WORKING PAPER 2019:7

A multi-sensory tutoring program for students at-risk of reading difficulties

Evidence from a randomized field experiment

Martin Bøg
Jens Dietrichson
Anna Aldenius

IFAU
INSTITUTE FOR
EVALUATION OF
LABOUR MARKET AND
EDUCATION POLICY

The Institute for Evaluation of Labour Market and Education Policy (IFAU) is a research institute under the Swedish Ministry of Employment, situated in Uppsala.

IFAU's objective is to promote, support and carry out scientific evaluations. The assignment includes: the effects of labour market and educational policies, studies of the functioning of the labour market and the labour market effects of social insurance policies. IFAU shall also disseminate its results so that they become accessible to different interested parties in Sweden and abroad.

Papers published in the Working Paper Series should, according to the IFAU policy, have been discussed at seminars held at IFAU and at least one other academic forum, and have been read by one external and one internal referee. They need not, however, have undergone the standard scrutiny for publication in a scientific journal. The purpose of the Working Paper Series is to provide a factual basis for public policy and the public policy discussion.

More information about IFAU and the institute's publications can be found on the website www.ifau.se

ISSN 1651-1166

A multi-sensory tutoring program for students at-risk of reading difficulties[a]

Evidence from a randomized field experiment

by

Martin Bøg[b], Jens Dietrichson[c], Anna Aldenius[d]

April 01, 2019

Abstract

Although reading is a fundamental skill, many students leave school without being proficient readers. We examine a literacy program targeting students most at-risk of reading difficulties in kindergarten and first grade. The program includes multi-sensory learning methods, which focus on phonological awareness and phonics and are delivered in a one-to-one or one-to-two tutoring setting. Using a randomized field experiment with 161 students in 12 Swedish schools, we find large positive effects on our two primary outcomes measures: a standardized test of decoding and a standardized test of letter knowledge. We also find positive effects on measures of phonological awareness and self-efficacy and small and statistically insignificant effects on measures of enjoyment and motivation. The program compares favorably to similar programs in terms of cost-effectiveness.

Keywords: phonological awareness, phonics, tutoring, multi-sensory, kindergarten, first grade, Sweden
JEL-codes: I00, I20, I24, I3, J24, Z18

[a] This study would not have been possible without the help and dedication of teachers and staff at the participating schools and municipalities, and we are very grateful for their efforts. We are also indebted to Åsa Malmberg and Christina Rick for helping us conduct the tests. We thank Mimmi Råback for providing us with background data. We are grateful to Trine Filges, Caroline Hall, Mette Kjærgaard Thomsen, Björn Öckert, and seminar participants at the TrygFondens Center for Child Research, the Institute for Evaluation of Labour Market and Education Policy (IFAU), and Copenhagen Education Network Workshop for helpful comments. All errors are our own. IFAU grant 198/2016 is gratefully acknowledged. Declaration of interest: The third author developed the program under study, which is commercially available for schools. The analysis of the data was carried out independently by the first two authors.

[b] Bøg: Lundbeck A/S, Ottiliavej 9, DK 2500, Valby, Denmark; martin.bog@gmail.com

[c] Dietrichson: Corresponding author. VIVE–The Danish Center for Social Science Research, Herluf Trolles gade 11, DK 1052 Copenhagen, Denmark; jsd@vive.dk

[d] Isaksson: S:t Jörgens plats 8, 252 20 Helsingborg, Sweden; anna@lararkonsult.se

Table of contents

1 Introduction 3
1.1 The program 7

2 Methods 9
2.1 Outcome measures 9
2.2 Participants 11
2.3 Control group instruction 18
2.4 Implementation fidelity 19
2.5 Analysis plan 19

3 Results 26
3.1 Pre- and post-test distributions of the outcome variables 26
3.2 Main results 28
3.3 Sensitivity analyses 29
3.4 Exploratory analyses of the short-term effects 31
3.5 Costs and cost-effectiveness 32
3.6 Results at follow-up 33

4 Discussion 35
4.1 Effect sizes and cost-effectiveness in perspective 35
4.2 Study design and effect sizes 37
4.3 Limitations 38

5 Conclusions 40

6 References 41

7 Appendix 52
7.1 Additional descriptive statistics 52
7.2 Sensitivity analyses of the short-term effects 53
7.3 Additional exploratory analyses of the short-term effects 56
7.4 Sensitivity analyses for the effects at follow-up 57

1 Introduction

Being able to read well is a fundamental skill and reading skills are highly correlated with labor market outcomes (e.g., OECD, 2016a). Yet on average around 20 percent of 15-year-olds are not proficient readers in the OECD countries, and all countries have a substantial minority of students who are not proficient (OECD, 2016b). Among adults, a similar share is still classified as poor readers (OECD, 2016a). Finding cost-effective methods to help students at-risk of reading difficulties should therefore be a key focus area for researchers and policy makers.

The previous literature provides reasons to be optimistic about improving this situation. Reviews comparing the effects of targeted (selected or indicated) interventions for at-risk groups suggest that some types of interventions can meaningfully improve literacy skills (e.g., Slavin et al., 2011; Dietrichson et al., 2017). Comparing instructional methods, the effect sizes are typically largest for tutoring, that is, intensive, small-group instruction by adults (ibid.).[1] As for content, phonological awareness and phonics are important building blocks of children's early literacy development (e.g., Bus & van IJzendoorn, 1999; Snow et al., 1998; Ehri et al., 2001; Snowling & Hulme, 2011; Statens beredning för medicinsk utvärdering, 2014), and explicit instruction in these areas may be especially important for children at-risk of reading difficulties (e.g., Lundberg & Hoien, 1996; Machin et al., 2018). Phonological awareness allows children to distinguish phonemes, the distinct units of sound that are the smallest building blocks of language and combine them into larger units such as syllables. Phonics training gives children their knowledge of letters and the ability to match letters or letter patterns with sounds.

However, important pieces of evidence are missing. There is much variation in effect sizes, even within categories of seemingly similar interventions, and most of this variation cannot be explained by observable intervention characteristics (e.g., Dietrichson et al., 2017). Furthermore, few studies have evaluated long-term effects and cost-effectiveness.[2]

[1] Tutoring in varying forms has also received support in the reviews of Cohen et al. (1982), Elbaum et al. (2000), Ritter et al. (2006), and Fryer (2017). Effective whole-school reform concepts often employ tutoring as one of their components (e.g., Borman et al., 2003). High-dosage tutoring was the only component significantly associated with better results in a meta-study of so-called "No excuses" charter schools, which have shown promising results for at-risk students (Chabrier et al., 2016).

[2] For example, studies of interventions targeting low socioeconomic status students rarely include information about costs (Dietrichson et al., 2017). Suggate (2016) reviews longer-term effects of reading interventions, but most included studies evaluate effects less than one year after the end of the intervention. Really long-term studies of the effects on for instance employment and earnings, which exist for targeted preschool programs (e.g. Heckman et al., 2010, Gertler et al., 2014) and class size interventions (Chetty et al., 2011; Fredriksson et al., 2013), are to the best of our knowledge entirely absent from the literature on targeted school-interventions for at-risk readers.

As the short-term effect sizes of targeted interventions are reasonably similar throughout primary and secondary school, the optimal time for intervention remains an open question.[3] In sum, there is still much to learn about the design of effective interventions.

We use a randomized field experiment with 161 students in 12 Swedish schools to examine the short-term impact of a literacy program—*Läsklar* ("ready-to-read")—developed by the third author. The program, which targets the students most at-risk of reading difficulties in kindergarten and first grade, combines three main components: (1) it uses multi-sensory methods aimed at activating many senses and thereby enhancing memory formation and retrieval; (2) it focuses instruction on phonological awareness and phonics; and (3) students are tutored one-to-one or one-to-two by teachers.

Students train with the program in kindergarten and first grade, which may have at least three important advantages. First, as students who cannot read after first grade are highly overrepresented among poor readers throughout compulsory school (e.g., Francis et al., 1996; Denton et al., 2004), early interventions may prevent students from experiencing long periods of school failure.

Second, some studies suggest that children start seeing outcomes as diagnostic of their ability around the start of primary school (e.g., Nicholls, 1978; Butler, 1999; Muenks & Miele, 2017). Furthermore, children are seldom confronted by situations where their academic achievements are systematically compared to their peers before the start of primary school (Poskiparta et al., 2003). Interventions implemented in kindergarten and first grade may therefore improve at-risk students' skills before reading difficulties become an entrenched part of the students' view of their own ability.

Third, if later interventions have to tackle both reading difficulties and motivational-emotional problems,[4] they may need to be more comprehensive and more expensive to succeed (Cook et al., 2014; Vaughn et al., 2015).[5] Furthermore, if skills beget skills, successful early interventions may be important for the success of later interventions,

[3] Average effect sizes of interventions for students with low socioeconomic status do not differ much between elementary and middle school in Dietrichson et al. (2017), and they are in turn similar to average effect sizes of interventions for more general groups of at-risk students in grade 7-12 (Dietrichson et al., 2019). See Hill et al. (2008) for similar results using more diverse student populations.

[4] Reading difficulties have been linked to the development of motivational-emotional vulnerability (Poskiparta et al., 2003) and reduced self-regulation (Connor et al., 2016), and there is a strong association between the development of academic achievement and school motivation (e.g., Archambault et al., 2010; Taylor et al., 2014). Lovett et al. (2017) find larger effects of the same intervention in first grade compared to second and third grade.

[5] In line with this reasoning, targeted interventions typically include more sessions and more intervention hours, and go on for more weeks in higher grades compared to lower (e.g., Dietrichson et al., 2017, 2019).

(e.g., Stanovich, 1986; Cunha & Heckman, 2007). Thus, early interventions have a higher cost-effectiveness potential.

We find positive and statistically significant effects on our two pre-registered primary outcome measures, a standardized test of decoding and a standardized test of letter knowledge. We also find positive and significant effects on measures of phonological awareness and self-efficacy. The effects are insignificant on measures of enjoyment and motivation, but both measures have high risk of ceiling effects. Almost all students are highly motivated and think it is going to be fun to learn to read at pre-, post-, and follow-up tests. These results support the idea that we implemented the intervention before students have become de-motivated.

The statistically significant effects are large. They are for example two to four times larger than the 0.25 standard deviations considered substantial effects by What Works Clearinghouse (2014). The effects are also large in comparison to average effect sizes obtained in other interventions, including tutoring interventions. Furthermore, the program compares favorably to similar programs in terms of cost-effectiveness. We discuss these comparisons at length in the penultimate section.

We contribute to the literature on targeted reading interventions for at-risk students in the following ways. First, we examine a program that adds multi-sensory methods to the more well studied intervention components of tutoring, phonological awareness, and phonics. Neuroscientific evidence suggests that multi-sensory stimulus engage more learning mechanisms than uni-sensory stimulus (Shams & Seitz 2008; Shams et al., 2011) and children perform better on categorization and numerical matching tasks after being exposed to multi-sensory rather than uni-sensory stimulus (Jordan & Baker, 2011; Broadbent et al., 2018). Multi-sensory methods have long been considered good practice when teaching students with dyslexia (Snowling & Hulme, 2011), but systematic evidence comparing multi-sensory methods to other methods for at-risk students is scarce (Broadbent et al., 2018; Snowling & Hulme, 2011).[6]

Second, we implemented the intervention in a way that closely resembles how schools normally would implement the program. We did not monitor sessions or support schools more than would have been the case outside of the study. Furthermore, we did not exclude

[6] See for example Sadoski & Willson (2006) and Ehri et al. (2007) for examples of successful reading interventions that use multi-sensory methods. Note that our experiment examines the impact of the full program, i.e., the joint effect of all its components, and not the components in isolation.

the lowest performing students, but actively sought to include them, regardless of the reason for their at-risk status. This is also in line with how schools typically work, whereas research studies often exclude students who for example are learning the majority language, have diagnoses and impairments, or perform at the bottom of the achievement distribution (e.g., Amendum et al., 2011; Burns et al., 2003; Lovett et al., 2017; Vadasy et al., 2006).

Third, we provide an estimate of the cost-effectiveness of the program and test the effects on measures of students' self-reported attitudes to reading, which there are few examples of in the previous literature.[7] Tutoring by teachers is resource-intensive in general, but we show that it is possible to improve student achievement substantially with relatively small means using a short but intensive intervention.

Lastly, in a literature dominated by studies from English speaking countries–in particular the United States (US)–we provide evidence from a different context. There is no guarantee that interventions that work well in one context also work well in others. Swedish has for example a more transparent orthography than English (but more opaque than e.g., Finnish and German), which could make phonological instruction less important.[8]

We are only aware of three previous studies from the Nordic countries on the effects of reading interventions directly targeting at-risk students in kindergarten and first grade. Our own pilot study of *Läsklar* use a quasi-experimental design to estimate the effects in three Swedish schools. For the comparable at-risk group and tests, we found similar short-term effects on decoding and smaller effects on letter knowledge (Bøg et al., 2018). Elbro and Petersen (2004) find positive effects seven years after an intervention designed to improve phonological awareness for Danish kindergarten children to dyslexic parents. Rogde et al. (2016) find positive, short-term effects on expressive language skills of a small group intervention aimed to improve general language skills for second-language learners in Norwegian kindergartens.[9]

[7] See e.g., Hollands et al. (2013, 2016) and Jacob et al. (2016) for examples of cost-effectiveness analyses and Morgan et al. (2008) and Fives et al. (2013) for exceptions using self-reported attitude measures with first grade children. A reason for the lack of effects on attitudes may be that there are few validated measures for this age group. We were unable to find validated measures in Swedish regarding the concepts of self-efficacy, enjoyment, and motivation, which is a limitation.

[8] See e.g., Furnes and Samuelsson (2010) and Landerl et al. (2013) for discussions.

[9] Fälth et al. (2017) examine a class level intervention focusing on phonological awareness training in four Swedish kindergarten classes. The intervention did not explicitly target at-risk children, but they find positive effects also for this group. Wolff (2011, 2016) study a tutoring intervention with phonological training for students with reading

The next section describes the program. Section 2 provides a description of the data we use and how we measure literacy skills and estimate the effects of the program. Section 3 presents the results. Section 4 compares the results to the earlier literature and discusses some limitations of the study. Section 5 provides concluding remarks.

1.1 The program

Läsklar has three main aims. The first is to make the students realize that words denote both a content and has a form composed of phonemes. The second is to help students grasp the connection between letters and sounds. The third is to show students how to decode simple words (i.e., to translate written words into sounds). In the "simple view of reading", efficient decoding is a main determinant of reading comprehension (e.g., Hoover & Tunmer, 2018).

Students trained together with a tutor either one-to-one or in groups of two (16 students trained in groups of two). The intended frequency and duration of the program was 3-4 times a week for about 8-10 weeks, over a total of 30-35 sessions where each session lasts about 10-15 minutes. The average received number of sessions in the treatment group was 32.6, the median 32, and the minimum and maximum number was 21 and 50, respectively.[10]

The tutors, teachers or special educators at the participating schools, were trained using an online course, which consists of short films where the third author shows how to conduct the sessions. The third author also visited each school one or two times to show in person how to conduct the sessions and answer questions.

The program had three types of sessions. In the first type, the tutor uses clay figurines representing 25 of Swedish language sounds and letters, a box with a compartment for each letter, and a laminated sheet with the fingerspelling alphabet. A session starts with the tutor selecting three figurines. The students' task is to help the figurine to move in to its designated compartment, or "house". To do so, students need to figure out a code. The first part of the code is the sound of first letter in the figurine's name, for example, "ll" in "lion", and the second part consists of fingerspelling the letter. When the students have figured out that the word "lion" begin with an "ll" sound, they should make the sound for

difficulties in third grade. The intervention group had significantly better reading skills in a number of areas immediately after the intervention, and five years later on a decoding test.

[10] We lack information about the number of sessions from eight treated students. In the few cases when the number of sessions was reported as a range, we use the midpoint of that range.

a while, while at the same time fingerspelling the letter. When this is accomplished, the figurine can move into its house. Once all three figurines have moved into their houses, the procedure is repeated once and then the session ends. This usually takes about 10-15 minutes. During the next session, the students start with the previously completed figurines, and then continued with three new ones.

The second type of session starts after the students have learned about 15 of the letter sounds, which typically takes less than a month. The tutor use images representing short words pasted on a white card. The tutor acts as a secretary to the students when they sound out the phonemes in the word, after which the tutor writes the word on the front of the card. Each word-picture card is put into a binder. The binder is gradually filled until it contains about 25 words. Using the binder, the students can practice themselves, as they are able to verify whether they read the word correctly by turning the card. If the students need more practice, the old word-picture cards are tied together into a book and afterwards the students began filling the binder with new words.

The third type of session aims to build decoding speed and fluency by practicing the decoding of short syllables with the help of flashcards with syllables. This type of session is therefore not used with students who do not reach the level where they are able to decode simple words. For the students that do, the task is to decode as many syllables as possible during one minute. The tutor registers the result on a chart and then the same syllables are repeated two more times in the same session. This procedure usually yields a clear improvement, which students can follow on the chart. Making improvements visible for the students usually motivates them to practice more.

The theory of action behind *Läsklar* is that by focusing on phonological awareness and phonics students will understand that the audio stream from the spoken language can be divided into a number of letter sounds, and learn how to differentiate them from one another. The small group instruction allows for an individualized pace and support, and frequent feedback from the tutor. The multi-sensory methods aim to involve as many senses as possible to activate the brain’s learning mechanisms and strengthen memory capacity and retrieval. Placing the figurines into their houses aims to activate the kinesthetic, visual, tactile, and auditory senses, while at the same time creating a little storyline about the figurines moving into their house that may activate the episodic memory (the memory of autobiographical events; e.g., Tulving, 2002). The consistent use of finger-spelling may also activate several senses and the students can use this tool as a memory aid when they proceed to regular text reading. Working with figurines may also strengthen the students' motivation, as they may enjoy helping the figurines find their houses more than just thinking about how the beginning of a word sounds.

2 Methods

This section describes the outcome measures, the participants, the assignment mechanism, the control group instruction, fidelity of implementation, and the analysis plan.

2.1 Outcome measures

Our two primary outcome measures, a standardized test of decoding skills and a standardized test of how many letters the students know, correspond to the primary objectives of *Läsklar*. As a secondary outcome measure, we use a test of phonological awareness, developed by the third author. To examine if the program also affects attitudes to reading, we use three measures related to self-efficacy, enjoyment, and motivation.

We administered all tests pre-randomization (*pre-test*) to both the treatment and control groups and again to all students shortly after the treatment group had trained with program (*post-test*), and then a third time after the control group had also trained with the program (*follow-up*). Schools that implemented the program in kindergarten could start in both fall and spring semesters, and we conducted pre-tests between late August and early March. All schools that implemented in first grade started early in first grade and we conducted pre-tests from late August to early October. The average number of days between pre- and post-tests was 103 and the average number of days between post- and follow-up was 102.[11]

Almost all students were tested by the same tester (a former teacher), who was blind to the treatment status of students. For one testing round at one school, we used another tester, who was also blind to treatment status. The test administrators tested the students individually, which took around 10-15 minutes per student. The total score for each test,

[11] There are no substantial or statistically significant differences between the treatment and control group regarding the number of days between the tests.

occasion, and student plus anonymized student identifiers were transferred to an Excel-file and sent to the researchers.

2.1.1 Decoding

LäSt is a test battery used for identification and diagnosis of basic literacy skills (Elwér et al., 2016). The material includes subtests of decoding skills, spelling and reading comprehension. We use the test of decoding skills which has two parts ("Avkodning ord", parts A and B), where students read short individual words (real words, not pseudowords)[12] of increasing difficulty for 45 seconds in each part, and points are given for each correct word (the maximum is 200).

The test–re-test reliability of both part A and B of the decoding test is 0.93 and the correlations between the two parts and two others commonly used Swedish tests ("Ordkedjor" and "H4") range from 0.59 to 0.88 (Elwér et al., 2016). Moreover, the construction of the test is like for example the commonly used Word Reading Efficiency subtest of the Test of Word Reading Efficiency (Torgesen et al., 1999) where students are asked to read isolated words of increasing difficulty for 45 seconds.

2.1.2 Letter knowledge

We use a diagnostic test from the standardized test material *LäsEttan* (Johansson, 2009) to test how many letters of the alphabet students could name. The letter knowledge subtest takes about 1-2 minutes to complete. The maximum score is 27, one for each letter included in the test (Johansson, 2009). Information about reliability and validity is not included in the test material. The test is similar to for example the Letter Naming Fluency subtest in the Dynamic Indicators of Basic Literacy Skills (e.g., Good et al., 2002), a commonly used test battery in American studies.

2.1.3 Phonological awareness

We use a researcher-developed, short test of phonological awareness that examines if students know the first sound in 10 words represented by a picture of an object or an animal. The student is asked to first identify the depicted animal or object and then to say the sound the word starts with. Each correct answer is worth one point and the maximum

[12] We were recommended by teachers involved in our pilot study (Bøg et al., 2018) to not include a pseudo-word test that is included in *LäSt*. The motivation was that students in Swedish kindergarten have typically not learnt how to read orthographically yet, and it may be an important realization for beginning readers when they understand that the letters they have decoded actually mean something. Using pseudo-words might therefore be confusing for the students.

score is therefore 10. This test was developed by the third author for this intervention with the primary aim of discerning which students who needed of extra help. The test has a low maximum score and as phonological awareness is often a strong focus of regular instruction in Swedish schools, it may, especially in first grade, imply considerable ceiling effects. For these reasons, we consider this test a secondary outcome.

2.1.4 Self-efficacy, enjoyment, and motivation

We use three simple questions posed to the participating students to measure impacts on attitudes and beliefs related to self-efficacy, enjoyment, and motivation (Swedish wording in parentheses):

- How easy or hard do you think it will be to learn how to read? (Hur lätt eller svårt är det att lära sig läsa, tror du?)
- How much fun do you think it will be to learn how to read? (Hur roligt tycker du att det ska bli att lära sig läsa?)
- How much would you like to learn how to read? (Hur gärna vill du lära dig läsa?)

For each question, the test administrator read the question and students indicated their answers on a Visual Analogue Scale (VAS) with the two end points represented by a sad and a happy smiley, respectively. The students were told that the sad smiley represent, e.g., that it will be no fun at all learning to read and the happy smiley that it will be a lot of fun. The students' answers are then transformed into a scale with a minimum of 0 and a maximum of 10.

We are not aware of validated, student-rated tests for these constructs and this age group in Swedish, but similar tests have previously been used in for example Tideman et al. (2011). We wanted to use student-rated tests because teachers are aware of treatment status, and teacher ratings may interact with teacher acceptance of the program.

2.2 Participants

2.2.1 Schools and tutors

We recruited 12 schools from 7 municipalities in the southwestern part of Sweden (Skåne) sequentially during the fall of 2016 and the spring of 2017. The six first schools began implementing the program for their kindergarten students either late in 2016, or during spring of 2017. The six schools recruited last began the implementation in kindergarten/first grade in the fall of 2017.[13] Four schools ran a second round of the program for a new cohort, so there were in total 16 rounds. The last students finished their training with the program in early June 2018.[14]

Table 1 compares the participating schools to the national average across socioeconomic and achievement variables and displays descriptive statistics on tutor characteristics. The participating schools have a lower share of parents with any tertiary education, a higher share of students with a foreign background, a lower share of students that pass all subjects in grade 9, and a lower average score on both the Swedish and mathematics parts of the national tests taken in grade 6. This suggests that the participating schools are on average relatively disadvantaged. There is however in-sample variation, as shown by the relatively large range between the minimum in column (3) and maximum in column (4).

In total 23 teachers/special educators participated in the study as tutors. The tutors were experienced teachers (they have on average worked for 23 years as teachers), but only about a third had previously worked with a similar program (including three that worked with *Läsklar* in the pilot study).

2.2.2 Students

We randomized in total 161 students to either the treatment or the waiting list control group (see Section 2.2.3 for a description of the assignment), 130 in kindergarten and 31 in first grade. We analyze them together in the primary analysis and examine differences in an exploratory analysis.

In the primary analysis, we focus on the sample with complete pre- and post-test scores. Attrition from pre- to post-test was low: three students moved to other schools after randomization (one from the treatment group and two from the control group), one treated student did not have complete pre-test scores, and one control group student was absent at post-test (but present at follow-up). In addition to the students who moved and

[13] Kindergarten (or preschool class, "förskoleklass") is the first year of Swedish primary school. It was not compulsory until the fall of 2018, but attendance was nearly universal already before this decision (Pontoppidan et al., 2018).

[14] Our initial power analysis, based on the effect sizes and correlations between covariates and outcomes in the pilot study and in a recent review (Bøg et al., 2018; Dietrichson et al., 2017), indicated that we needed to recruit 10 schools with around 10-15 students in each to find effect sizes around 0.3-0.4 with 80% power. As 10-15 students turned out to be more than most schools had the resources to work with, we recruited two more schools and allowed schools that wanted to run a second round for a new cohort to do so. Our pre-registered analysis plan contains an updated power analysis using the number of schools and students actually recruited, which implies a minimum detectable effect size of 0.33 (see the American Economic Association's RCT Registry, registration number AEARCTR-0002750).

did not have complete pre-test scores, eight more treatment group students and eight more control group students left the study between the post- and follow-up tests.

Table 1 School and tutor characteristics.

Variable	(1) National average	(2) Sample average	(3) Sample min	(4) Sample max	N
Schools					
Parents with any tertiary education	57%	40%	29%	60%	12
Foreign background	24%	52%	11%	91%	12
Proficiency in all subjects in grade 9	74%	56%	33%	73%	6
National test score in Mathematics in grade 6	12.1	10.4	6.4	13.0	11
National test score in Swedish in grade 6	13.0	12.3	8.5	15.1	10
Tutors					
Teaching experience (years)	-	23	8	42	21
Earlier experience with similar programs	-	33%	0	1	21
Special educator	-	45%	0	1	21

Note: Source for school variables: National Agency for Education (2018a). Statistics for school year 2016/2017 and for the K-6 part of each school, whenever the statistic was available for that part. Foreign background is defined as being born outside of Sweden or having two parents that were born outside of Sweden. Note that the variables Proficiency in all subjects in grade 9, National test score in Swedish in grade 6, and National test score in mathematics in grade 6 have missing values, because e.g., some schools are not K-9 schools, or have too few students taking a certain test. Source for tutor variables: interviews with and a survey of tutors. We were unable to obtain information from two tutors. Special educator denotes either "specialpedagog" or "speciallärare".

Table 2 shows student characteristics for the analysis sample of 156 students at pre- and post-test and 141 students at follow-up.[15] About 45 percent are girls, and 20 percent are first grade students. Participating students are the ones most at-risk of reading difficulties in their respective schools. Most students were not at-risk for a specified reason, although the sample includes students who very recently immigrated, have behavioral problems (e.g., ADHD or concentration difficulties), language difficulties, hearing impairment, cognitive disabilities, or suspected dyslexia.[16] In total, 24 percent of the sample had some form of specific risk.[17]

[15] In the Appendix, we show descriptive statistics for all randomized students. These statistics are nearly identical to the ones in Table 2.

[16] Formal tests of dyslexia are not regularly performed in these schools before grade 4 or 5 (personal communication with teachers). This is in line with general practice in Sweden, where dyslexia diagnoses are usually not given at early ages (Gustafson et al., 2007).

[17] We surveyed schools about the share of participating students that have another language than Swedish as their mother tongue. It was not possible to obtain this information from all schools (three are missing), but among those that answered around 60% of students had different mother tongue.

Table 2 Means and standard deviations at pre-, post-, and follow-up test for the analysis sample

Panel A: Pre-test

	Treatment			Control				
Variable	Mean	SD	n	Mean	SD	n	ES	p
Girl	0.39	0.49	80	0.51	0.50	76	-0.13	0.116
Grade 1	0.20	0.40	80	0.20	0.40	76	0.00	0.967
Specific risk	0.21	0.41	80	0.26	0.44	76	-0.05	0.460
Decoding	0.00	0.00	80	0.00	0.00	76	0.00	1.000
Letter knowledge	5.11	3.71	80	5.11	3.92	76	0.00	0.991
Phonological awareness	4.25	4.17	80	5.17	4.18	76	-0.22	0.171
Self-efficacy	4.87	3.88	80	4.62	3.55	76	0.07	0.675
Enjoyment	8.20	3.24	80	7.35	3.45	76	0.25	0.114
Motivation	8.12	3.41	80	8.27	3.10	76	-0.05	0.773

Panel B: Post-test

	Treatment			Control				
Variable	Mean	SD	n	Mean	SD	n	ES	p
Decoding	6.63	7.47	80	1.28	3.55	76	0.90	0.000
Letter knowledge	17.86	6.29	80	10.58	7.12	76	1.08	0.000
Phonological awareness	9.00	2.08	80	7.25	3.86	76	0.57	0.001
Self-efficacy	6.61	3.45	80	4.66	3.11	76	0.59	0.000
Enjoyment	8.41	3.01	80	8.26	3.06	76	0.05	0.749
Motivation	9.14	2.20	80	8.78	2.35	76	0.16	0.315

Panel C: Follow-up

	Treatment			Control				
Variable	Mean	SD	n	Mean	SD	n	ES	p
Decoding	11.79	11.21	72	9.91	12.50	69	0.16	0.349
Letter knowledge	21.29	5.99	72	20.23	6.22	69	0.17	0.304
Phonological awareness	9.61	1.01	72	9.16	2.27	69	0.26	0.127
Self-efficacy	5.26	3.56	72	6.64	3.08	69	-0.41	0.015
Enjoyment	7.70	3.27	72	8.84	2.30	69	-0.40	0.019
Motivation	8.52	2.87	72	9.14	2.14	69	-0.24	0.147

Note: Mean, standard deviation (SD), sample size (n), and the difference between treatment and control groups expressed as effect sizes (ES), and p-values from a two-sided t-test of equal means/proportions. The effect size is for all variables, except Girl, Grade 1, and Specific risk, Hedges' g; i.e., the difference between treatment and control group in standard deviations, adjusted for the small sample, see Equation (2). The effect sizes for Girl, Grade 1, and Specific risk are expressed as the differences in shares.

2.2.3 Assignment to treatment

We randomly assigned students to treatment and waiting list control groups. We chose a waiting list design for primarily two reasons: First, no student is denied the program. This feature likely made recruitment easier and also made schools more willing to accept randomization and comply with the assignment. Second, a control group (or their parents and teachers) that knows it will get an intervention at a later date may be less likely to seek help elsewhere, compared to a group that knows it does not get anything. Thus, treatment spillover effects may be less likely with a waiting list design.

We used an assignment procedure with three-steps, where students were assigned separately within each participating school (i.e., the randomization was blocked by school). Figure 1 illustrates the procedure and in addition shows the attrition at different stages. We applied the procedure sequentially, that is, as soon as a school had agreed to participate we performed the following steps.

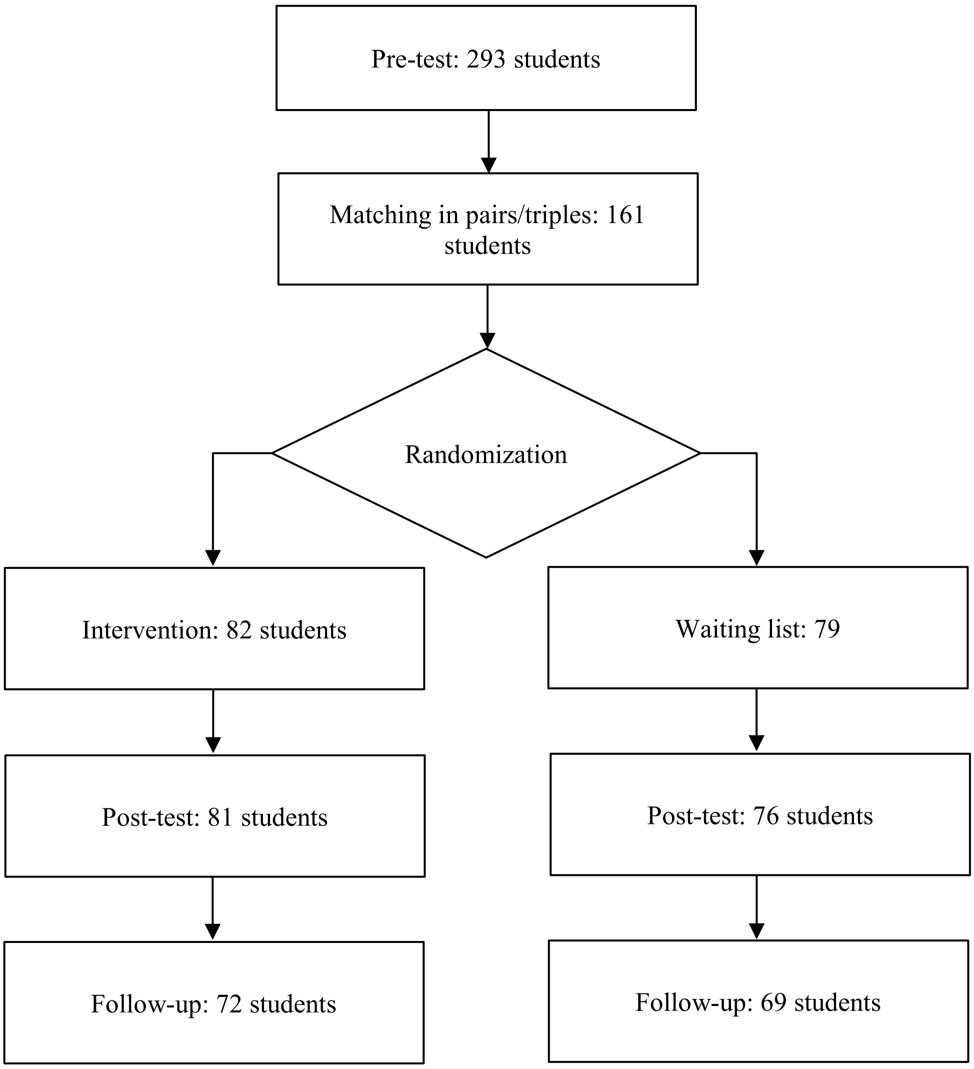
First, teachers and special educators in each school (and round) selected the group of students believed to be most in need of extra help, regardless of the reason for their difficulties. This group was then tested using all outcome measures. Note that a few schools took the opportunity to test all their students, so the number tested at pre-test in Figure 1 is larger than the group believed to be at-risk.

Second, based on the results on the pre-tests and inputs from the third author the school selected the final group that they wanted to include in the intervention. If this group was smaller than the group believed to need extra help (e.g., because of resource constraints), the students with the lowest scores on the letter knowledge and the phonological awareness tests were selected (all students score zero on the decoding test).

Third, within each school, the researchers matched students into pairs and, when the number of students were uneven, one triple. Pre-matching increases the chances of obtaining balanced treatment and control groups and may also improve statistical power when sample sizes are not very large (e.g., Bruhn & McKenzie, 2009; Imai et al., 2009). The matching of students into pairs/triples was done by the nearest neighbor method, where closeness was based on the two primary outcome variables. However, as no students knew how to decode at pre-test, the letter knowledge tests determined the matches in practice. If there was a tie, we used the test of phonological awareness to determine the closest match. One student in each pair, or one or two students in each triple, was then randomly assigned to receive the program in the first period (i.e., was assigned to the treatment group) and the rest were put on a waiting list and received the program in the second period.[18]

[18] At one school's request (for scheduling reasons), we stratified students by class first, and then used the same matching and randomization procedure.

Figure 1 Random assignment to treatment and control groups.



The second author conducted the randomization to treatment and waiting list control groups using the random number generator in Microsoft Excel. Students' identities were anonymized and the only information available at randomization was the pre-test scores.

2.2.4 Treatment and control group balance

Properly conducted randomization ensures that the treatment assignment is statistically independent of student characteristics. However, even proper randomization procedures can produce imbalances between treatment and control groups in finite samples, which may bias statistical tests (e.g., Roberts & Torgerson, 1999).

Panel A in Table 2 shows means and standard deviations of student characteristics and pre-test scores. All students score zero on the decoding test on the first test occasion; i.e., no student could read any words before the intervention. On average, the participants knew around five letters of the alphabet and got around half the sounds correct on the phonological awareness test (five out of a maximum of ten). Most students believed that it would be fun to learn to read and they want to learn: the scores on the enjoyment and motivation measures are around eight out of a maximum of ten. However, more students believe reading will be difficult to learn, the average on the self-efficacy measure is around five (out of ten).

None of the treatment and control group means in Panel A, Table 2 are significantly different from each other ($p > 0.10$). The groups are virtually identical on the two primary outcomes, decoding and letter knowledge. This is not surprising given the matching procedure and because all students score zero on the decoding pre-test. The two largest differences on the pre-tests, for phonological awareness and enjoyment have different signs, which indicates that the differences are not systematic.

Lack of individual significance does not rule out that covariates may jointly predict treatment or outcomes. However, using a logit model with the treatment indicator as the outcome variable and all variables included in Table 2 Panel A, as explanatory variables, we cannot reject the null hypothesis of the coefficients being jointly equal to zero ($\chi^2(8) = 9.23$, $p = 0.323$).[19]

In Table 3, we show the results from another prediction exercise. We use the covariates in Table 2, Panel A, in a linear regression with the post-test scores as the outcome variables and then use the results to calculate predicted means and standard deviations for each outcome in the treatment and control group. The differences in predicted means have different signs, are small for all outcomes, and no means are significantly different from each other.

In sum, pre-treatment characteristics and pre-test scores are not systematically different, do not predict well who ends up in the treatment group, and do neither predict large nor significant differences at post-test. We conclude that the randomization appears to have created a balanced treatment and control group.

[19] One variable is individually significant on the 5 percent level in this specification, Enjoyment ($p = 0.045$).

Table 3 Predicted means and standard deviations in the treatment and control group

	Treatment			Control				
Variable	Mean	SD	N	Mean	SD	n	ES	p
Decoding	4.06	3.33	80	3.97	3.39	76	0.027	0.868
Letter knowledge	14.53	4.68	80	14.08	4.77	76	0.095	0.551
Phonological awareness	8.06	1.62	80	8.24	1.62	76	-0.113	0.480
Self-efficacy	5.65	0.68	80	5.67	0.72	76	-0.035	0.825
Enjoyment	8.39	1.11	80	8.28	0.98	76	0.110	0.492
Motivation	8.94	0.75	80	9.00	0.69	76	-0.082	0.610

Note: Mean, standard deviation (*SD*), sample size (*n*), and the difference between treatment and control groups expressed as effect sizes (*ES*). The *p*-value in the rightmost column is from a *t*-test of different means in the treatment and control group.

2.3 Control group instruction

The waitlist control group received business-as-usual reading instruction. The regular kindergarten instruction in most schools was based on some version of the Bornholm model (e.g., Lundberg et al., 1988; Häggström, 2007). The Bornholm model uses structured metalinguistic games and exercises that aim to help students discover the phonological structure of language. The games include listening to verbal and nonverbal sounds, rhyming games and rhymed stories, segmentation of sentences into words and investigation of word length, clapping hands to mark syllables, recognizing phonemes in the initial position of and within words, and prosodic games (see, e.g., Lundberg et al., 1988). The Bornholm model was always implemented in whole class.

Regular instruction in first grade was based on the national Swedish curriculum (National Agency for Education, 2018b). The curriculum emphasizes among other things alphabet knowledge, letter-sound correspondence, and handwriting. Special education services ("särskilt stöd") are rarely used in first grade in Swedish schools. Only 1.6 percent of students in first grade have an individualized education plan, which is a requisite for receiving special education services, and 0.5 percent receives one-to-one instruction (National Agency for Education, 2018c).

Kindergarten is not covered by the national statistics, but no school reported using any programs specifically targeting at-risk students in kindergarten (before starting to use *Läsklar*). Some schools reported using one-to-one/small group extra instruction for students having reading difficulties in the latter part of first grade. However, all first-grade students in the treatment group trained with the program during the first half of first grade. Therefore, it seems unlikely that the study had negative spillovers to the control group, in the sense that control group students would have gotten special education services or small group instruction had there not been a study.

2.4 Implementation fidelity

We did not monitor the fidelity during the implementation to make the trial as naturalistic as possible. Tutors could contact the third author in case they had questions during the implementation, but this would also have been possible if schools had implemented the program outside of the study (3-5 tutors used this opportunity).

We interviewed a selection of tutors after the follow-up test (we were unable to meet with tutors at five schools). Most schools that had not used *Läsklar* before were not used to tutoring and organizing the sessions was initially seen as difficult, as there were no established routines. However, the schools were able to solve these problems, as shown by the mean number of received sessions being within the range of intended sessions. Very few students received a number of sessions that was substantially below the intended 30-35 (e.g., 3 students received less than 25 sessions).

At one school, both tutors left for work at other schools after the treatment group had received the program. The school was therefore unable to give the program to the waiting list control group and we conducted no follow-up tests. One other school had more students in need of the program than they were able to tutor. We agreed with the school to let the third author tutor eight students. They are included in our preferred estimations, but we show in Section 3.3 that our main results are very similar if we exclude them.

2.5 Analysis plan

The analysis follows our registered analysis plan, which can be found in the American Economic Association's RCT Registry (www.socialscienceregistry.org) with registration number AEARCTR-0002750. Due to the sequential recruitment of schools, the plan was registered after the trial had started but before the researchers performing the analysis (the first and second author) had examined the data. The plan specifies an ordered testing procedure for the primary analysis designed to adjust for multiple hypothesis testing and control the family-wise error rate at a level of 0.05. As we will see, our results in the primary analysis are either statistically significant on a much lower level and would survive a more conservative adjustment (e.g., a Bonferroni-type), or are not statistically significant on conventional levels without adjustment. We therefore refrain from describing the procedure.

2.5.1 Estimation of short-term effects

We start by showing the full pre- and post-test distributions of the outcome variables. We then estimate the short-term effects using linear regressions of the following type:

$$y_{i,post} = \alpha + \beta Treatment_i + \boldsymbol{y}_{i,pre}\boldsymbol{\lambda} + \mu_i + \varepsilon_i \qquad (1)$$

where $y_{i,post}$ is one of our six outcome variables measured at post-test, α is a constant, $Treatment_i$ is an indicator for whether student *i* was randomized to the treatment group or not, $\boldsymbol{y}_{i,pre}$ is a vector of pre-test scores,[20] β and $\boldsymbol{\lambda}$ are parameters to be estimated, μ_i is a pair/triple fixed effect, and ε_i is a random error term. Because we are not interested in more than controlling for the pair/triple fixed effects, we use the *ivreg2* command (Baum et al., 2010), for Stata 14 (StataCorp, 2013) to partial out the fixed effects, which improves efficiency.[21]

Our primary analysis uses the sample of students with a complete set of pre-test scores and at least one post-treatment score. Treatment is not defined by having received training, only by being randomized to treatment. However, all students in the treatment group received some training, and most received close to the intended number of sessions.

We calculate an effect size measure, Hedges' *g*, to compare the magnitude of our effects with other interventions. This measure includes a small sample correction (Hedges, 1981), which is appropriate in our case (the measure is also recommended by e.g., What Works Clearinghouse, 2014). We calculate *g* according to the following formula (Lipsey & Wilson, 2001):

$$g = \left(1 - \frac{3}{4N-9}\right) \times \frac{\hat{\beta}}{SD_{post}};\ SD_{post} = \sqrt{\frac{(n_T-1)s_T^2 + (n_C-1)s_C^2}{N-2}} \qquad (2)$$

where the term in the first parenthesis is the small sample size correction, $N = n_T + n_C$ is the total sample size where n_T is the number of treated students and n_C is the number of control students, $\hat{\beta}$ is a treatment effect estimate and SD_{post} is the weighted unadjusted

[20] We include all pre-test scores except the decoding test, as all children score zero at pre-test.

[21] Our results are not sensitive to this choice or to the inclusion of pair/triple fixed effects, see section 3.3.

post-test standard deviation in the treatment and control group. We use the post-test standard deviation for comparison purposes and because all students score zero at the decoding pre-test (see Panel A, Table 2).[22]

2.5.2 Sensitivity analyses

Our sensitivity analyses include specifications without any covariates, which allows us to include the student without a complete set of pre-test results. We also show results from a specification with pre-tests but without pair/triple fixed effects. Furthermore, for students matched in triples the probability of receiving treatment is not 0.5, and we therefore test if our results are robust to weighting observations with the probability of being treated. As mentioned, we also test if our results are sensitive to excluding students who were tested by a different tester and the few students instructed by the third author.

The random assignment of treatment was done at the individual level and most students received the program individually (one-to-one), but 16 students in the treatment group received the program in instruction groups of two. Weiss et al. (2016) show that models using a cluster-robust variance estimator (or a random group effect) typically overestimate variance in cases when the level of instruction is different from the level of assignment. On the other hand, not adjusting for clustering typically underestimate the variance. As per our analysis plan, we report standard errors that are robust to heterogeneity of unknown form in our primary analysis and cluster the standard errors on the instruction groups (using the cluster-option in Stata) in a sensitivity analysis. The instruction group is then a singleton for control group students, students tutored individually, and groups of two for other tutored students.[23]

We use randomization (or permutation based) inference methods, as described by e.g., Athey and Imbens (2017) and Young (2018), to assess whether our results are sensitive to the assumptions made on the distribution of the standard errors when using regular inference methods. We confine this test to the significant coefficients, as the regular inference procedure is unlikely to have overestimated the standard errors.

[22] It is possible to calculate the small sample correction exactly (see Hedges, 1981) and use the exact pooled standard deviation in the treatment and control groups. However, we want to compare our effect sizes to those reported in related reviews and meta-analyses, which often use this formula and the post-test standard deviation (e.g., Dietrichson et al., 2017). We therefore think the effect sizes are more comparable when calculated in this way.

[23] Because treatment is randomly assigned on the individual level (within each pair/triple), the common environmental factors stemming from treatment and control students being in same “cluster” (e.g., school) should cancel out. Clustering on a higher level, e.g., the pair/triple, the class, or the school would therefore likely overestimate the standard errors for the treatment variable (e.g., Lohr et al., 2014; Cameron & Miller, 2015; Abadie et al., 2017).

In schools where teachers were tutors, classes were pooled so that one teacher at a time could tutor the participating students. As the class sizes therefore became bigger, the quality of regular instruction may have suffered. However, *Läsklar* was typically not implemented during periods with formal reading instruction. Nevertheless, we provide a partial check on this concern. We use a variant of the main specification where we only include the schools that used special educators as tutors. Because the control group did not experience larger classes in this sample (if anything, slightly smaller), if we see positive effects in this specification then reduced quality of regular instruction is an unlikely explanation of the effects. [24] All schools implementing *Läsklar* in first grade used special educators as tutors. We omit all first-grade students from this specification to avoid confounding with grade effects.

It is common practice in educational research to treat, as we do in the primary analysis, test scores as an interval scale (i.e., every increment represents an equally large change of the measured skill). Although most test scores are ordinal (e.g., Jacob & Rothstein, 2016), it may be an acceptable simplification when evaluating program effects. In our case however, a change from a 0 to 1 may represent a more important improvement than going from for example 4 to 5, since a change from 0 to 1 means that the students grasp what letters represent for the first time, or how to decode. Only the decoding test has a large share of zeros at pre-test in our sample and we focus this sensitivity analysis on that test. We run similar specifications to Equation (1) but use a series of 20 separate linear probability models (LPM) and outcome variables equal to 1 if the student's decoding post test score is at least 1, at least 2, and so on up to at least 20 and over. We exclude the pre-tests in these models to avoid problems with continuous covariates in LPMs.

The decoding test contains 7 words out of 200 that overlap with the words used in the second type of session in *Läsklar*. Personal communication with our main tester revealed that very few students could read orthographically at post-test, i.e., students do not recognize words on a visual basis but identify words phonologically based on letter-sound conversion (Gustafson et al., 2000). It is therefore unlikely that our results are affected by students learning these words by heart. Furthermore, only two of the seven words appear early in the test and it seems exceedingly unlikely that the students would reach more

[24] As mentioned in section 2.3.1, special educators are rarely used in kindergarten and early in first grade when we implemented the intervention. Therefore, it is unlikely that the intervention worsened the quality of instruction for the control group by reducing access to special educators.

than those two without being able to decode.[25] The sensitivity test using a binary outcome variable is also a test of whether the overlapping words between the *Läsklar* material and the decoding test is a major influence of our results. If the effects are driven by the treatment group decoding one or two overlapping words correct, we should see a large dip in the treatment effects after those scores. We also run a second sensitivity test where we simply reduce the treatment group's decoding scores by one and two points (but no further than to zero points). This is a very conservative test, as it assumes that all treated students learnt these words by heart and that no control students did.[26]

2.5.3 Exploratory analyses of the short-term effects

Our analysis plan specified an analysis of the heterogeneity of the short-term effects across pre-test scores.[27] We expected most students' baseline decoding scores to be very close to zero, and the letter knowledge and phonological awareness to be highly correlated. We therefore limited the use of test score moderators to the letter knowledge test. We similarly expected the three questions pertaining to motivation, enjoyment, and self-efficacy to be highly correlated, and use the average score across the three measures. For both moderators, we create a variable indicating whether a student is below the median and interact this indicator with the treatment indicator. Both moderators are included in the same regression.

We also perform four exploratory analyses that we did not pre-register. We examine if results are different for students who had specific reasons for being at-risk, and if there are differences in the effects in kindergarten and first grade. Two schools have prior experience with the program (for earlier cohorts) and some schools chose to participate with a second cohort. We test if the results are heterogeneous over the level of experience by interacting the treatment indicators for students in these schools.

[25] The third overlapping word is number nine in the word list of part A. Because the instructions tell the tester to wait for more than five seconds when the student hesitates at a word and the time limit for each part is 45 seconds this word should be difficult to reach for students that do not know how to decode any words.

[26] Our pre-registered plan contains a test of the sensitivity to missing observations by dropping all pairs/triples with missing outcomes and thus balancing treatment and control groups regarding attrition. However, all students with missing outcomes were matched in pairs, which means that there is no within-pair variation in treatment status in pairs with missing outcomes. Because we include pair/triple fixed effects in our primary analysis, dropping these pairs yields the exact same estimates.

[27] Our pre-registered plan also includes a test of the effects of getting one session more using the treatment indicator as an instrumental variable (IV) for the number of sessions, as suggested by e.g., Angrist et al. (1996). However, the intention was to get closer to a treatment-on-treated effect in case some treated students got very few sessions. As mentioned, no student got very few sessions. Although the results are supportive of the hypothesis that more sessions are better, these estimates are of limited value as the variation in the number of sessions in the treatment group is relatively low. Therefore, we omit these results (they are available on request).

2.5.4 Costs and cost-effectiveness

The aim of the cost-effectiveness evaluation is to get an estimate of how much it would cost to replicate the implementation and get similar effects on the students' literacy skills. We follow Hollands et al. (2013, 2016), Levin and Belfield (2015), Jacob et al. (2016) and express cost-effectiveness as the costs per unit increase in effect size. We calculate the costs of the program for a typical school using three main categories of costs: *personnel*, *facilities*, and *materials and equipment*.

There are three items related to personnel costs. The first measures the time it took tutors to learn how to use the program. The second item is the time used to prepare and implement the sessions. The third item relates to the screening of students. We include time for taking the pre-treatment test plus time for schools to make the initial selection of students. As the post-treatment tests were part of the study but not the program, we do not include the costs of these tests. Such testing is either something schools do already or do not want to do; either way the extra costs are zero.

We gathered information about personnel costs by interviewing the tutors and our main tester, and by a short survey to the tutors. We convert teacher hours to SEK/USD/Euro using information about average national wages for teachers and special educators with similar level of experience as our tutors,[28] and information about Purchasing Power Parity adjusted exchange rates for 2017 (OECD, 2018). Because the exchange rates may fluctuate between years, we also convert all costs to tutor hours using a weighted cost per hour that reflect the three types of tutors in our study. In the few cases where we lack information, we either impute values by the average for the school or for all schools (e.g., if information was not possible to obtain from a whole school).

We do not have access to estimates of the opportunity costs of using school facilities. It is unclear if the facilities would have been used for something else, but we use an estimate of 6 USD per student. This cost amounts to a weighted average of the facility cost per student in the two most similar programs in Hollands et al. (2013; 3 USD per student

[28] The average national wages are from November 2017 for experienced teachers and two types of special educators ("specialpedagog", "speciallärare") in the early years of primary school (personal communication with Lärarförbundet, a teachers union). We calculate an hourly cost for the three categories as follows: ((11 × monthly wage + 1 month of holiday wage) × 1.3142 [employers' fees] × 1.05 [employers' insurance costs])/1760 [yearly work hours] (see e.g., Bolagsverket, 2018). The calculation yields 328 SEK, 375 SEK (specialpedagog), and 369 SEK (speciallärare) per hour for the three categories.

in *Stepping Stones*, which lasts 5 weeks, and 11 USD per students in *Sound Partners*, which lasts 18 weeks).

Materials and equipment include the cost of program material and the tutor training course, the costs of making copies for the pictures used in the second type of session, and the costs of test materials.

Two schools had already used the program before and three other schools used the program in a second round for a new cohort of students. Some features like training and material were therefore not relevant for them, so we assign the average costs of these items to those schools. We calculate total costs as the weighted average over schools, with weights corresponding to the schools' shares of treated students. The total costs should therefore be interpreted as the average opportunity cost of implementing the program per treated student the first time a sample-typical school uses it.

2.5.5 Effects at follow-up

We test hypotheses about the effects of program timing– whether it matters to have been waitlisted–using the same regression framework as described above. The outcome variables are the same measures used to evaluate the short-term effects, measured at follow-up. We also perform similar sensitivity analyses as for the short-term effects. However, as the same tester conducted all follow-up tests and the control group had also trained with the program, examining whether having a different tester matters and whether the quality of regular instruction for the control group worsened is not relevant.

3 Results

3.1 Pre- and post-test distributions of the outcome variables

Figure 1 and 2 displays the pre- and post-distribution of the outcome variables in the treatment and control group. Raw test scores are shown on the x-axis and the fraction of students in a certain bin on the y-axis. The grey bars represent the treatment group and the white bars the waiting list control group. All distributions are similar across the two groups at pre-test. The figures furthermore indicate that the treatment group improve their scores more than the control group from pre- to post-test over large parts of the distributions on the three literacy tests and the self-efficacy measure. There are small differences both over time and between the groups on the enjoyment and motivation measures. Most students score high already at pre-test and the scores increase somewhat on the post-test. That is, there is a clear risk of ceiling effects on these two measures.

Figure 2 Pre- and post-test distributions of the literacy tests

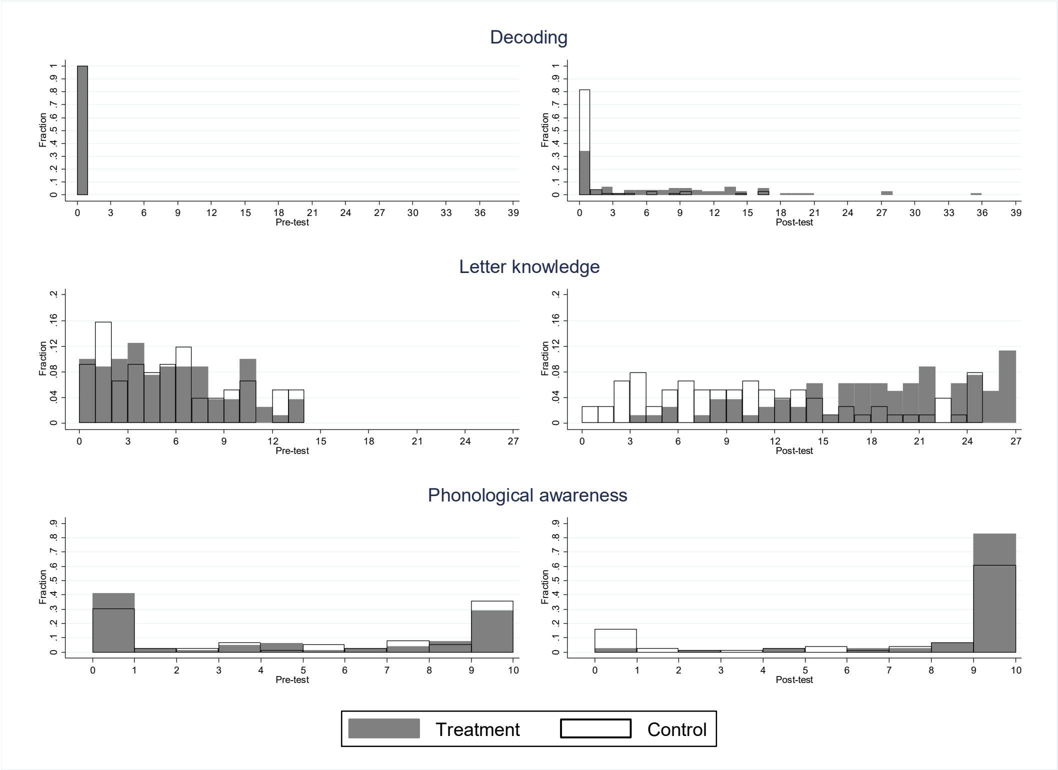
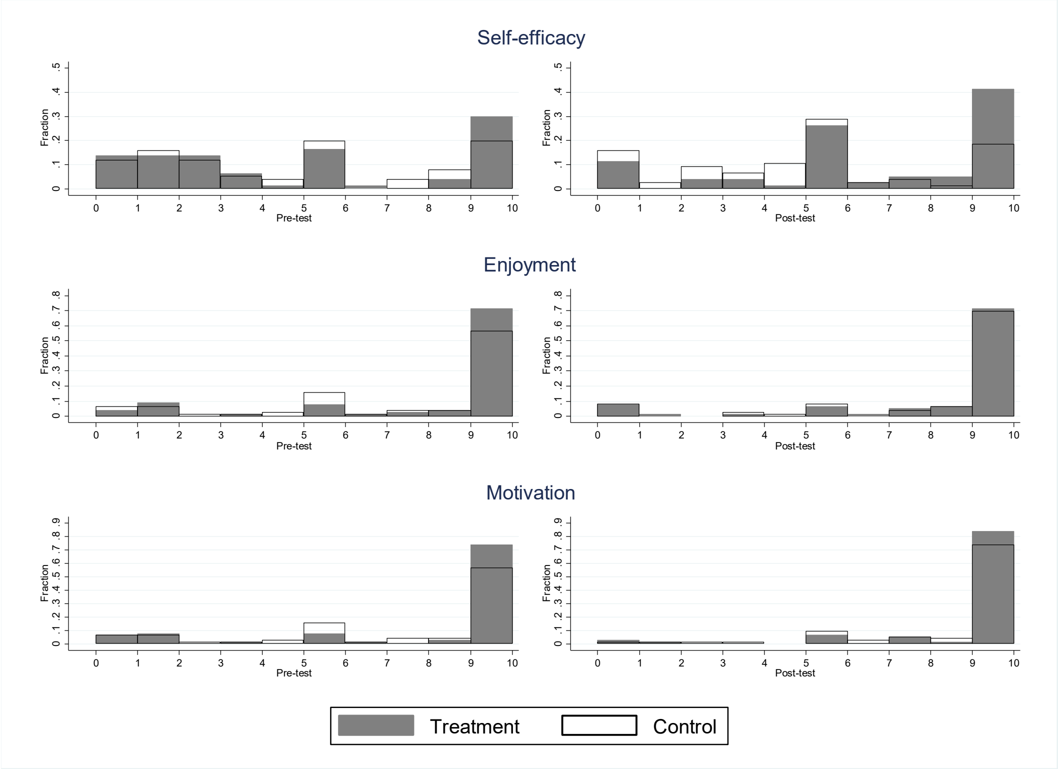


Figure 3 Pre- and post-test distributions of the attitude measures



In fact, a reasonably large share of the students is at the max score also on the letter knowledge and phonological awareness tests, and the measure of self-efficacy. Most control group students still score zero on the decoding post-test. The share of students scoring zero is much smaller in the treatment group; around 65 percent can decode at least one word.

3.2 Main results

Table 4 presents the estimates of the short-term effects of Läsklar. Each column corresponds to one of the six outcome measures and the coefficients of the treatment variable in Table 4 express the treatment effects in raw scores (robust standard errors are in parentheses below each coefficient).

Table 4 Main results: short-term effects

Variables	Decoding (1)	Letter knowledge (2)	Phonological awareness (3)	Self-efficacy (4)	Enjoyment (5)	Motivation (6)
Treatment	6.333***	6.911***	1.731***	1.767***	-0.240	0.342
	(1.123)	(0.835)	(0.454)	(0.542)	(0.486)	(0.382)
Pre-test covariates						
Letter knowledge	-0.619	0.234	0.282	-0.694**	0.0256	-0.00180
	(0.575)	(0.479)	(0.211)	(0.325)	(0.254)	(0.169)
Phonological	0.470***	0.354**	0.333***	0.158	-0.0670	0.0168
Awareness	(0.148)	(0.157)	(0.0777)	(0.117)	(0.111)	(0.0612)
Self-efficacy	-0.265	-0.181	-0.0596	-0.114	-0.0977	-0.0468
	(0.167)	(0.149)	(0.0891)	(0.0992)	(0.0938)	(0.0566)
Enjoyment	-0.327	0.403*	0.124	0.135	0.182	0.0438
	(0.298)	(0.227)	(0.117)	(0.123)	(0.117)	(0.0805)
Motivation	0.662*	-0.0423	-0.0791	0.394**	0.250*	0.147
	(0.343)	(0.158)	(0.113)	(0.179)	(0.134)	(0.0894)
Hedges' *g*	1.07	1.03	0.56	0.57	-0.08	0.11
Pair/triple FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	156	156	156	156	156	156

Note: The table displays the coefficients and robust standard errors (in parentheses) from linear regression models based on Equation (1), including pre-tests and pair/triple fixed effects as covariates (the pair/triple fixed effects are partialled out). We calculate Hedges' g for the treatment effect according to Equation (2) and use the post-test standard deviations of each outcome measure. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

The treatment effects are positive and statistically significant in columns (1) and (2). Students who trained with the program can decode, on average, about six more words and know about seven more letters than the control group ($p < 0.001$ for both measures). The effect sizes, displayed in the bottom part of the table, are $g = 1.07$ and $g = 1.03$.

The effects are also positive and significant on the phonological awareness test ($g = 0.56$, $p < 0.001$) and the self-efficacy measure ($g = 0.57$, $p < 0.001$). The estimates are smaller and not significant on the enjoyment ($g = -0.08$, $p = 0.623$) and motivation measures ($g = 0.11$, $p = 0.374$).

We also show the coefficients for the pre-test scores, which are included as covariates in the specifications. As much of the variation is captured by the pair/triple fixed effects, few pre-tests are significantly associated with the outcomes. The main exception is the phonological awareness pre-test, which is significantly associated with the three literacy tests.

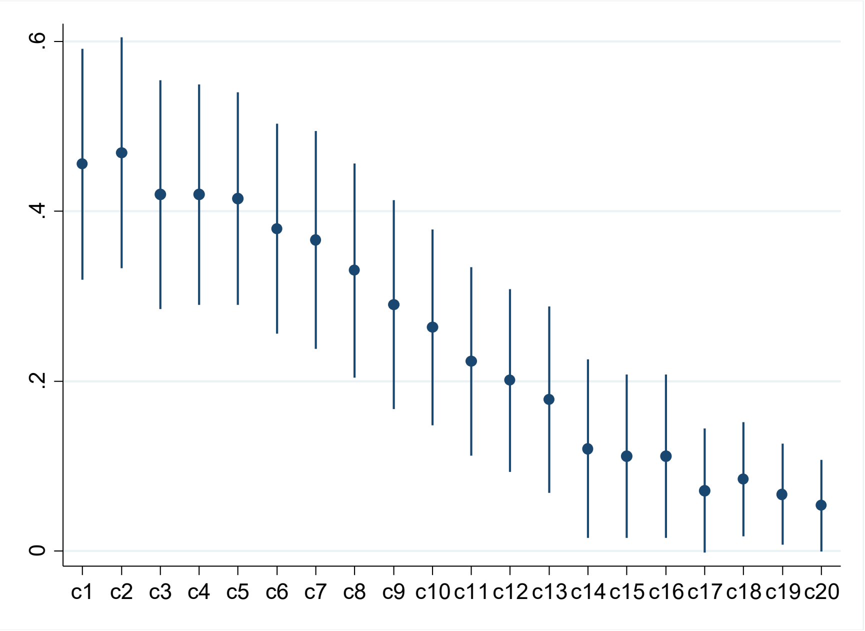
3.3 Sensitivity analyses

Appendix section A2 contains the results of the sensitivity analyses. Our results are robust to excluding covariates completely (and therefore to including the student without complete pre-test scores) and to excluding pair/triple fixed effects from the regressions. Excluding students who either were tested under different circumstances or were instructed by third author does not change our results in any meaningful way. The statistically significant estimates retain their significance when we cluster standard errors on instructional groups ($p < 0.001$ in all cases) and when we use randomization tests to perform inference ($p = 0.002$ or lower). Because the effects are of similar size or larger in the schools where special educators were tutors, we find little evidence that our results are driven by reduced quality of instruction for the control group. Moreover, as *Läsklar* was typically not conducted during regular reading instruction periods, the scope for negative spillovers of this kind ought to be small.

Our most important outcome measure, the decoding test, may not be well approximated by an interval scale. In addition, the test has a few words that overlap with the words used in the practice sessions of *Läsklar*. However, the pre- and post-distributions shown in Figure 1 indicate that many more students in the treatment group than in the control group can read more than just a few words, and that the effect is not driven by a few outliers that know many words.

This pattern is shown more clearly in Figure 3. The figure displays the coefficients and 95-percent confidence intervals of the treatment effect from a series of regressions where the outcome variable is a binary variable equal to one if a student could decode the number of words indicated by the category, or more words. The difference between the treatment and control group is very large for the lower categories. It is for example 42-47 percentage points in categories 1-5, which can be compared to the control group average that range from 11 to 18 percent in the same categories. The effects are statistically significant ($p < 0.05$) in all regressions except for category 17 and 20. Note also that there are no sudden drops, which we would have expected, had the treatment group just learnt some of the overlapping words by heart. The effect instead declines rather smoothly.

Figure 4 Results from LPM regressions with binary decoding variables as outcomes



Note: The figure displays coefficients and 95-percent confidence intervals (based on robust standard errors) of a series of 20 LPM models. The models are versions of our main specification that exclude covariates (but include pair/triple fixed effects) and use binary versions of the decoding post-test as the outcome variables. Treatment effects are on the y-axis and each category on the x-axis. In the category 1 (c1) specification, the decoding variable is equal to 1 if the student could decode one word or more at post-test and zero otherwise; in category 2, the decoding variable is equal to 1 if the student could decode two or more words at post-test and zero otherwise; and so on up to the last category, where the outcome variable is equal to 1, if the student could decode 20 or more words.

The effects on the decoding test are also robust to deducting one and two points from each student in the treatment group. These conservative tests yield effects that are still positive, significant, and large (reducing by one-point yields $\hat{\beta} = 5.7$, $p < 0.001$, $g = 0.99$, and by two points $\hat{\beta} = 5.0$, $p < 0.001$, $g = 0.90$).[29] We conclude that the results do not seem to be driven to any substantial degree by students learning the overlapping words by heart. Recall also that this interpretation is corroborated by information from our tester: very few students were able to read orthographically at post-test.

[29] As deducting points reduces the standard deviations as well, the reduction in effect sizes becomes very small.

3.4 Exploratory analyses of the short-term effects

We find relatively weak evidence of heterogeneity across pre-test scores. There is a tendency in Table 5 for the treatment effects to be smaller for students with below median letter knowledge scores at pre-test on the decoding test and higher on the letter knowledge and phonological awareness test, as well as the measure of self-efficacy. However, no interaction is significant at the 5 percent level (*p* = 0.06 for the decoding test). There are no significant differences between the groups with above and below median pre-test scores on the composite attitude measure and the sign of the interaction effects differ.

Table 5 Heterogeneity of treatment effects across pre-test scores

Variables	Decoding (1)	Letter knowledge (2)	Phonological awareness (3)	Self-efficacy (4)	Enjoyment (5)	Motivation (6)
Treatment	6.390***	6.429***	1.378*	1.031	-0.141	0.706
	(2.144)	(1.699)	(0.761)	(1.308)	(0.603)	(0.519)
Treatment *x* Low letter	-3.841*	2.784	1.458	1.879	-0.188	-0.198
Knowledge	(2.014)	(1.849)	(0.941)	(1.198)	(0.641)	(0.504)
Treatment *x* Low attitude	3.941	-1.514	-0.702	-0.218	0.413	-0.498
	(2.979)	(2.153)	(1.110)	(1.481)	(0.790)	(0.708)
p(T+Low Letter knowledge)	0.087	0.000	0.004	0.005	0.549	0.444
p(T+Low attitude)	0.000	0.001	0.342	0.348	0.618	0.635
Pre-tests	Yes	Yes	Yes	Yes	Yes	Yes
Pair/triple FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	156	156	156	156	156	156

Note: The table displays the coefficients and robust standard errors (in parentheses) from linear regression models including pre-tests and pair/triple fixed effects as covariates (the pair/triple fixed effects are partialled out). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Appendix Table A3 contains four additional exploratory analyses that we did not preregister. In brief, we find again only weak evidence of heterogeneous effects. Treatment effects are not significantly different (*p* > 0.05) between students in kindergarten and first grade, between students tutored by tutors with some or no prior experience of the program, or between students with or without specific reasons for being at-risk.

3.5 Costs and cost-effectiveness

Table 6 displays the costs of implementing the program per treated student. It took on average 10.7 hours to learn the method, prepare and implement the sessions, and screen students.[30] As mentioned, we use an estimate of school facility costs of 6 USD per treated student. Materials and equipment cost 2,202 SEK per treated student, which include the cost of the program material, the tutoring training course, copies, and the costs of test materials. We use the price schools would pay to buy the program outside of the study, an estimate of the copying cost, and the price paid per school by the study for the set of test materials.[31] Total costs amount to 6,040 SEK/703 USD/627 Euro, or 17 tutor hours per treated student.

Table 6 Costs per treated student

Cost items	SEK	USD	Euro	Hours	%
Personnel costs	*3786.5*	*442.9*	*392.9*	*10.7*	*63%*
Tutor training	329.7	38.6	34.2	0.9	5%
Implementation	3456.8	404.4	358.7	9.8	57%
Screening students	116.9	13.7	12.1	0.3	2%
School facilities	*51.3*	*6.0*	*5.3*	*0.1*	*1%*
Material and equipment	*2202.2*	*254.3*	*228.5*	*6.2*	*36%*
Program material and course	1935.8	226.4	200.8	5.5	32%
Copies	14.0	1.6	1.5	0.0	0%
Screening tests	252.4	26.2	26.2	0.7	4%
Grand total	**6404.0**	**703.2**	**626.7**	**17.1**	**100%**

Note: Costs calculated as described in Section 2.4.4. We use the following currency conversion rates: USD/SEK = 8.549, Euro/SEK = 9.638 (OECD, 2018).

Note that this overestimates the costs for schools that keep on using the program. The cost refers to the first time a school implements the program. In total 36 percent of the costs are fixed and would only have to be paid once (unless new tutors are trained).

In the discussion section we compare *Läsklar* to the cost-effectiveness of a few other US programs. We then use the average effect size for the decoding and letter knowledge test, as these tests are most comparable to the tests of "alphabetics" used in e.g., Hollands

[30] The screening hours refer to the time it would take a teacher to test an average student. As our tester had to travel to the schools it cost more, but this is not the relevant opportunity cost for a school, where regular teachers who are always travelling to the school can perform the screening tests.

[31] This item only includes the costs of the two standardized tests; all other test materials can be used freely.

et al. (2013) and Jacob et al. (2016). The costs per unit increase in effect size for this measure is 5,768 SEK/671 USD/598 Euro.

3.6 Results at follow-up

The raw means displayed in Panel C of Table 2 indicate that the treatment group have further improved their scores on the three literacy measures. This is important as it suggests that the improvement of the students' literacy skills was not transient. The measures of attitudes have decreased somewhat compared to the post-test, although they are still higher than at pre-test for the self-efficacy and motivation measures. The enjoyment and motivation measures are still relatively close to the maximum score for most students. The control group have improved all scores from post- to follow-up test. They lag the treatment group somewhat on the literacy tests but have higher scores on the attitude measures.

In principle, our design provides a formal test of program timing, i.e., whether it is better to receive the program in the first or in the second period. However, for three reasons we caution against strong interpretations of the differences between the groups at follow-up. First, treated students who were not able to decode at post-test were in many schools given more sessions. The timing effect is therefore confounded with the effect of getting extra sessions.[32] Second, two schools that ran a second round of the program wanted to use the method in larger groups and included all students after the treatment group had received the program. As this experiment would not affect our estimations of the short-term effect, we did not want to discourage the schools from spreading the program to more students. All students on the waiting list who needed one-to-one instruction still got this, but some control group students in these two schools received a slightly different treatment. Third, the ceiling effects are more problematic at follow-up than at post-test for all tests except the decoding test.[33]

Table 7 displays the effects at follow-up. The treatment group has higher scores on the three literacy measures, and lower scores on the attitude measures. No difference is statistically significant, although for decoding, phonological awareness, self-efficacy, and enjoyment the differences are reasonably large (*g* is slightly below 0.3 for these tests).

[32] We lack information about the number of extra sessions.

[33] There is also more attrition at follow-up and one school withdrew from the study between post- and follow-up tests. Overall attrition is still relatively low, 12 percent, and there is hardly any differential attrition between the treatment and control group, so this ought to be a minor problem.

The sensitivity tests shown in Appendix Table A4 do not change our interpretation that the evidence of differences between the treatment and waiting list control group are relatively weak.

Table 7 Effects at follow-up

Variables	Decoding (1)	Letter knowledge (2)	Phonological awareness (3)	Self-efficacy (4)	Enjoyment (5)	Motivation (6)
Treatment group	3.314	0.433	0.488	-0.894	-0.818	-0.366
	(2.016)	(0.734)	(0.294)	(0.611)	(0.594)	(0.437)
Pre-test covariates						
Letter knowledge	-0.00656	-0.322	0.0280	-0.236	0.157	0.0774
	(0.869)	(0.405)	(0.120)	(0.311)	(0.298)	(0.293)
Phonological	0.797***	0.325*	0.150**	0.223*	0.0867	0.00409
Awareness	(0.250)	(0.188)	(0.0573)	(0.130)	(0.140)	(0.0967)
Self-efficacy	-0.553**	-0.183	-0.0256	0.0636	-0.137	-0.131*
	(0.246)	(0.130)	(0.0398)	(0.105)	(0.114)	(0.0738)
Enjoyment	-0.331	0.276	-0.0477	-0.190	0.0470	0.00331
	(0.649)	(0.240)	(0.0560)	(0.187)	(0.195)	(0.107)
Motivation	0.455	-0.141	-0.121	0.220	0.210	0.293**
	(0.768)	(0.166)	(0.0743)	(0.170)	(0.224)	(0.126)
Hedges' *g*	0.28	0.07	0.28	-0.27	-0.29	-0.14
Pair/triple FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	141	141	141	141	141	141

Note: The table displays the coefficients and robust standard errors (in parentheses) from linear regression models based on Equation (1), including pre-tests and pair/triple fixed effects as covariates (the pair/triple fixed effects are partialled out). Hedges' *g* for the treatment effect is calculated according to Equation (2) using the standard deviations at follow-up. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

4 Discussion

This section discusses the magnitudes of the effects and compares the cost-effectiveness of *Läsklar* to a few other programs for which costs have been estimated using similar methods as we use. We also discuss how features of the study design may have influenced the effect sizes as well as some limitations of our study.

4.1 Effect sizes and cost-effectiveness in perspective

To put the effects of *Läsklar* in perspective, we compare them to 1) statements from norm-giving organizations; 2) gaps between student groups; 3) how much students typically learn in a year; and 4) to effect sizes from other interventions. We limit the discussion to the literacy tests, as none of benchmarks 1)–3) are available for the attitude measures and it is unclear how our measures compare to other measures used in the earlier literature.

What Works Clearinghouse in the US maintains that effect sizes over 0.25 should be regarded as "substantively important" (What Works Clearinghouse, 2014, p. 23). Lipsey et al. (2012) show that US students improve on average 1.52 standard deviations between kindergarten and first grade and 0.97 between first and second grade on standardized tests in reading. The gap between low and high SES students are around 0.7-0.8 in grade 4 (information is not available for grade K-3 and we are not aware of similar statistics for Sweden). The effect sizes of *Läsklar* appear large in comparison to these three benchmarks.

We want to acknowledge some important caveats to between-study comparisons, before we compare the effect sizes and cost-effectiveness of *Läsklar* to other interventions. There are often differences between the duration, intensity, target population, the control group condition, and between the outcome measures and the formulas used to estimate the effects and calculate effect sizes. It is well known that many of these features moderate effect sizes (e.g., Cheung & Slavin, 2016), but more difficult to say by exactly how much when comparing two individual studies (we discuss this issue further in the next section). The comparison problems are worse in cost-effectiveness analyses, which are affected by differences in the estimations of both effect sizes and costs, as well as by differences in input prices across countries and regions.

We match how we calculate our effect sizes to the methods used in Dietrichson et al. (2017). The two primary outcomes we use are standardized tests, which are the only ones included Dietrichson et al.'s review of interventions for students with low SES in grade

K-8. This target group is furthermore reasonably close to ours. For these reasons, we believe this review is the closest comparison.[34] The average effect size of all reading interventions in that review is $g = 0.09$. The average effect size of all tutoring interventions in reading is $g = 0.30$, which in turn is close the average effect size for the six studies of reading interventions involving tutoring and, at least in part, kindergarten students ($g = 0.31$).[35] Thus, our effect sizes appear large also in comparison to the typical effect sizes found for other interventions and to similar tutoring interventions.[36]

We focus the cost-effectiveness comparison on programs involving tutoring and similar student groups and grades to minimize the between-study differences. We also focus on the decoding and letter knowledge tests (both are "alphabetics" tests in the terminology of Hollands et al., 2013), which are our two primary outcomes, are standardized, and have reasonably similar counterparts among standardized tests used in previous interventions from the US.

Hollands et al. (2013) estimate the cost per unit increase of effect size of three comparable programs: *Stepping Stones*, a 5-week tutoring program using effect sizes derived from at-risk kindergarten students in Nelson et al. (2005a, b); *Sound Partners*, an 18-week program for at-risk kindergartners, evaluated in Vadasy and Sanders (2008); and *Reading Recovery*, a 20-week program targeting at-risk students in first grade and evaluated in Schwartz (2005). *Stepping Stones* is slightly more cost-effective than *Läsklar* at 570 USD per unit increase in effect size in alphabetics,[37] while both *Sound Partners* and *Reading Recovery* are less cost-effective (2,093 and 1,480 USD per unit increase in effect size calculated with alphabetics tests).

Two grade 3 programs (*Corrective Reading* and *Wilson Reading*) examined in Hollands et al. (2013, 2016) are in comparison considerably less cost-effective, but also target a broader set of reading skills.[38] Jacob et al. (2016) evaluate a version of *Reading*

[34] The results for tutoring interventions in Dietrichson et al. (2017) are close to those obtained by Slavin et al. (2011).

[35] This effect size is based on the following articles: Burns et al. (2003), Vadasy & Sanders (2008, 2010), Amendum et al. (2011), Nielsen et al. (2012), and Apel & Diehm (2014). Some of these studies also include students in higher grades.

[36] Our effect sizes also seem large in comparison to the effect sizes of school interventions in general (see e.g., Fryer, 2017, for a review). See also the webpage Evidence for ESSA (www.evidenceforessa.org, accessed 2018-10-18), which is continuously adding information about programs/evaluations from the US and include effect sizes calculated in a reasonably comparable way to ours.

[37] The effect sizes are very similar, but *Stepping Stones* costs less, mainly due to the short intervention period.

[38] Corrective Reading costs 45,945 USD (38,135 USD for alphabetics) and Wilson Reading 20,291 USD (13,392 USD for alphabetics) per unit increase in effect size, mainly due to their more intensive tutoring (1 hour per day, 5 days per week), much longer duration (28 weeks), and lower effect sizes.

Partners, which uses volunteer tutors and includes students in grades 2-5. Volunteers reduce the costs attributable to schools, but the cost per unit increase of effect size is still much larger than for *Läsklar*, 6,455 USD for the closest test (the reason is mainly that the effect sizes are small).[39] Thus, *Läsklar* compares favorably to other programs both in terms of the magnitudes of the effects and the costs incurred when producing those effects.[40]

4.2 Study design and effect sizes

Effect sizes may be influenced by the study design (e.g., Cheung & Slavin, 2016), rather than the program components. We discuss four such features below and compare to the six most similar tutoring studies–called the "comparison studies" below–included in Dietrichson et al. (2017; see footnote 39 for references).

First, the control group received business-as-usual instruction while waiting to get the program. The usual instruction did not include supplementary in-school instruction like tutoring. Knowing that they would receive the program may have reduced the risk of the control group getting more help outside of school compared to studies where the control group does not get the intervention at all. That is, our control group is close to a no-treatment control group, which ought to produce larger effect sizes than when the control group receives at least some extra help. Of the comparison studies, the control condition in Vadasy and Sanders (2008, 2010) and Apel and Diehm (2014) is similar to ours, while it is unclear what the control group gets in the others. The control group in the comparison studies improves on average 0.68 post-test standard deviations between pre- and post-tests. In comparison, our control group improves 0.55 standard deviations averaged over our two primary outcomes.

Second, our primary outcomes are standardized tests and not constructed for the study. They are still well aligned with the content of *Läsklar*, which tend to produce larger effects (e.g., Madden & Slavin, 2011). However, our control group received instruction in these areas too, as discussed in Section 2.3.2. The average effect size in the comparison

[39] Total costs for all involved parties, including opportunity costs for tutors, are much larger and cost-effectiveness decreases considerably to 32,820 USD.

[40] Scammaca et al. (2007) include cost estimates for 12 programs in grade K-3. Because they only include personnel costs based on a standard hourly wage, their estimates are not fully comparable to ours. But 10 out of 12 programs have higher personnel costs than *Läsklar* and are less cost-effective. Note that some of the programs and target groups differ from *Läsklar* and that not all tests are alphabetics tests, so the effect sizes are also less comparable.

studies is also based on standardized tests. Over 80 percent seem similarly well aligned and no test is not at all aligned.

Third, our study was more of an effectiveness study than an efficacy trial. We provided minimal extra support and monitoring during the implementation, which mimicked how the program is implemented when schools use it outside the study. We expect this feature to result in smaller effect sizes (see e.g., Thomas et al., 2018, for an interesting discussion and results in line with this hypothesis). All comparison studies monitor and support their tutors more than we do or have personnel hired by the study that perform the intervention.

Fourth, we include the students most at-risk of reading difficulties. Vadasy and Sanders (2008), and Nielsen et al. (2012) use selection procedures that seem to result in a similar at-risk group, whereas the other comparison studies either exclude the students most at-risk or include students who are less at risk. It is not clear though whether it is harder or easier to produce large effect sizes with students that are more at-risk.

In sum, our study design features are not only in favor of larger effects. The combination of program components in *Läsklar* may therefore be the reason for the large effect sizes.[41] However, our design does not allow us to disentangle the effects of the intervention components, not from each other and not from other study features. Multi-sensory methods, tutoring group sizes, and the type of tutors would all be interesting components to examine in a design where intervention components are tested against each other while other study features are held constant between components.

4.3 Limitations

We intended to get as close possible to how the schools would have implemented the program, had there not been a study. However, to examine the effects we had to perform measurements, which should be regarded as an interference over and above what schools normally would have done, and the tutors were aware that they participated in the study. As parents were informed of the study, they were also aware. This may have created Hawthorne effects – e.g., tutors may have tried harder because they knew they were being studied – but this is very difficult to avoid in educational interventions and our study is not different from others in this regard.

[41] Interventions using teachers as tutors has larger effect sizes according to Slavin et al. (2011). Tutoring delivered by school professionals and one-to-one tutoring is associated with larger effect sizes than non-professionals and 1:2-5 tutoring in Dietrichson et al. (2017), but not significantly so. Four comparison studies use one-to-one or one-to-two tutoring, while the others use somewhat larger groups.

The waiting list design precludes us from estimating longer-term effects of the program. There are few studies that have followed participants of reading interventions over longer periods than a year after the end of intervention. Suggate (2016) documents a positive but decreasing effect over time: the longer-term average effect size is about 60 percent of the short-term effect size. The two longest follow-up studies of interventions directly targeting similar at-risk students as we do–Blachman et al. (2014) and Elbro and Petersen (2004)–find significant effects ten and seven years after their interventions, respectively.[42] These results indicate that the effects of targeted interventions can be long lasting.

It is also reassuring that the treatment group continue to improve on all literacy measures from post-test to follow-up and that effect sizes amounting to 60 percent of our short-term effects would still be sizeable. Two aspects of *Läsklar* may furthermore help students remember what they learnt: using fingerspelling helps them remember the letter-sound correspondence and the timing of the intervention makes sure that letter knowledge and decoding is still in focus in the regular instruction. That is, students are reminded of what they learnt also during regular instruction.

Our outcome measures have some limitations. The attitude measures have not been examined for validity and reliability. Therefore, the results based on those measures should be interpreted with caution regarding what underlying psychological constructs they capture. The decoding test contains a couple of words that overlap with the ones used during instruction with *Läsklar*. However, we show that these words are not driving our results. Our measures may furthermore have problems with floor (the decoding test) and ceiling effects (all the others, but in particular the enjoyment and motivation tests and more so at follow-up than post-test). Floor effects may overestimate the effects if more control group students are close to scoring one instead of zero and underestimate the effects if the treated students scoring zero are closer to cracking the reading code. Conversely, ceiling effects are likely to underestimate the effects, at least on letter knowledge, phonological awareness, and self-efficacy, as more treated students reach the maximum score on those outcomes.

[42] Blachman et al. (2014) implement their tutoring intervention in second and third grade and Elbro and Petersen (2004) do not provide end of intervention results, which is why we do not include them among the comparison studies in the discussion of effect sizes.

Our cost estimates are based on recall, not a cost diary. However, most of the studies we benchmark our costs to use similar methods (in some cases with longer recall periods, see Hollands et al., 2013). We underestimate (overestimate) the cost-effectiveness of the program, if there are for example positive (negative) peer effects, which our study was not designed to estimate. We are not aware of any study that have provided credible estimates of the peer effects of a similar targeted intervention.

5 Conclusions

We examine a literacy program–*Läsklar*–targeting the students most at-risk of reading difficulties in kindergarten and first grade. We find large positive effects on our two primary outcome measures, a standardized test of decoding (g = 1.07) and a standardized test of letter knowledge (g = 1.03), and positive effects on measures of phonological awareness (g = 0.56) and self-efficacy (g = 0.57). The effects are insignificant on measures of enjoyment and motivation, which may be explained by ceiling effects. The program costs about 6,400 SEK/700 USD/630 Euro/17 teacher hours per treated student the first time a school implements it (and gets cheaper in further iterations), which together with the large effect sizes implies that the program compares favorably to most other similar programs in terms of cost-effectiveness.

We believe these results are very promising for the students who are most at-risk of reading difficulties. Our sample is geographically concentrated and include schools that selected into the sample, perhaps because they are, on average, more disadvantaged than the average Swedish school. Although we see few obstacles to implementing the program in other schools and most schools have at least a few students in need of extra help, the results may not be fully generalizable. Future studies that try to replicate our results in other contexts would therefore be important. Examining the longer-term cost-effectiveness and the separate contribution of program components would also be interesting future research.

References

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. (2017). *When should you adjust standard errors for clustering?* (NBER Working Paper no. 24003). Retrieved from http://www.nber.org/papers/w24003.

Amendum, S. J., Vernon-Feagans, L., & Ginsberg, M. C. (2011). The effectiveness of a technologically facilitated classroom-based early reading intervention: The targeted reading intervention. *Elementary School Journal*, *112*(1), 107–131.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*(434), 444–455.

Apel, K., & Diehm, E. (2014). Morphological awareness intervention with kindergarteners and first and second grade students from low SES homes: A small efficacy study. *Journal of Learning Disabilities*, *47*(1), 65–75.

Archambault, I., Eccles, J. S., & Vida, M. N. (2010). Ability self-concepts and subjective value in literacy: Joint trajectories from grades 1 through 12. *Journal of Educational Psychology*, *102*(4), 804–812.

Athey, S. & Imbens, G. (2017). The econometrics of randomized experiments. In: Banerjee, A. & Duflo, E. (Eds.), *Handbook of Field Experiments, Volume 1*. Amsterdam: Elsevier.

Baum, C. F., Schaffer, M. E., & Stillman, S. (2010). *ivreg2: Stata module for extended instrumental variables/2SLS, GMM, and AC/HAC, LIML and k-class regression.* Retrieved from http://ideas.repec.org/c/boc/bocode/s425401.html.

Blachman, B. A., Schatschneider, C., Fletcher, J. M., Murray, M. S., Munger, K. A., & Vaughn, M. G. (2014). Intensive reading remediation in grade 2 or 3: Are there effects a decade later? *Journal of Educational Psychology*, *106*(1), 46–57.

Bolagsverket (2018). Räkna ut vad en anställd kostar. www.verksamt.se. Last accessed 2018-09-28.

Borman, G. D., 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.

Broadbent, H. J., White, H., Mareschal, D., & Kirkham, N. Z. (2018). Incidental learning in a multisensory environment across childhood. *Developmental Science*, *21*(2), e12554.

Bruhn, M., & McKenzie, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics*, *1*(4), 200–232.

Burns, M. K., Senesac, B. V., & Symington, T. (2003). The effectiveness of the HOSTS program in improving the reading achievement of children at-risk for reading failure. *Literacy Research and Instruction*, *43*(2), 87–103.

Bus, A. G., & IJzendoorn, M. H. (1999). Phonological awareness and early reading: A meta-analysis of experimental training studies. *Journal of Educational Psychology*, *91*(3), 403–414.

Butler, R. (1999). Information seeking and achievement motivation in middle childhood and adolescence: The role of conceptions of ability. *Developmental Psychology*, *35*(1), 146.

Bøg, M., Dietrichson, J., & Isaksson, A. A. (2018). A multi-sensory literacy program for at-risk students in kindergarten – Promising results from a small-scale Swedish intervention. Unpublished manuscript.

Cameron, A. C., & Miller, D. L. (2015). A practitioner’s guide to cluster-robust inference. *Journal of Human Resources*, *50*(2), 317–372.

Chabrier, J., Cohodes, S., & Oreopoulos, P. (2016). What can we learn from charter school lotteries?. *Journal of Economic Perspectives*, *30*(3), 57–84.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *Quarterly Journal of Economics*, *126*(4), 1593–1660.

Cheung, A., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, *45*(5), 283–292.

Cohen, P. A., Kulik, J. A., & Kulik, C-L. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal*, *19*(2), 237–248.

Connor, C. M., Day, S. L., Phillips, B., Sparapani, N., Ingebrand, S. W., McLean, L., Barrus, A., & Kaschak, M. P. (2016). Reciprocal effects of self-regulation, semantic knowledge, and reading comprehension in early elementary school. *Child Development*, *87*(6), 1813–1824.

Cook, P. J., K. Dodge, G. Farkas, R.J. Fryer, J. Guryan, J. Ludwig, ..., & L. Steinberg (2014). The (surprising) efficacy of academic and behavioral intervention with disadvantaged youth: Results from a randomized experiment in Chicago (NBER Working Paper no. 19862). Retrieved from http://www.nber.org/papers/w19862.

Cunha, F., & Heckman, J. (2007). The technology of skill formation. *American Economic Review*, *97*(2), 31–47.

Denton, C. A., Anthony, J. L., Parker, R., & Hasbrouck, J. E. (2004). Effects of two tutoring programs on the English reading development of Spanish-English bilingual students. *Elementary School Journal*, *104*(4), 289–305.

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.

Dietrichson, J., Filges, T., Klokker, R. H., Viinholt, B. C. A., Bøg, M., & Højmark, U. J. (2019). Targeted school-based interventions for improving reading and mathematics for students with or at-risk of academic difficulties in grade 7 to 12: A systematic review. Unpublished manuscript.

Ehri, L., Nunes, S. R., Willows, D. M, Valeska Schuster, B., 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.

Ehri, L. C., Dreyer, L. G., Flugman, B., & Gross, A. (2007). Reading rescue: An effective tutoring intervention model for language-minority students who are struggling readers in first grade. *American Educational Research Journal*, *44*(2), 414–448.

Elbaum, B., Vaughn, S., Tejero Hughes, M., & Watson Moody, S. (2000). How effective are one-to-one tutoring programs in reading for elementary students at risk for reading failure? A meta-analysis of the intervention research. *Journal of Educational Psychology*, *92*(4), 605–619.

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.

Elwér, Å., Fridolfsson, I., Samuelsson, S., & Wiklund, C. (2016). *LäSt – Test i läsförståelse, läsning och stavning för åk 1–6*. Hogrefe Psykologiförlaget: Stockholm.

Fives, A., Kearns, N., Devaney, C., Canavan, J., Russell, D., Lyons, R., ... & O'Brien, A. (2013). A one-to-one programme for at-risk readers delivered by older adult volunteers. *Review of Education*, *1*(3), 254–280.

Francis, D. J., Shaywitz, S. E., Stuebing, K. K., Shaywitz, B. A., & Fletcher, J. M. (1996). Developmental lag versus deficit models of reading disability: A longitudinal, individual growth curves analysis. *Journal of Educational Psychology*, *88*(1), 3–17.

Fredriksson, P., Öckert, B., & Oosterbeek, H. (2012). Long-term effects of class size. *Quarterly Journal of Economics*, *128*(1), 249–285.

Fryer, R. G. (2017). The production of human capital in developed countries: Evidence from 196 randomized field experiments. In Banerjee, A. and Duflo, E. (Ed.) *Handbook of Field Experiments Volume 2* (pp. 95-322), Amsterdam: Elsevier.

Furnes, B., & Samuelsson, S. (2010). Predicting reading and spelling difficulties in transparent and opaque orthographies: A comparison between Scandinavian and US/Australian children. *Dyslexia*, *16*(2), 119–142.

Fälth L., Gustafson S., & Svensson, I. (2017). Phonological awareness training with articulation promotes early reading development. *Education¸ 137*(3), 261–276.

Gertler, P., Heckman, J., Pinto, R., Zanolini, A., Vermeersch, C., Walker, S., ... & Grantham-McGregor, S. (2014). Labor market returns to an early childhood stimulation intervention in Jamaica. *Science*, *344*(6187), 998–1001.

Good, R. H., & Kaminski, R. A. (red.). (2002). *Dynamic Indicators of Basic Early Literacy Skills* (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement.

Gustafson, S., Ferreira, J., & Rönnberg, J. (2007). Phonological or orthographic training for children with phonological or orthographic decoding deficits. *Dyslexia*, *13*(3), 211–229.

Gustafson, S., Samuelsson, S., & Rönnberg, J. (2000). Why do some resist phonological intervention? A Swedish longitudinal study of poor readers in grade 4. *Scandinavian Journal of Educational Research*, *44*(2), 145–162.

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P., & Yavitz, A. (2010). Analyzing social experiments as implemented: A reexamination of the evidence from the HighScope Perry Preschool Program. *Quantitative Economics*, 1, 1–46.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics*, *6*(2), 107–128.

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.

Hollands, F. M., Kieffer, M. J., Shand, R., Pan, Y., Cheng, H., & Levin, H. M. (2016). Cost-effectiveness analysis of early reading programs: A demonstration with recommendations for future research. *Journal of Research on Educational Effectiveness*, *9*(1), 30–53.

Hollands, F. M., Pan, Y., Shand, R., Cheng, H., Levin, H. M., Belfield, C. R., Kieffer, M. J., Bowden, A. B., & Hanisch-Cerda, B. (2013). *Improving early literacy: Cost-effectiveness analysis of effective reading programs.* New York: Center for Benefit-Cost Studies of Education, Teachers College, Columbia University.

Hoover, W. A., & Tunmer, W. E. (2018). The simple view of reading: Three assessments of its adequacy. *Remedial and Special Education*, *39*(5), 304–312.

Häggström, I. (2007). Att förebygga läs- och skrivsvårigheter med språklekar. In Granström, K. (Ed.) *Forskning om lärares arbete i klassrummet, Forskning i fokus, nr 33*, Stockholm: Myndigheten för Skolutveckling.

Imai, K., King, G., & Nall, C. (2009): The essential role of pair matching in cluster-randomized experiments, with application to the Mexican universal health insurance evaluation. *Statistical Science*, *24*, 29–53.

Jacob, R., Armstrong, C., Bowden, A. B., & Pan, Y. (2016). Leveraging volunteers: An experimental evaluation of a tutoring program for struggling readers. *Journal of Research on Educational Effectiveness*, *9*(sup1), 67-92.

Jacob, B., & Rothstein, J. (2016). The measurement of student ability in modern assessment systems. *Journal of Economic Perspectives*, *30*(3), 85–108.

Jordan, K. E., & Baker, J. (2011). Multisensory information boosts numerical matching abilities in young children. *Developmental Science*, *14*(2), 205–213.

Johansson, M-G. (2009). *LäsEttan*. Stockholm: Natur & Kultur.

Landerl, K., Ramus, F., Moll, K., Lyytinen, H., Leppänen, P. H., Lohvansuu, K., ... & Kunze, S. (2013). Predictors of developmental dyslexia in European orthographies with varying complexity. *Journal of Child Psychology and Psychiatry*, *54*(6), 686–694.

Levin, H. M., & Belfield, C. (2015). Guiding the development and use of cost-effectiveness analysis in education. *Journal of Research on Educational Effectiveness*, *8*(3), 400–418.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., …, & Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms* (NCSER 2013-3000). Washington DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncser/.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Applied Social Research Methods Series, v. 49. Thousand Oaks, CA: Sage Publishing, Inc.

Lohr, S., Schochet, P. Z., & Sanders, E. (2014). *Partially nested randomized controlled trials in education research: A guide to design and analysis* (NCER 2014-2000). Washington, DC: National Center for Education Research.

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.

Lundberg, I., Frost, J. & Petersen, O. (1988). Effects of an extensive program for stimulating phonological awareness in preschool children. *Reading Research Quarterly*, *23*, 263–284.

Lundberg, I., & Hoien, T. (1996). Levels of approaching reading and its difficulties. In B. Ericson & J. Rönnberg (Eds.), *Reading disabilities and its treatment* (pp. 11–33). Linköping, Sweden: Linköping University.

Machin, S., McNally, S., & Viarengo, M. (2018). Changing how literacy is taught: evidence on synthetic phonics. *American Economic Journal: Economic Policy*, *10*(2), 217–241.

Slavin, R., & Madden, N. A. (2011). Measures inherent to treatments in program effectiveness reviews. *Journal of Research on Educational Effectiveness*, *4*(4), 370–380.

Muenks, K., & Miele, D. B. (2017). Students' thinking about effort and ability: The role of developmental, contextual, and individual difference factors. *Review of Educational Research*, *87*(4), 707–735.

Morgan, P. L., Fuchs, D., Compton, D. L., Cordray, D. S., & Fuchs, L. S. (2008). Does early reading failure decrease children's reading motivation?. *Journal of Learning Disabilities*, *41*(5), 387–404.

National Agency for Education (2018a). http://www.skolverket.se/statistik-och-utvardering/statistik-i-databaser. Data accessed 2018-02-26.

National Agency for Education (2018b). https://www.skolverket.se/undervisning/grundskolan/laroplan-och-kursplaner-for-grundskolan/laroplan-lgr11-for-grundskolan-samt-for-forskoleklassen-och-fritidshemmet. Data accessed 2018-09-13.

National Agency for Education (2018c). *Statistik om särskilt stöd i grundskolan.* Retrieved from https://www.skolverket.se/skolutveckling/statistik/arkiverade-statistiknyheter/statistik/2018-04-27-statistik-om-sarskilt-stod-i-grundskolan. Data accessed 2019-02-14.

Nelson, J. R., Benner, G. J., & Gonzales, J. (2005a). An investigation of a prereading intervention on the early literacy skills of children at risk of emotional disturbance and reading problems. *Journal of Emotional and Behavioral Disorders*, *13*(1), 3–12.

Nelson, J. R., Stage, S. A., Epstein, M. H., & Pierce, C. D. (2005b). Effects of a prereading intervention on the literacy and social skills of children. *Exceptional Children*, *72*(1), 29–45.

Nicholls, J. G. (1978). The development of the concepts of effort and ability, perception of academic attainment, and the understanding that difficult tasks require more ability. *Child Development*, *49*(3), 800–814.

Nielsen, D. C., & Friesen, L. D. (2012). A study of the effectiveness of a small-group intervention on the vocabulary and narrative development of at-risk kindergarten children. Reading Psychology, *33*(3), 269–299.

OECD (2016a). *Skills Matter: Further Results from the Survey of Adult Skills, OECD Skills Studies*. OECD Publishing, Paris. Retrieved from https://www.oecd-ilibrary.org/education/skills-matter_9789264258051-en.

OECD (2016b). *PISA 2015 results (Volume I): Excellence and equity in education*. Paris: PISA, OECD Publishing. Retrieved from http://www.oecd.org/publications/pisa-2015-results-volume-i-9789264266490-en.htm.

OECD (2018). Exchange rates. https://data.oecd.org/conversion/exchange-rates.htm. Last accessed 2018-09-28.

Pontoppidan, M., Keilow, M., Dietrichson, J., Solheim, O. J., Opheim, V., Gustafson, S., & Andersen, S. C. (2018). Randomised controlled trials in Scandinavian educational research. *Educational Research*, *60*(3), 311–335.

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.

Ritter, G., Denny, G., Albin, G., Barnett, J., & Blankenship, V. (2006). The effectiveness of volunteer tutoring programs: A systematic review. Campbell Systematic Reviews, 2006:7. Retrieved from https://www.campbellcollaboration.org/library/effectiveness-of-volunteer-tutoring-programmes.html.

Roberts, C., & Torgerson, D. J. (1999). Understanding controlled trials: baseline imbalance in randomised controlled trials. *BMJ*, *319*(7203), 185–185.

Rogde, K., Melby-Lervåg, M., & Lervåg, A. (2016). Improving the general language skills of second-language learners in kindergarten: A randomized controlled trial. *Journal of Research on Educational Effectiveness*, *9*(sup1), 150–170.

Sadoski, M., & Willson, V. L. (2006). Effects of a theoretically based large-scale reading intervention in a multicultural urban school district. *American Educational Research Journal*, *43*(1), 137–154.

Scammacca, N., Vaughn, S., Roberts, G., Wanzek, J., & Torgesen, J. K. (2007). *Extensive reading interventions in grades k-3: From research to practice*. Portsmouth, NH: RMC Research Corporation, Center on Instruction.

Shams, L., & Seitz, A. (2008). Benefits of multisensory learning. *Trends in Cognitive Science*, *12*(11), 411–417.

Shams, L., Wozny, D. R., Kim, R., & Seitz, A. (2011). Influences of multisensory experience on subsequent unisensory processing. *Frontiers in Psychology*, *2*(264), 1–9.

Slavin, R. E., Lake, C., Davis, S., & Madden, N. (2011). Effective programs for struggling readers: A best-evidence synthesis. *Educational Research Review*, *6*(1), 1–26.

Snow, C., Burns, S., Griffin, P., & the Committee on the Prevention of Reading Difficulties in Young Children, National Research Council (1998). *Preventing reading difficulties in young children*. Washington DC: The National Academies Press.

Snowling, M., & Hulme, C. (2011). Evidence-based interventions for reading and language difficulties: Creating a virtuous circle. *British Journal of Educational Psychology*, *81*(1), 1–23.

Stanovich, K. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*(4), 360–407.

Statens beredning för medicinsk utvärdering (2014). *Dyslexi hos barn och ungdomar – tester och insatser. En systematisk litteraturöversikt* (SBU-rapport nr 225). Stockholm: Statens beredning för medicinsk utvärdering (SBU). Retrieved from http://www.sbu.se/sv/publikationer/SBU-utvarderar/dyslexi-hos-barn-och-ungdomar--tester-och-insatser/.

StataCorp (2013). *Stata statistical software: Release 13*. College station, TX: StataCorp LP.

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.

Taylor, G., Jungert, T., Mageau, G. A., Schattke, K., Dedic, H., Rosenfield, S., & Koestner, R. (2014). A self-determination theory approach to predicting school achievement over time: The unique role of intrinsic motivation. *Contemporary Educational Psychology*, *39*(4), 342–358.

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*, forthcoming.

Tideman, E., Vinneljung, B., Hinze, K., & Isaksson, A. A. (2011). Improving foster children's school achievement: Promising results from a Swedish intensive study. *Adoption & Fostering*, *39*(1), 44-56.

Torgesen, J. K., Rashotte, C. A., & Wagner, R. K. (1999). *TOWRE: Test of word reading efficiency*. Austin, TX: Pro-ed.

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology*, *53*, 1–25.

Vadasy, P. F., Sanders, E. A., & Peyton, J. A. (2006). Paraeducator-supplemented instruction in structural analysis with text reading practice for second and third graders at risk for reading problems. *Remedial and Special Education*, *27*(6), 365–378.

Vadasy, P. F., & Sanders, E. A. (2008). Code-oriented instruction for kindergarten students at risk for reading difficulties: A replication and comparison of instructional groupings. *Reading and Writing: An Interdisciplinary Journal*, *21*(9), 929-963.

Vadasy, P. F., & Sanders, E. A. (2010). Efficacy of supplemental phonics-based instruction for low-skilled kindergartners in the context of language minority status and classroom phonics instruction. *Journal of Educational Psychology*, *102*(4), 786–803.

Vaughn, S., Roberts, G., Wexler, J., Vaughn, M. G., Fall, A-M., & Schnakenberg, J. B. (2015). High school students with comprehension difficulties: Results of a randomized controlled trial of a two-year reading intervention. *Journal of Learning Disabilities*, *48*(5), 546–558.

What Works Clearinghouse (2014). *Procedures and Standards Handbook Version 3.0.* Retrieved 2017-02-20 from: https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_v3_0_standard-_handbook.pdf.

Weiss, M. J., Lockwood, J. R., & McCaffrey, D. F. (2016). Estimating the standard error of the impact estimator in individually randomized trials with clustering. *Journal of Research on Educational Effectiveness*, *9*(3), 421–444.

Wolff, U. (2011). Effects of a randomised reading intervention study: An application of structural equation modelling. *Dyslexia*, *17*(4), 295–311.

Wolff, U. (2016). Effects of a randomized reading intervention study aimed at 9-year-olds: A 5-year follow-up. *Dyslexia*, *22*(2), 85–100.

Young, A. (2018). Channeling Fisher: Randomization tests and the statistical significance of seemingly significant experimental results. *Quarterly Journal of Economics*, forthcoming.

Appendix

The appendix presents additional descriptive statistics, sensitivity analyses of the short-term effects, additional exploratory analyses of the short-term effects, and sensitivity analyses for the effects of program timing.

5.1 Additional descriptive statistics

Table A1 shows descriptive statistics for the full sample of randomized students; i.e., the table corresponds to Table 2 in the main text but include all students who were randomized and not just the analysis sample. The sample is exactly the same for the follow-up test so we do not repeat those statistics here.

Table A1. Means and standard deviations at pre-, post-, and follow-up test for the full sample

Panel A: Pre-test								
		Treatment			Control			
Variable	*Mean*	*SD*	*n*	*Mean*	*SD*	*n*	*ES*	*p*
Girl	0.39	0.49	82	0.52	0.50	79	-0.13	0.102
Grade	0.20	0.40	82	0.19	0.39	79	0.01	0.933
Specific risk	0.22	0.42	82	0.27	0.44	79	-0.05	0.496
Decoding	0.00	0.00	82	0.00	0.00	79	0.00	1.000
Letter knowledge	4.99	3.75	82	5.00	3.89	79	0.00	0.984
Phonological awareness	4.27	4.20	82	4.99	4.20	79	-0.17	0.279
Self-efficacy	4.81	3.89	81	4.52	3.55	79	0.08	0.624
Enjoyment	8.24	3.21	82	7.34	3.49	79	0.27	0.087
Motivation	8.02	3.51	81	8.22	3.16	79	-0.06	0.701

Panel B: Post-test								
		Treatment			Control			
Variable	*Mean*	*SD*	*n*	*Mean*	*SD*	*n*	*ES*	*p*
Decoding	6.54	7.46	81	1.3	3.55	76	0.89	0.000
Letter knowledge	17.72	6.39	81	10.6	7.12	76	1.05	0.000
Phonological awareness	8.91	2.21	81	7.3	3.86	76	0.53	0.001
Self-efficacy	6.53	3.50	81	4.7	3.11	76	0.56	0.001
Enjoyment	8.43	2.99	81	8.3	3.06	76	0.06	0.717
Motivation	9.12	2.20	81	8.8	2.35	76	0.15	0.349

Note: Mean, standard deviation (*SD*), sample size (*n*), and the difference between treatment and control groups expressed as effect sizes (*ES*). The effect size is for all variables, except Girl, Grade 1, and Specific risk, Hedges' *g*; i.e., the difference between treatment and control group in standard deviations, adjusted for the small sample, see Equation (2). The effect sizes for Girl, Grade 1, and Specific risk is expressed as the differences in shares.

5.2 Sensitivity analyses of the short-term effects

This section shows the sensitivity analysis reported in Section 3.3 of the main text. Table A2 contains variations of our main specification used in the primary analysis, plus the main specification itself for easy reference in column (1), Panel A. For brevity, we show only the treatment coefficient and standard errors for each specification. Each row corresponds to an outcome variable and each column to a specification. In panel A, column (2) excludes all covariates, and column (3) excludes the pair/triple fixed effects. Column (4) includes randomization weights. In panel B, column (5) excludes students who were tested under different circumstances. Column (6) excludes students who were instructed by the third author. Column (7) clusters the standard errors on the instructional groups, and column (8) examines if reduced quality of regular instruction due to increased class sizes are driving our results (note that we omit first grade students from the specification in column 8).

Comparing the estimates in column (2)–(7) to the primary analysis, we can see that they are all close in magnitude. Furthermore, the standard errors and consequently the statistical significance are also similar. We conclude that our results are not sensitive to these changes of the main specification. Column (8) includes only the schools that used special educators as tutors. The treatment effects are larger than in the primary analysis for three out four significant outcomes and larger than the effects for kindergarten students for our two primary outcome measures (compare Table A3, panel B). As class sizes were slightly reduced in these schools (because treated students were pulled-out of classrooms), reduced quality of regular instruction for the control group is an unlikely explanation of these results.

Table A2. Sensitivity analysis for the short-term effects

Panel A: Sensitivity analyses

Outcome variable	**Primary analysis** **(1)**	**No covariates** **(2)**	**No FE** **(3)**	**Weights** **(4)**
Decoding	6.333***	5.267***	5.863***	6.092***
	(1.123)	(0.924)	(0.899)	(1.052)
Letter knowledge	6.911***	7.137***	7.440***	6.978***
	(0.835)	(1.082)	(0.778)	(0.825)
Phonological awareness	1.731***	1.664***	2.075***	1.749***
	(0.454)	(0.506)	(0.457)	(0.456)
Self-efficacy	1.767***	1.873***	1.938***	1.758***
	(0.542)	(0.528)	(0.540)	(0.556)
Enjoyment	-0.240	0.176	0.150	-0.277
	(0.486)	(0.483)	(0.493)	(0.496)
Motivation	0.342	0.341	0.459	0.351
	(0.382)	(0.364)	(0.381)	(0.381)
Pre-tests	Yes	No	Yes	Yes
Pair/Triple FE	Yes	No	No	Yes
Observations	156	157	156	156

Panel B: Sensitivity analyses

Outcome variable	**Test differences** **(5)**	**Instruction differences** **(6)**	**Clustered** **(7)**	**Special educators** **(8)**
Decoding	6.130***	6.217***	6.333***	6.352***
	(1.190)	(1.197)	(1.148)	(1.305)
Letter knowledge	6.925***	6.672***	6.911***	7.294***
	(0.873)	(0.866)	(0.837)	(1.211)
Phonological awareness	1.848***	1.870***	1.731***	1.295*
	(0.508)	(0.471)	(0.458)	(0.709)
Self-efficacy	1.911***	1.691***	1.767***	2.159**
	(0.502)	(0.563)	(0.534)	(0.898)
Enjoyment	0.0472	-0.253	-0.240	-0.653
	(0.504)	(0.494)	(0.489)	(0.810)
Motivation	0.468	0.318	0.342	-0.0702
	(0.408)	(0.414)	(0.384)	(0.400)
Pre-tests	Yes	Yes	Yes	Yes
Pair/Triple FE	Yes	Yes	Yes	Yes
Observations	140	148	156	68

Note: The table displays the treatment coefficient and robust standard errors in parentheses from linear regression models based on Equation (1) but with the following modifications: Column (1) includes the primary analysis for comparison purposes. Column (2) excludes all covariates and pair/triple fixed effects. Column (3) excludes the pair/triple fixed effects. Column (4) includes randomization weights. Column (5) excludes students who were tested under different circumstances. Column (6) excludes students who were instructed by the third author. Column (7) clusters the standard errors on the instructional groups. Column (8) estimates the treatment effects on the kindergarten sample of schools that used special educators as tutors. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

5.2.1 Randomization inference for the short-term effects

The results in Young (2018) indicate that conventional inference methods may often overstate statistical significance in experiments. To assess whether our results are sensitive to the assumptions made on the distribution of the standard errors, we use randomization (or permutation based) inference methods, as suggested by e.g., Young (2018) and Athey and Imbens (2017).

We derive *p*-values for the statistically significant treatment coefficients using the following procedure (we restrict the procedure to the significant coefficient because statistical significance is unlikely to be underestimated in the primary analysis). Step 1: The matching procedure is done exactly as in the analysis of main results. That is, we match the same pairs/triples, including the four students with missing observations. Step 2: We randomly assign treatment within each pair/triple using the same probability of treatment as in the analysis of main results. The probability of treatment is 1/2 in each pair, and 1/3 and 2/3 in the triples where one and two students were assigned to treatment in the main results, respectively. Step 3: Using the new treatment variable, we run regressions corresponding to those in Table 4 and register the absolute value of the *t*-statistic of the treatment coefficient from each regression (to get a two-sided test).[43]

Step 2 and 3 are repeated 10,000 times to obtain an empirical distribution of *t*-statistics. We use the *t*-statistic instead of e.g., the beta-coefficient, as recommended by Young (2018) and Mackinnon and Webb (2016). The *p*-values obtained from this procedure are then equal to the distribution rank of the *t*-statistics of the coefficients divided by 10,000. The results indicate that our statistically significant estimates of the short-term effects of *Läsklar* are robust. There is no configuration that yield a larger *t*-statistic than the primary analysis for the decoding, letter knowledge, and phonological awareness tests (i.e., rank = 1 and p = 0.0001), and 19 configurations yield larger *t*-statistics for the self-efficacy measure (rank 20 and p = 0.0020).

[43] Students with incomplete pre- or post-tests are therefore included in the matching procedure but not in the regressions, just as in the primary analysis.

5.3 Additional exploratory analyses of the short-term effects

Table A3 contains three additional exploratory analyses that we did not pre-register and which are described in Section 3.4.

Table A3. Not pre-registered exploratory analyses of short-term effects

Panel A: Are the effects larger for students with specific risks?						
		Letter	**Phonological**	**Self-**		
	Decoding	**knowledge**	**awareness**	**efficacy**	**Enjoyment**	**Motivation**
Variables	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Treatment	7.298***	6.937***	1.285**	2.088***	-0.810	0.268
	(1.446)	(1.061)	(0.553)	(0.699)	(0.498)	(0.401)
Treatment *x*	-3.850*	-0.607	1.701	-1.133	2.025	-0.0283
Specific risk	(2.225)	(2.557)	(1.479)	(1.627)	(1.618)	(0.974)
p(TME)	0.0378	0.0042	0.0192	0.4711	0.4079	0.7920

Panel B: Are the effects larger in first grade?						
		Letter	**Phonological**	**Self-**		
	Decoding	**knowledge**	**awareness**	**efficacy**	**Enjoyment**	**Motivation**
Variables	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Treatment	5.352***	7.392***	1.830***	2.174***	-0.236	0.197
	(0.945)	(0.940)	(0.568)	(0.646)	(0.609)	(0.465)
Treatment *x*	4.384	-2.150	-0.441	-1.818	-0.0178	0.648
Grade 1	(3.011)	(2.210)	(0.944)	(1.114)	(0.783)	(0.663)
p(TME)	0.0020	0.0092	0.0528	0.6946	0.6211	0.0935

Panel C: Are the effects larger in schools with prior experience?						
		Letter	**Phonological**	**Self-**		
	Decoding	**knowledge**	**awareness**	**efficacy**	**Enjoyment**	**Motivation**
Variables	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Treatment	5.890***	6.323***	2.262***	1.375**	-0.317	0.397
	(1.477)	(1.058)	(0.582)	(0.620)	(0.580)	(0.528)
Treatment *x*	1.254	1.665	-1.503*	1.110	0.218	-0.157
Prior experience	(1.981)	(1.592)	(0.864)	(1.148)	(1.048)	(0.680)
p(TME)	0.0000	0.0000	0.2530	0.0140	0.9108	0.5989

Panel D: Are the effects larger in schools with any experience?						
		Letter	**Phonological**	**Self-**		
	Decoding	**knowledge**	**awareness**	**efficacy**	**Enjoyment**	**Motivation**
Variables	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Treatment	6.374***	6.733***	2.177***	1.352**	-0.0393	0.625
	(1.622)	(1.164)	(0.693)	(0.655)	(0.724)	(0.667)
Treatment *x*	-0.0827	0.358	-0.895	0.832	-0.403	-0.569
Any experience	(1.798)	(1.673)	(0.841)	(1.036)	(0.963)	(0.748)
p(TME)	0.0000	0.0000	0.0182	0.0111	0.4935	0.8759
Pre-tests	Yes	Yes	Yes	Yes	Yes	Yes
Pair/triple FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	156	156	156	156	156	156

Note: The table displays coefficients, robust standard errors (in parentheses), and the p-values on the total marginal effects (TME) of the interactions from linear regression models including pre-tests and pair/triple fixed effects as covariates (the pair/triple fixed effects are partialled out). Panel A examines whether the effects are larger or smaller for students with specific risks. Panel B shows results where we include an interaction between the treatment indicator and an indicator for being in first grade. Panel C and D reports tests of whether schools with prior and more experience have larger effects. *** p < 0.01, ** p < 0.05, * p < 0.1.

Panel A examines whether the effects are larger or smaller for students with specific risks. Panel B shows results where we include an interaction between the treatment indicator and an indicator for being in first grade. Panel C and D report two tests of whether schools with prior or more experience have larger effects.

We find only weak evidence of heterogeneity. No interaction effect is significant at the 5 percent level, and only two on the 10 percent level (students with specific risks have lower decoding scores and students being tutored by tutors with prior experiences have lower scores on phonological awareness).

5.4 Sensitivity analyses for the effects at follow-up

Table A4 displays the results of the sensitivity tests for the effects at follow-up. All estimates have a similar sign across specifications and most are of reasonably similar magnitude as in the primary analysis. A few estimates become significant: the treatment group has significantly higher scores on decoding when we apply randomization weights ($p < 0.10$), on phonological awareness when we do not include pair/triple fixed effects ($p < 0.05$) and when we exclude students instructed by the third authors ($p < 0.10$). The self-efficacy and enjoyment measures are significantly higher in the control group when we exclude all covariates and when we exclude pair/triple fixed effects ($p < 0.05$). However, no estimate is consistently significant across specifications. In sum, the evidence of differences between the treatment and control group at follow-up is relatively weak.

Table A4. Sensitivity analysis for the effects at follow-up

Outcome variable	Primary analysis (1)	No covariates (2)	No FE (3)	Weights (4)	Instruction differences (5)	Clustered (6)
Decoding	3.314	1.879	2.752	3.453*	3.477	3.314
	(2.016)	(2.002)	(1.811)	(2.060)	(2.152)	(2.041)
Letter	0.433	1.060	1.306	0.444	0.627	0.433
knowledge	(0.734)	(1.029)	(0.930)	(0.720)	(0.761)	(0.728)
Phonological	0.488	0.452	0.676**	0.484	0.513*	0.488
awareness	(0.294)	(0.299)	(0.323)	(0.295)	(0.303)	(0.326)
Self-efficacy	-0.894	-1.381**	-1.226**	-0.961	-0.854	-0.894
	(0.611)	(0.560)	(0.554)	(0.596)	(0.627)	(0.627)
Enjoyment	-0.818	-1.139**	-1.125**	-0.708	-0.685	-0.818
	(0.594)	(0.475)	(0.471)	(0.604)	(0.606)	(0.676)
Motivation	-0.366	-0.624	-0.417	-0.396	-0.335	-0.366
	(0.437)	(0.425)	(0.447)	(0.443)	(0.458)	(0.478)
Pre-tests	Yes	No	Yes	Yes	Yes	Yes
Pair/triple FE	Yes	No	No	Yes	Yes	Yes
Observations	141	141	141	141	133	141

Note: The table displays the treatment coefficient and robust standard errors in parentheses from linear regression models based on Equation (1) but with the following modifications: Column (2) excludes all covariates; column (3) excludes the pair/triple fixed effects; column (4) includes randomization weights; column (5) excludes students who were instructed by the third author; and column (6) clusters the standard errors on the instructional groups.. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.