DMT (word reading)		AVI (text reading)	
Fixed Part	B	S.E.	B	S.E.	B	S.E.	B	S.E.	B	S.E.
Intercept	-1.203	2.434	1.481	2.085	-2.237	2.239	1.759	2.927	30.271	28.832
Score on SBG-1 pretest	0.222	0.194	-0.254	0.161	0.248	0.172	0.078	0.136	-0.412	1.063
Score on SBG-2 pretest	0.030	0.242	0.313	0.210	0.191	0.271	0.179	0.196	1.522	3.058
Score on TVK pretest	-0.122	0.240	-0.119	0.209	0.163	0.264	-0.059	0.254	-9.090	8.897
Scoring low on only SBG pretest[a]	0.001	0.481	0.314	0.452	0.314	0.487	0.075	0.346	11.090	10.956
Scoring low on only TVK pretest[a]	-0.175	0.553	0.733	0.489	-0.233	0.514	-0.137	0.416	-2.572	4.613
Gender (girl)[a]	0.010	0.357	-0.541	0.303	-0.025	0.350	-0.380	0.223	-11.056	8.898
Age	0.020	0.034	-0.025	0.029	0.031	0.031	-0.025	0.042	-0.352	0.387
Disadvantaged background 0.3[a]	-0.515	0.481	0.375	0.435	-0.363	0.487	0.081	0.356	6.212	5.098
Disadvantaged background 1.2[a]	0.136	0.672	-0.929	0.613	0.219	0.625	-0.459	0.638	-6.052	6.805
Tutoring dosage	-0.012	0.014	0.011	0.013	-0.011	0.013	0.000	0.011	-0.209	0.248
Random Part	Variance	S.E.	Variance	S.E.	Variance	S.E.	Variance	S.E.	Variance	S.E.
Level-two (classes)	0.035	0.089	0.000	0.000	0.004	0.034	0.000	0.000	66.295	160.538
Level-one (students)	0.847	0.207	0.696	0.162	0.723	0.153	0.402	0.236	n.a.	n.a.

[a]Respectively 'scoring low on both pretests', 'boy', and 'no disadvantaged background' were the reference categories.

Table 5
Cohort 2: Multilevel models predicting reading achievement of tutored students ($n = 44$) at the end of Grade 1.

	Reading outcome									
	SBG-1 (phonological awareness)		SBG-2 (letter knowledge)		BL (reading comprehension)		DMT (word reading)		AVI (text reading)	
Fixed Part	B	S.E.	B	S.E.	B	S.E.	B	S.E.	B	S.E.
Intercept	1.943	2.605	1.657	3.651	1.674	3.152	0.948	2.174	513.721	533.594
Score on SBG-1 pretest	0.119	0.151	-0.044	0.200	-0.066	0.188	-0.164	0.130	-175.364	113.416
Score on SBG-2 pretest	0.361	0.223	0.422	0.261	0.443 *	0.219	0.204	0.153	34.400	56.832
Score on TVK pretest	0.216	0.171	0.120	0.233	-0.064	0.262	-0.246	0.156	-20.637	62.825
Scoring low on only SBG pretest[a]	-0.082	0.446	-0.488	0.593	0.472	0.628	0.697	0.401	-59.533	145.951
Scoring low on only TVK pretest[a]	-0.324	0.577	-0.842	0.768	-0.190	0.809	0.101	0.448	225.668	148.897
Gender (girl)[a]	0.023	0.327	-0.373	0.434	0.284	0.390	-0.010	0.271	-9.924	77.982
Age	-0.030	0.032	-0.015	0.045	-0.028	0.037	-0.013	0.026	-3.363	5.490
Disadvantaged background 0.3[a]	0.486	0.587	-0.277	0.767	-0.363	0.715	-0.153	0.424	-77.899	167.458
Disadvantaged background 1.2[a]	0.259	0.594	0.462	0.744	-0.086	0.634	-0.344	0.435	138.708	231.217
Tutoring dosage	0.013	0.013	-0.004	0.016	-0.007	0.014	-0.035 *	0.012	-14.882 *	6.218
Random Part	Variance	S.E.	Variance	S.E.	Variance	S.E.	Variance	S.E.	Variance	S.E.
Level-two (classes)	0.148	0.132	0.124	0.187	0.255	0.250	0.473	0.275	150,031.224	165,144.976
Level-one (students)	0.570	0.146	1.115	0.269	0.639	0.197	0.347	0.086	n.a.	n.a.

*$p < .05$.
[a]Respectively 'scoring low on both pretests', 'boy', and 'no disadvantaged background' were the reference categories.

tutoring intervention significantly predicted (at the $p < .05$ level) reading achievement at the end of Grade 1 in these models.

Of the 75 students at risk of reading problems in the SfA condition in cohort 2, 44 students received tutoring (59%). The average dosage of the tutoring intervention these students received was 21.4 h in total ($SD = 12.7$). The multilevel models predicting reading achievement of tutored students in cohort 2 can be found in Table 5. Regarding the covariates, only the SBG-2 pretest turned out to be a significant predictor of BL achievement ($p = .05$). The intensity of the tutoring intervention was negatively associated with DMT achievement ($p < .01$) and AVI achievement ($p = .04$) at the end of Grade 1. No significant effects of tutoring intensity on SBG-1, SBG-2 and BL achievement were found.

4. Discussion

4.1. Interpretation of findings

In this study, the effects of the SfA program on the reading achievement of Dutch first-grade students at risk of reading problems were investigated. Recently it was decided to develop a Dutch version of SfA, because many students in the Netherlands demonstrate insufficient reading skills and little motivation for reading (Mullis et al., 2017; OECD, 2019), and because the gap between disadvantaged and non-disadvantaged students widens (Engzell et al., 2021; Inspectie van het Onderwijs, 2018; Van Huizen, 2018).

Our hypothesis, that students at risk at SfA schools would perform better in reading than students at risk at control schools, can be partially confirmed. We found promising results in the first cohort: a significant effect of SfA on BL (reading comprehension) (ES = +0.26). This is an interesting finding, as in the analyses of the full sample, no significant effect of SfA on reading comprehension was observed (Mullender--Wijnsma et al., 2020). In earlier studies on SfA, stronger effects for low achievers were also observed (Cheung et al., 2021). Furthermore, in the current study SfA seemed to improve AVI (text reading skills) (ES = +0.51) of student at risk, but this effect was statistically only significant at the $p < .10$ level, not at the $p < .05$ level. The positive effect of SfA on text reading skills is in line with the findings for the full sample (Mullender-Wijnsma et al., 2020). On the other three reading outcomes that were analyzed in the current study (SBG-1 (phonological awareness), SBG-2 (letter knowledge), and DMT (word reading)), small, positive, non-significant effects (effect sizes ranging from +0.06 to +0.11) of SfA were found. In the second cohort, the positive significant effect of SfA on BL was not replicated. Also on the other reading outcomes, we observed no significant differences between the SfA condition and the control condition. Effect sizes for the second cohort ranged from − 0.08 to + 0.22. All observed effects can be considered small to medium, according to Cohen (1988). However, using Cohen's rules of thumb for interpreting effect sizes ignores the context, as is pointed out in more detail by Cheung and Slavin (2016) and Hill et al. (2008). When comparing our results to other studies about reading interventions in the early grades (e.g. Gersten et al., 2020; Quint et al., 2015), the effect sizes are quite similar. Furthermore, these small to medium effect sizes can have significant implications for educational practice. For example, an effect of +0.26 on reading comprehension in cohort 1 means that students in the SfA condition had an average score of 108 on the original scale, while students in the control condition had an average score of 99. Although both groups still perform below the population average, this 9-point difference is of practical interest, as the scores are normally distributed with a mean of 117 and a standard deviation of 28 (Jolink et al., 2015). An effect of +0.51 on text reading in cohort 1 means that 68% of the students at risk in the SfA condition was reading on grade level, in contrast to 43% of the students at risk in the control condition, which we consider to be a relevant difference.

Regarding the second research question, exploring the relationship between tutoring intensity and reading outcomes, we found significant negative relationships between tutoring intensity and AVI and DMT achievement in the second cohort. This finding might be partly explained by the idea underlying RTI that the most intensive types of intervention are provided to the most struggling students (Fuchs & Fuchs, 2006). In a recent meta-analysis of Roberts et al. (2022), it was found that the relationship between reading intervention dosage and achievement of students with reading difficulties seems to follow a nonlinear pattern. Although Roberts et al. (2022) observed that effect sizes increased up to a dosage of approximately 40 h, which is substantially higher than the dosage students in our sample received, the targeted students in the included studies might have had milder problems than those in our sample.

4.2. Limitations and directions for future research and development

When interpreting the conclusions from the current study, some limitations should be taken into account. Firstly, power was limited by the relatively small sample sizes. With 19 available classes to assess effects of the SfA program a priori the power is apparently limited, but notice that we were able to find significant effect sizes as small as.26. To detect effect sizes as small as. 20 with a power of .80 in future studies using the design and covariates as employed in our study, the sample should at least consist of around 50 classes (Spybrook et al., 2011). By

implication the corresponding sample size to detect small dosage effects would then increase to around 125 students, with an effective sample, given the hierarchical structure of the data, of around 100. In that case enough power would be assured in future studies.

Secondly, our mixed findings are presumably affected by implementation issues. As the study concerned the first years of implementation of SfA in the Netherlands, the program was still under development, and we therefore consider the findings to be preliminary. Because the translation of a comprehensive school reform program to another context and another language is laborious, it was not possible to introduce all of SfA's components at the same time. Flexible regrouping of students by reading level was not realized yet, while this feature of SfA could be particularly beneficial for students at risk of reading problems. Furthermore, unlike the American version of SfA, there was no kindergarten program developed yet, while this could have been helpful in addressing emerging reading problems early.

The implementation of both the classroom program (Tier 1) and the tutoring intervention (Tier 2 and 3) varied between teachers and between schools, as is described in the treatment fidelity paragraph. Regarding tutoring, schools were given general guidelines, but they had a fair amount of freedom in how to organize it. Furthermore, the content of the sessions was flexible. This can be considered a strength (activities are tailored to students' individual needs), but also a weakness (designing and evaluating tutoring plans is a time-consuming task that requires expertise in children's reading development). Furthermore, because several SfA schools experienced staffing issues, not all students who would be eligible based on their test scores were assigned to tutoring. It is likely that the weakest performing students were given priority for tutoring. So, among the tutored students, there might be a relatively large group that is not very responsive to intervention, i.e. students whose reading problems turn out to be persistent. Therefore, it might have been difficult to detect a positive effect of tutoring intensity in this study. In future studies evaluating the SfA tutoring intervention, we recommend to increase the sample size and to involve a comparable control group of at-risk students receiving another treatment. Given the strong evidence for the effectiveness of a variety of tutoring interventions demonstrated in other, more rigorous studies (Dietrichson et al., 2017; Gersten et al., 2020; Neitzel et al., 2021), we would still advocate for providing tutoring to students at risk of reading problems as part of the SfA program. We recommend to develop a Dutch version of SfA's adaptive computer program *the Lightning Squad* (Success for All Foundation, 2017), which is expected to make the tutoring job easier and which is well suited for use as a small-group Tier 2 intervention. The labor-intensive one-to-one Tier 3 intervention would then only need to be provided to the students most struggling. This way, a more unified RTI model can be implemented in Dutch schools, which eases some of the faced challenges (Fuchs & Fuchs, 2009).

Although implementation problems seem to be common when implementing a CSR program such as SfA (Desimone, 2002; Klingner et al., 2006; Quint et al., 2015), it is likely that a higher level of treatment fidelity to the SfA program would have been achieved if facilitators, teachers and tutors had received more intensive training and coaching (Desimone, 2002; Durlak & DuPre, 2008). Although ongoing professional development is one of the components of SfA, it should be noted that the amount of training of coaching provided at SfA schools in the Netherlands was limited compared to SfA schools in the U.S. This might have affected the quality of the instruction and feedback students received during the reading lessons and tutoring sessions.

Initially, we expected to find greater effects of SfA in the second cohort, assuming that more experience with SfA would lead to better implementation of the program. This did not seem to be the case in our R&D project. Over the years, balancing between fidelity to the program and fit with local needs and resources remained a challenge. The phenomenon that implementation fidelity varies or even deteriorates over time occurs more frequently. After challenges are faced in the initial implementation phase, adaptations to the program are often made (Durlak & DuPre, 2008; Fixsen et al., 2005).

It is difficult to determine how much the daily practices at SfA schools actually differed from the 'business as usual' at the control schools. Based on policy documents and conversations with school staff, we know that additional support/tutoring for struggling readers was also provided at the control schools. Furthermore, both SfA schools and control schools used *Veilig Leren Lezen* for whole-class phonics instruction, which decreases the contrast between the two conditions even more. A recommendation for future research is to collect more detailed information at the classroom and student level at both intervention and control schools (alike Quint et al. (2015)). Important topics to address would be the time spent on reading instruction in general, the quality of instruction and feedback, the implementation of cooperative learning, and the exact amount and types of tutoring struggling readers receive (e.g. distinguishing Tier 2 and Tier 3 intervention). Furthermore, it is recommended that future studies include motivational measures. In particular for struggling readers, it is important to gain insight in the extent to which programs foster feelings of success, because of the reciprocal relationship between intrinsic motivation and reading achievement (Hebbecker et al., 2019).

The current study shows that SfA has potential to benefit students at risk of reading problems in the Netherlands, in particular in terms of improving reading comprehension skills. By broadening the scope of future evaluations, more insight will be obtained in the similarities and differences between conditions, and thus in the added value of SfA. Continuing both the development and research of SfA will hopefully lead to sustainable school improvement and, as a result, to better readers.

Table A1
Percentages of imputed data for Grade 1 analyses.

	Cohort 1 (n = 148)	Cohort 2 (n = 151)
Variable	% of imputations	% of imputations
SBG-1 pretest	0.0	3.3
SBG-2 pretest	1.4	9.3
TVK pretest	13.5	17.9
SBG-1 posttest	13.5	12.6
SBG-2 posttest	12.2	13.9
BL posttest	13.5	14.6
DMT posttest	9.5	6.6
AVI posttest	15.5	33.8
Gender	0.0	0.0
Age	0.0	0.0
Disadvantaged background	0.7	0.0

Declarations of interest

None.

Appendix: Percentages of imputed data

See appendix Table A1.

References

Borman, B., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: a meta-analysis. *Review of Educational Research, 73*(2), 125–230. https://doi.org/10.3102/00346543073002125

Burns, M. K., Appleton, J. J., & Stehouwer, J. D. (2005). Meta-analytic review of responsiveness-to-intervention research: Examining field-based and research-implemented models. *Journal of Psychoeducational Assessment, 23*(4), 381–394. https://doi.org/10.1177/073428290502300406

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

Mullender-Wijnsma, M.J., Veldman, M.A., de Boer, H., van Kuijk, M.F., & Bosker, R.J. (2020). Implementatie en effecten van Success for All in Nederland. [Implementation and effects of Success for All in the Dutch context]. GION onderwijs/onderzoek.Author, 2020.

Centraal Bureau voor de Statistiek, 2016, Inkomen per gemeente en wijk, 2016. [Income per municipality and neighborhood, 2016]. Retrieved March 19, 2020 from https://www.cbs.nl/nl-nl/maatwerk/2019/02//inkomen-per-gemeente-en-wijk-2016.

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

Cheung, A. C., Xie, C., Zhuang, T., Neitzel, A. J., & Slavin, R. E. (2021). Success for All: A quantitative synthesis of U.S. evaluations. *Journal of Research on Educational Effectiveness, 14*(1), 90–115. https://doi.org/10.1080/19345747.2020.1868031

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed..,). Lawrence Erlbaum Associates,.

Cortiella, C., & Horowitz, S. H. (2014). *The State of Learning Disabilities: Facts, Trends and Emerging Issues*. National Center for Learning Disabilities,.

COTAN, 2010a, AVI-toets. [AVI-test]. Retrieved March 19, 2020 from https://www.cotandocumentatie.nl/beoordelingen/b/14564/avi-toets/.

COTAN, 2010b, Drie-Minuten-Toets, DMT. [Three-Minute-Test, DMT]. Retrieved March 19, 2020 from https://www.cotandocumentatie.nl/beoordelingen/b/14566/drie-minuten-toets/.

COTAN, 2011a, Screeningsinstrument Beginnende Geletterdheid groep 2 en 3, SBG. [Screening instrument early literacy for kindergarten and Grade 1, SBG]. Retrieved April 9, 2020 from https://www.cotandocumentatie.nl/beoordelingen/b/14610/screeningsinstrument-beginnende-geletterdheid-groep-2-en-3/.

COTAN, 2011b, Taal voor Kleuters. [Language for Kindergartners]. Retrieved April 9, 2020 from https://www.cotandocumentatie.nl/beoordelingen/b/14597/taal-voor-kleuters/.

CSRQ. (2006). *CSRQ Center Report on Elementary School Comprehensive School Reform Models*. Comprehensive School Reform Quality Center/American Institutes for Research,.

Desimone, L. (2002). How can comprehensive school reform models be successfully implemented. *Review of Educational Research, 72*(3), 433–479. https://doi.org/10.3102/00346543072003433

Dienst Uitvoering Onderwijs, 2016, Leerlingen bo naar gewicht en leeftijd 2016–2017. [Weighted funding and age of primary students 2016–2017]. Retrieved March 10, 2020 from https://duo.nl/open_onderwijsdata/databestanden/po/leerlingen-po/bo-sbo/po-owsoort-leeftijd.jsp.

Dienst Uitvoering Onderwijs, 2017, Leerlingen bo naar gewicht en leeftijd 2017–2018. [Weighted funding and age of primary students 2017–2018]. Retrieved March 10, 2020 from https://duo.nl/open_onderwijsdata/databestanden/po/leerlingen-po/bo-sbo/po-owsoort-leeftijd.jsp.

Deunk, M. I., Jacobse, A. E., de Boer, H., Doolaard, S., & Bosker, R. J. (2018). Effective differentiation Practices: A systematic review and meta-analysis of studies on the cognitive effects of differentiation practices in primary education. *Educational Research Review, 24*, 31–54. doi:10.1016/j.edurev.2018.02.002.

Dietrichson, J., Bøg, M., Filges, T., & Klint Jørgensen, A. M. (2017). Academic interventions for elementary and middle school students with low socioeconomic status: A systematic review and meta-analysis. *Review of Educational Research, 87*(2), 243–282. https://doi.org/10.3102/0034654316687036

Durlak, J. A., & DuPre, E. (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

Education Endowment Foundation. (2017). *Improving Literacy in Key Stage 1*. Education Endowment Foundation,.

Ehri, L. C., Nunes, S. R., Willows, D. M., Schuster, B. V., Yaghoub-Zadeh, Z., & Shanahan, T. (2001). 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

Engzell, P., Frey, A., & Verhagen, M. (2021). Learning loss due to school closures during the COVID-19 pandemic. *PNAS, 118*(17), Article e2022376118. https://doi.org/10.1073/pnas.2022376118

Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2015). COTAN review system for evaluating test quality. Retrieved January 25, 2023 from https://psynip.nl/wp-content/uploads/2022/05/COTAN-review-system-for-evaluating-test-quality.pdf.

Expertgroep Toetsen PO, 2016, Kwaliteitsoordeel Begrijpend Lezen 3.0 (groep 3 en 4). [Quality assessment Reading Comprehension 3.0 (Grade 1 and 2)]. Expertgroep Toetsen PO.

Fixsen, D. L., Naoom, S. F., Blasé, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation Research: A Synthesis of the Literature*. University of South Florida, Louis de la Parte Florida Mental Health Institute, The National Implementation Research Network.,.

Fletcher, J. M., Reid, G. R., Fuchs, L. S., & Barnes, M. A. (2018). *Learning Disabilities, Second Edition: From Identification to Intervention*. Guilford Press,.

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

Fuchs, L. S., & Fuchs, D. (2009). On the importance of a unified model of responsiveness to intervention. *Child Development Perspectives, 3*(1), 41–43. https://doi.org/10.1111%2Fj.1750-8606.2008.00074.x.

Fuchs, L. S., & Vaughn, S. (2012). Responsiveness-to-Intervention: A decade later. *Journal of Learning Disabilities, 45*(3), 195–203. https://doi.org/10.1177/0022219412442150

Gersten, R., Haymond, K., Newman-Gonchar, R., Dimino, J., & Jayanthi, M. (2020). Meta-analysis of the impact of reading interventions for students in the primary grades. *Journal of Research on Educational Effectiveness, 13*(2), 401–427. https://doi.org/10.1080/19345747.2019.1689591

Gijsel, M., Scheltinga, F., van Druenen, M., & Verhoeven, L. (2011). *Protocol leesproblemen en dyslexie groep 3. [Protocol reading problems and dyslexia for Grade 1]*. Expertisecentrum Nederlands,.

Griffin, B. A., Ridgeway, G., Morral, A. R., Burgette, L. F., Martin, C., Almirall, D., Ramchand, R., Jaycox, L. H., & McCaffrey, D. F. (2014). *Toolkit for Weighting and Analysis of Nonequivalent Groups (TWANG)*. RAND Corporation,.

Hagans, K. S., & Good, R. H. (2013). Decreasing reading differences in children from disadvantaged backgrounds: The effects of an early literacy intervention. *Contemporary School Psychology, 17*(1), 103–117. https://doi.org/10.1007/BF03340992

Hammerstein, S., König, C., Dreisörner, T., & Frey, A. (2021). Effects of COVID-19-Related School Closures on Student Achievement - A Systematic Review. *Front Psychol 12.* , Article 746289. https://doi.org/10.3389/fpsyg.2021.746289

Hebbecker, K., Förster, N., & Souvignier, E. (2019). Reciprocal effects between reading achievement and intrinsic and extrinsic reading motivation. *Scientific Studies of Reading, 23*(5), 419–436. https://doi.org/10.1080/10888438.2019.1598413

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*(3), 172–177. https://doi.org/10.1111/j.1750-8606.2008.00061.x

Hoover, J. (2010). Special education eligibility decision making in Response to Intervention models. *Theory into Practice, 49*(4), 289–296. https://doi.org/10.1080/00405841.2010.510752

Houtveen, T., & van de Grift, W. (2007). Reading instruction for struggling learners. *Journal of Education for Students Placed at Risk, 12*(4), 405–424. https://doi.org/10.1080/10824660701762001

Hughes, C. A., & Dexter, D. D. (2011). Response to Intervention: A research-based summary. *Theory into Practice, 50*(1), 4–11. https://doi.org/10.1080/00405841.2011.534909

Inspectie van het Onderwijs, 2018, De staat van het onderwijs: Onderwijsverslag over 2016/2017. [The state of education: Education report for 2016/2017]. Inspectie van het Onderwijs.

Jolink, A., Tomesen, M., Hilte, M., Weekers, A., & Engelen, R. (2015). Wetenschappelijke verantwoording Begrijpend Lezen 3.0 voor groep 3. [Scientific justification reading comprehension 3.0 for Grade 1]. Cito.

Klingner, J., Cramer, E., & Harry, B. (2006). Challenges in the implementation of Success for All in four high-need urban schools. *Elementary School Journal, 106*(4), 333–350. https://doi.org/10.1086/503635

Krom, R. , Jongen, I. , Verhelst, N. , Kamphuis, F. , & Kleintjes, F., 2010, Wetenschappelijke verantwoording DMT en AVI. [Scientific justification DMT and AVI]. Cito.

Lansink, N. & Hemker, B. (2010). Wetenschappelijke verantwoording van de toetsen Taal voor Kleuters voor groep 1 en 2 uit het Cito Volgsysteem primair onderwijs. [Scientific justification of the Language for Kindergartners tests from the Cito monitoring system for primary schools]. Cito.

Lexima, 2019, Bouw! ter voorkoming van leesproblemen. [Bouw! to prevent reading problems]. Retrieved April 2, 2020 from https://www.lexima.nl/preventie/bouw.

Lexima, 2020, BLOON Spelling. Retrieved April 2, 2020 from https://www.lexima.nl/remedieren/bloon.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical Meta-analysis*. SAGE Publications, Inc,.

Mullis, I.V.S., Martin, M.O., Foy, P., & Hooper, M. (2017). PIRLS 2016 international results in reading. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

National Reading Panel. (2000). *Teaching Children to Read: An Evidence-based Assessment of the Scientific Research Literature on Reading and its Implications for Reading Instruction*. National Institute of Child Health and Human Development,.

Neitzel, A. J., Lake, C., Pellegrini, M., & Slavin, R. E. (2021). A synthesis of quantitative research on programs for struggling readers in elementary schools. *Reading Research Quarterly, 57*(1), 149–179. https://doi.org/10.1002/rrq.379

O'Connor, R. E., & Vadasy, P. F. (2011). *Handbook of Reading Interventions*. The Guilford Press,.

OECD. (2016). *Low-performing students: Why they fall behind and how to help them succeed*. PISA: OECD Publishing,.

OECD. (2019). *PISA 2018 Results (volume I): What Students Know and Can Do*. OECD Publishing. https://doi.org/10.1787/5f07c754-en

Pillinger, R. (2011). *Weighting in MLwiN. Centre for Multilevel Modelling*. University of Bristol.

Quint, J., Zhu, P., Balu, R., Rappaport, S., & DeLaurentis, M. (2015). *Scaling up the Success for All Model of School Reform: Final Report from the Investing In Innovation (i3) Evaluation*. MDRC,.

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing,. ⟨https://www.R-project.org/⟩.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., & Healy, M. (2000). *A user's guide to MLwiN*. Centre for Multilevel Modeling,.

Reschly, A. L. (2010). Reading and school completion: Critical connections and Matthew effects. *Reading & Writing Quarterly, 26*(1), 67–90. https://doi.org/10.1080/10573560903397023

Roberts, G. J., Dumas, D. G., McNeish, D., & Coté, B. (2022). Understanding the dynamics of dosage response: a nonlinear meta-analysis of recent reading interventions. *Review of Educational Research, 92*(2), 209–248. https://doi.org/10.3102/00346543211051423

Rubin, D. B., & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association, 81*, 366–374. https://doi.org/10.1080/01621459.1986.10478280

Slavin, R.E., Madden, N.A., Chambers, B., & Haxby, B. (2009). 2 million children: Success for All (2nd ed.). Corwin Press.

Smits, A., & Braams, T. (2006). *Dyslectische kinderen leren lezen: Individuele, groepsgewijze en klassikale werkvormen voor de behandeling van leesproblemen. [Learning dyslexic children to read: Individual, small-group and whole-class practices for treating reading problems]*. Boom.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Sage.

Snow, C. E., Burns, S. M., & Griffin, P. (1998). *Preventing Reading Difficulties in Young Children*. National Academy Press,.

Spybrook, J., Bloom, H., Congdon, R., Hill, C., Martinez, A., & Raudenbush, S. (2011). Optimal design plus empirical evidence: documentation for the "Optimal Design" software. Western Michigan University.

Success for All Foundation, 2017, Tutoring with the Lightning Squad software user's guide. Success for All Foundation.

Van Bergen, E., De Jong, P. F., Maassen, B. A. M., & van der Leij, A. (2014). The effect of parents' literacy skills and children's preliteracy skills on the risk of dyslexia. *Journal of Abnormal Child Psychology, 42*(7), 1187–1200. https://doi.org/10.1007/s10802-014-9858-9

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Van Huizen, T. (2018). The Evolution of Achievement Gaps from Early Childhood to Adolescence in the Netherlands. In G. Passaretta, & J. Skopek (Eds.), *Roots and Development of Achievement Gaps. A Longitudinal Assessment in Selected European Countries* (pp. *50–87*). ISOTIS Report (D 1.3), Trinity College Dublin.

Vloedgraven, J., Keuning, J., & Verhoeven, L. (2011). Wetenschappelijke verantwoording Screeningsinstrument Beginnende Geletterdheid voor groep 2 en 3. [Scientific justification Screening instrument early literacy for kindergarten and Grade 1]. Cito.

Van Kuijk, M. F., Mullender-Wijnsma, M. J., & Bosker, R. J. (2021). A Systematic Review of Studies Addressing the Implementation of the Evidence-Based Whole-School Reform "Success for All. *ECNU Review of Education, 4*(1), 128–163. https://doi.org/10.1177/2096531120961521.

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

Zijlstra, A. H. (2015). *Early Grade Learning: The Role of Teacher-child Interaction and Tutor-assisted Intervention (Doctoral dissertation)*. Universiteit van Amsterdam,.

Zwijsen, 2003, Veilig Leren Lezen, 2e maanversie. [Learning to read safely, 2nd edition]. Zwijsen.