HCEO WORKING PAPER SERIES

Implementing Educational Interventions at Scale

Simon Calmar Andersen
Ulrik Hvidman

Working Paper

2020-039

05/2020; Revised 05/2024

HUMAN CAPITAL AND
ECONOMIC OPPORTUNITY
GLOBAL WORKING GROUP

The University of Chicago
1126 E. 59th Street Box 107
Chicago IL 60637

www.hceconomics.org

Implementing Home-Based Educational Interventions at Scale

Simon Calmar Andersen & Ulrik Hvidman*

May 15, 2024

Abstract

Although many educational programs have demonstrated the potential to increase student learning, few examples of successful scaling exist. We study the scalability of a home-based reading program that has shown promising results in an experiment within a local government. Using a nationwide experiment among the full population of 2nd-grade children in Danish public schools (n=51,312), we find that the intervention is less effective at large scale. We provide evidence on potential explanations for the lack of scalability, which suggests that low levels of take-up among both schools and parents were the most important barriers to successful scaling.

Keywords: implementation; education; experiments.

*Andersen: Department of Political Science and TrygFonden's Centre for Child Research, Aarhus University (email: sca@ps.au.dk); Hvidman: The Danish Center For Social Science Research (email: ulhv@vive.dk). The authors thank Hans Henrik Sievertsen, Phillip Heiler, and Kurt Houlberg for valuable comments and suggestions. We also thank Morten Bruntse and Marlene Vita Kristensen for outstanding research assistance on this project. Finally, we thank three anonymous reviewers for helpful comments. Andersen and Hvidman acknowledge financial support from TrygFonden for this project. This paper uses confidential data from Statistics Denmark that can be obtained by filing a request to Statistics Denmark through an authorized institution.

Parents play an important role in the educational development of their children. An increasing number of field experiments have demonstrated that interventions aimed at involving parents in supporting their children's learning can have positive effects (e.g., Andersen and Nielsen, 2016; Bergman, 2021; Bergman and Chan, 2021; Bergman and Rogers, 2017; Doss et al., 2019; York, Loeb and Doss, 2019). The evidence on the effectiveness of these early interventions suggests that there could be substantial returns from investing in parent-targeted educational programs at a large scale. However, results from small-scale experiments may not generalize to a large scale (Al-Ubaydli, List and Suskind, 2017).[1]

An important next step for the research on early interventions is to systematically study barriers to scalability. First, subjects who experience the largest effects of the interventions may be more likely to select into experimental studies (Heckman, 1992, 2020; Heckman and Smith, 1995; Allcott, 2015). Second, results from small-scaled, published trials may not replicate because, conditional on being published and showing statistically significant effects, there is a relatively high probability that the trials overestimate the effect sizes (Gelman and Carlin, 2014). Third, changing the scale of the intervention may also affect the implementation of the program (Duflo, Glennerster and Kremer, 2008).[2] Unfortunately, the use of science to systematically compare these "scalability" problems is at such an early stage that little is known about the relative importance of these problems (Czibor, Jimenez-Gomez and List, 2019).

To study these scalability problems, we fielded an experiment of a reading program called

[1]The fact that the effectiveness of interventions may depend on scale is supported by an emerging body of empirical studies that provide evidence on how effects from smaller-scale behavioral interventions may not replicate at scale (DellaVigna and Linos, 2022; Ganimian, 2020; Kizilcec et al., 2020)

[2]These and related challenges to scaling are also discussed by Banerjee et al. (2017). Moreover, research in developing countries has shown that nationwide implementation tends to reduce effect sizes. In a meta-analysis, Vivalt (2020) shows that effects of randomized controlled trials tend to be smaller when the intervention is implemented by governments rather than non-governmental organizations (NGOs). Bold et al. (2018) randomized whether the same intervention was implemented by an NGO or by the national government in Kenya. While they found positive effects when the NGO implemented the policy, effects were small and not statistically significant when implemented by the government.

READ. READ aims at improving 2nd-grade students' literacy skills. Families receive books as well as information that contains three messages. First, the information emphasizes a growth theory of abilities by explaining to the parents that their child's reading ability can be improved regardless of the child's literacy skills. Second, the material encourages parents to support the child's autonomous engagement with the books by asking parents to talk to the child about the content. Third, the parents are encouraged not to correct their child if it reads incorrectly, unless it affects the child's understanding of the text.

The national government in Denmark invited by email a random sample of the full population of schools with more than 50,000 2nd-grade students to participate in READ. The intervention had previously demonstrated positive effects on 2nd-grade students' reading and writing skills in a field experiment run by a local government (Andersen and Nielsen, 2016). This original trial included 1,587 children in 72 classrooms that were randomly assigned to the READ program or a control condition (i.e., business as usual) and found that the program improved standardized test scores in reading by 0.26 standard deviations after two months and 0.12 standard deviations after seven months.

To analyse the scale-up of READ, we use administrative data collected independently of the experiment to measure the effect on students' reading skills as well as the socio-economic characteristics of the full population of schools and families. These features make the trial a natural field experiment in the typology of Harrison and List (2004), meaning that it was conducted in the natural environment in which it would be used if implemented nationally, that there was no self-selection into the trial, and that subjects were unaware that researchers evaluated the effects of the program.

Results show no statistically significant intention-to-treat effect when the intervention is scaled up (measured on average three months after the beginning of the intervention), and the estimate is close to zero and precise enough to rule out even modest effects. To understand this finding, we examine the three scalability problems. First, the features of the natural field experiment allow us to study which schools opted into the program and compare participants in the original, local

trial to the national trial. Interestingly, in the national trial, the groups of schools and parents that opted into the program were not much different in terms of socio-economic status and previous test scores in comparison to those who opted out. Moreover, treatment effects did not seem to be heterogeneous with respect to socio-economic background. Since the original, local study was also conducted among a relatively large sample of students with variation in both socio-economic background and ethnicity, selection into the original trial does not appear to explain the scalability problem.

To study the replicability, we evaluated the effect of the program when the local government, which ran the first READ trial, subsequently put the program into operation. We use a difference-in-differences (DiD) design to compare schools within the local government that chose to either adopt or not adopt the READ program for a new cohort of students. We find effects that are of similar size as in the original, experimental trial. These results suggest that the program has the potential to improve student learning and that the lack of scalability was not merely due to a statistical artifact.

Instead, low levels of take-up at both the school and the family level appear to be important barriers for the effectiveness of the program. Only 31 % of the invited schools accepted the invitation for the program. Yet, even within the participating schools in the national trial, take-up among families was low. Data from a smartphone app—in which invited parents could sign up—provide a behavioral measure of take-up. These data demonstrate that the local government succeeded in making twice as many parents sign up in the app (24 % in the nationwide experiment, and 48 % in the local government intervention). Of course, the relatively low levels of take-up in the national trial—both at the school level and among families within the participating schools—reduce the statistical power of the study and may provide one explanation for the lack of positive results in the national trial. However, supplementary analyses may suggest that the program was effective once it was implemented appropriately. Thus, we find that test scores were higher among treated schools in the national trial with a relatively large proportion of parents downloading the app—a finding that is robust to various specifications.

Qualitative information on the implementation process may suggest that the way the program was implemented is an important factor to explain the differences in the results between the local and the national setting. Notably, the local implementation differed from the nationwide implementation in important ways (e.g., administrative support from the municipality). Thus, the results of the analyses, coupled with these qualitative insights, suggest that implementation is crucial for scalability. Furthermore, our findings indicate that successful implementation is not solely achieved by securing administrative support among schools but also requires active engagement with families.

The remainder of the article is organized as follows. In the second section, we review and analyze existing research on parent-aimed and other educational interventions. The third section presents the results of the large-scale randomized trial at the national level in Denmark. The fourth section studies potential explanations of scalability problems. The final section concludes.

1 Parent-Aimed Programs at Scale

1.1 Evidence on the Effectiveness of Parent-Aimed Programs

Given that family investments and resources matter for children's skill development, one important question is whether and how governments can support and encourage parent engagement. A growing body of experimental studies supports the notion that parent-directed interventions can increase parental involvement in their children's learning and that they have the potential to improve child learning. Although parent-aimed interventions often try to enhance ability, knowledge, and motivation, initiatives differ in their focus on providing additional learning material (i.e., resources) and merely providing information to families.

One set of parent-aimed programs provide resources such as books or tablets to families in order to encourage them to enact learning activities at home. A number of parent-aimed programs have been successful in encouraging parents to read or do math with their children—often in close collaboration with their teachers. In a randomized controlled trial including 284 immigrant chil-

dren from 61 child care centers, Jakobsen and Andersen (2013) found that providing families with children's books and games (and collaboration with pre-school teachers on the development of their children) increased the language test scores of children with low-educated mothers as assessed by their pre-school teachers. In the study of the READ program, on which we build the current study, Andersen and Nielsen (2016) found that the program led to an increase in students' standardized test scores in reading of 0.12-0.26 standard deviations. The family-aimed treatment consisted of books and information on how to use dialogue-based reading. 1,587 students from 28 schools (within one local government) participated in the randomized controlled trial. Berkowitz et al. (2015) administered a tablet app to parents that helped them do math activities at home with their 1st-grade children. The study included 22 schools and 587 families and found that the more parents used the app, the better the child performed in a math test administered by trained researchers.

Taken together, these studies suggest that when parents are encouraged to read or do math with their children at home, there is a potential for improving children's learning. In contrast, Guryan, Hurst and Kearney (2008) did not detect a positive effect of a reading intervention in their experimental study, even though the intervention resembles the other programs previously described. In their study, books were sent to students during the summer to encourage them to keep reading during the school break. Parents were invited to an after school family literacy event where they learned about the program. Including 5,319 students from 59 schools, this study is relatively large, which may be an indication of the challenges in scaling up programs that provide books and other resources to the families and facilitates collaboration between schools and families.

Another group of parent-aimed interventions focus on school-to-parent communication. Since the cost of sending information to parents is often lower than providing parents with learning materials, it is generally cheaper to scale up information-based interventions. In a study of 1,031 parents from one school district, York, Loeb and Doss (2019) found that sending parents text messages with advice on how to support the development of their children improved the pre-school children's early literacy by about 0.11 standard deviations. In another experiment, Bergman (2021)

studied the impact of emails, text messages, and phone calls from teachers to parents with information about missed assignments and grades. The effect among the 462 participating students (all from one school) was an increase in the grade point average (GPA) of about 0.20 standard deviations. In a related study, Bergman and Chan (2021) used automated text messages to scale up the intervention (i.e., sending parents information about missed assignments and grades). In a sample of 22 schools (1,137 students), they found positive effects on GPA but did not find significant effects on state-administered test scores.

In a randomized controlled trial among 6,976 students in 12 schools, Bergman and Rogers (2017) use automated text message alerts to inform parents if their child had a missing assignment, a class absence, or a low average course grade. They found effects of information on GPA of about 0.06 of the standard deviations of the control group at baseline. In a study, Rogers and Feller (2018) sent parents of 28,080 12^{th}-grade students (from one large school district) information about their children's school absences. The most effective treatment in the experiment reduced absences by 1.1 day, a reduction of 6.5 percent compared to the control group. However, they were not able to detect a statistically significant effect on test scores.

1.2 Scalability of Educational Programs

To more systematically examine the relationship between scale and the effectiveness of education interventions, we conducted a meta-analysis of randomized controlled trials that examine the effect on standardized test outcomes.[3] In Figure 1, we plot the number of participants in the studies on a logarithm scale against the estimated, standardized effect sizes. Three patterns in Figure 1 are

[3]Our analysis is based on three data sources. First, we included the studies from Noble et al. (2019), who systematically review parent-aimed programs targeting shared book reading. Second, we included studies from Lortie-Forgues and Inglis (2019), who study all interventions commissioned by the Education Endowment Foundation (EEF) in the UK and the National Center for Education Evaluation and Regional Assistance (NCEE) in the US . Third, we supplemented these two meta-analyses with a systematic search for parent-aimed education interventions. A more detailed description of the search strategy can be found in Appendix A. A full list of all data and publications included in our meta-analysis is available in a separate supplementary materials document.

worth emphasizing. First, the larger studies tend to produce smaller effects on standardized student outcomes. For studies with around 1,000 or more participants, the average effect size is close to zero. Although the interventions and study samples differ in several respects, this relationship could indicate that there are some challenges in scaling up such educational interventions. Second, the scale of the program evaluations have typically been rather small. Third, there is variation in scale across type of intervention. Many evaluations of parent-aimed programs have included less than 300 participants, whereas the educational interventions included have been tested at larger scales, although seldom with more than 10,000 individuals. These results support Kraft (2020), who suggests that effect sizes should be evaluated in relation to the scalability of the educational interventions and that for low-cost, scalable interventions, effect sizes as small as 0.05 standardized achievement outcomes may be valuable.

Overall, the meta-analysis demonstrates that there are few successful examples of introducing parent-aimed educational programs at scale. Moreover, the negative relationship between the size of the study sample and the effect size may suggest that scaling educational programs is difficult. In sum, we see a growing body of evidence that small-scale parent-aimed interventions provide a promising strategy for delivering improvement in student learning. However, there is little empirical evidence as to what are the most important barriers for scale-up.

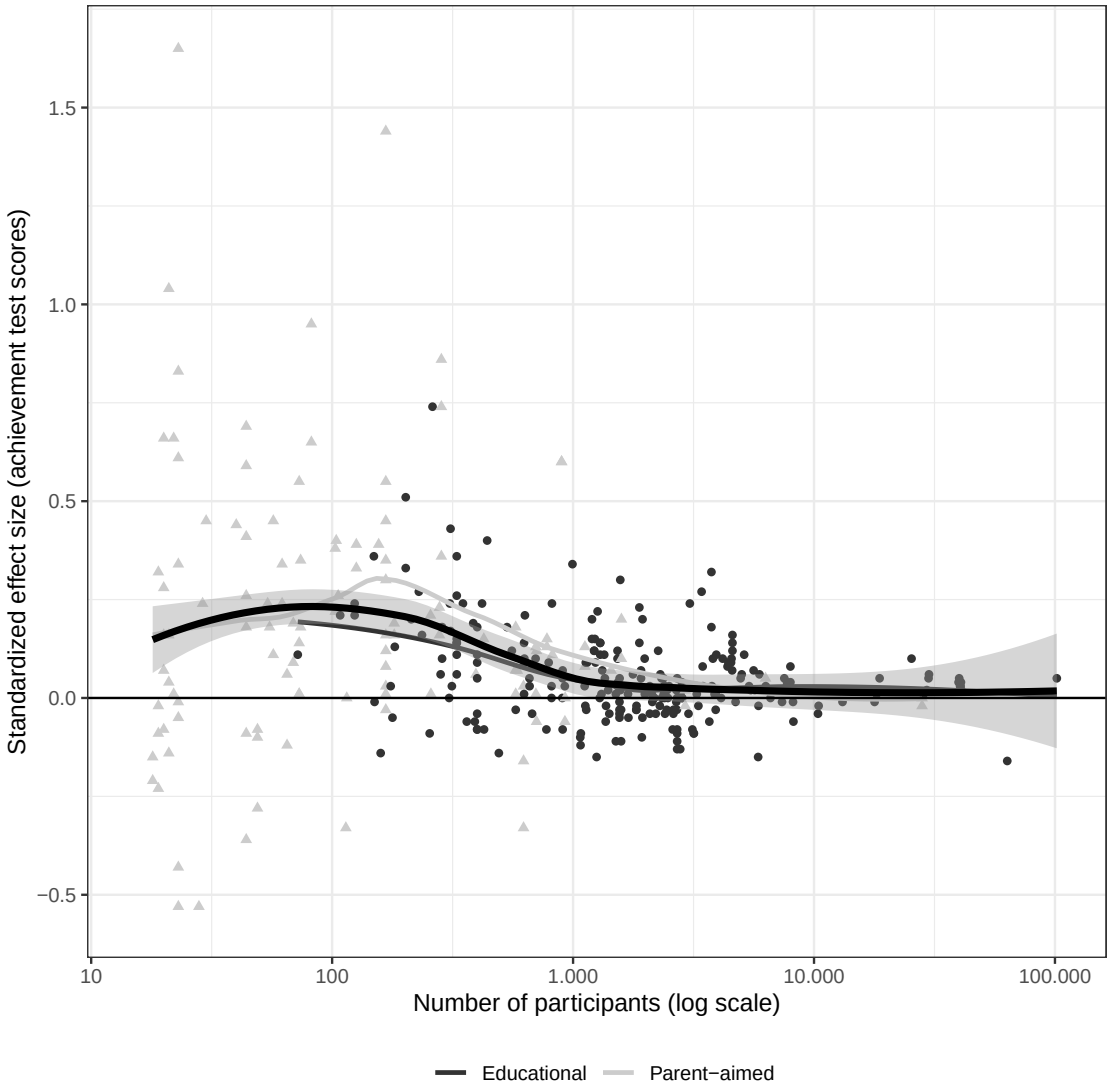


Figure 1: Number of Participants (Log Scale) against Standardized Achievement Test Score Effect Sizes in Randomized Controlled Trials of Parent-Aimed and General Educational Interventions.

Note: Black dots designate general educational interventions commissioned by the Educational Endowment Foundation and the National Center for Educational Evaluation and Regional Assistance (Lortie-Forgues and Inglis, 2019, see Appendix A for details). Grey triangles indicate parent-aimed interventions (Noble et al., 2019, and own review, see Appendix A). The bold black line and the shaded area represent local polynomial regression with 95 % confidence intervals for all observations. The thin grey line and the thin black line are local polynomail regressions based on the general educational and the parent-aimed interventions, respectively.

2 A Full-Scale Natural Field Experiment

2.1 The READ Program

In collaboration with the Ministry of Education in Denmark, we conducted an experiment of the READ program among the full population of Danish public schools with 2nd-grade students.

READ was developed by a team of educational researchers at VIA University College and TrygFonden's Centre for Child Research in collaboration with the local government of Aarhus (Aarhus is the second-largest city in Denmark with approximately 350,000 inhabitants). The program aims at improving 2nd-grade students' literacy skills. As part of the program, families receive four books[4] and information on how to find other reading material at the library, at the school, or in newspapers. Parents are also provided with a booklet and access to an online video (all information was translated into 10 languages).

The booklet and the video based on three key components. First, they underscore a growth theory of abilities, explaining to parents that their child's literacy skills can be enhanced irrespective of the current level (Dweck, 1999, 2006). Second, the materials advocate for a constructive, mastery-oriented approach, urging parents to support their child's autonomous engagement with books (Pomerantz, Moorman and Litwack, 2007; Moorman and Pomerantz, 2010). Third, the parents are encouraged not to correct their child's reading, unless it interferes the child's comprehension of the material (Haimovitz and Dweck, 2016).

2.2 The Original, Local Experiment

A previous study in the local government of Aarhus yielded encouraging results (Andersen and Nielsen, 2016). The randomized controlled trial included 1,587 children in 72 classrooms from 28 schools. Based on cluster randomization at the classroom level, the 1,587 children were assigned to treatment (i.e., READ) or control (i.e., business as usual). The duration of the program was 16

[4]Due to changes in the available budget and the available books, we included four books (compared to three in the original trial).

weeks, and the average costs per child approximately DKK 500 (USD 76).

The READ treatment improved standardized test scores in reading significantly, with an estimated effect size of 0.26 standard deviations after two months and 0.12 standard deviations after seven months. The treatment also improved children's expressive language skills as measured by a writing test by 0.16 standard deviations.[5]

2.3 Experiment at a Nationwide Scale

To study the effectiveness of READ at a nationwide scale, we worked with the Danish Ministry of Education to randomly assign all Danish public schools to receive the READ program or to a control condition (i.e., business as usual). In Denmark, most children are enrolled in basic education in the summer of the year they turn six. Danish basic schooling covers a preschool year and nine years of compulsory education. Although parents can choose to enroll their children in a self-governing school or educate them at home, most children attend a public school (in 2017, 79 %). Schools are governed by 98 local governments—comparable to school districts in the US—but the national government (i.e., the Ministry of Education) formulates the general rules and can initiate policies for all public schools.

Population and Randomization

The Danish Ministry of Education provided a list that included all public schools with 2nd-grade students in the school year 2017/2018.[6] As the City of Aarhus was implementing the READ

[5]We can only speculate which components of the intervention were driving the effect. We think that a key ingredient of the intervention was the advice to parents on the value of shared book reading and guidance on how to do so. The original trial found larger effects for children of low-educated parents that, at baseline, held the belief that reading skills tended to be fixed. However, the books in the intervention may play a vital role in getting parents' attention. Only providing advice to parents (online or in a booklet) may not have the same effect if it does not gain their attention (for a thorough description of the intervention and the results, see Andersen and Nielsen, 2016).

[6]We excluded schools without 2nd-grade students and self-governing schools not governed by a local government.

program among a subgroup of their schools simultaneously with the national study, we excluded all schools from the City of Aarhus from the randomization (N = 46).[7] We end up with a sample of 1,142 public schools. Figure 2 presents the enrollment and participant flow of the experiment.

[7] In section 3.2, we describe the observational replication study that evaluated the new implementation by the local government of Aarhus.

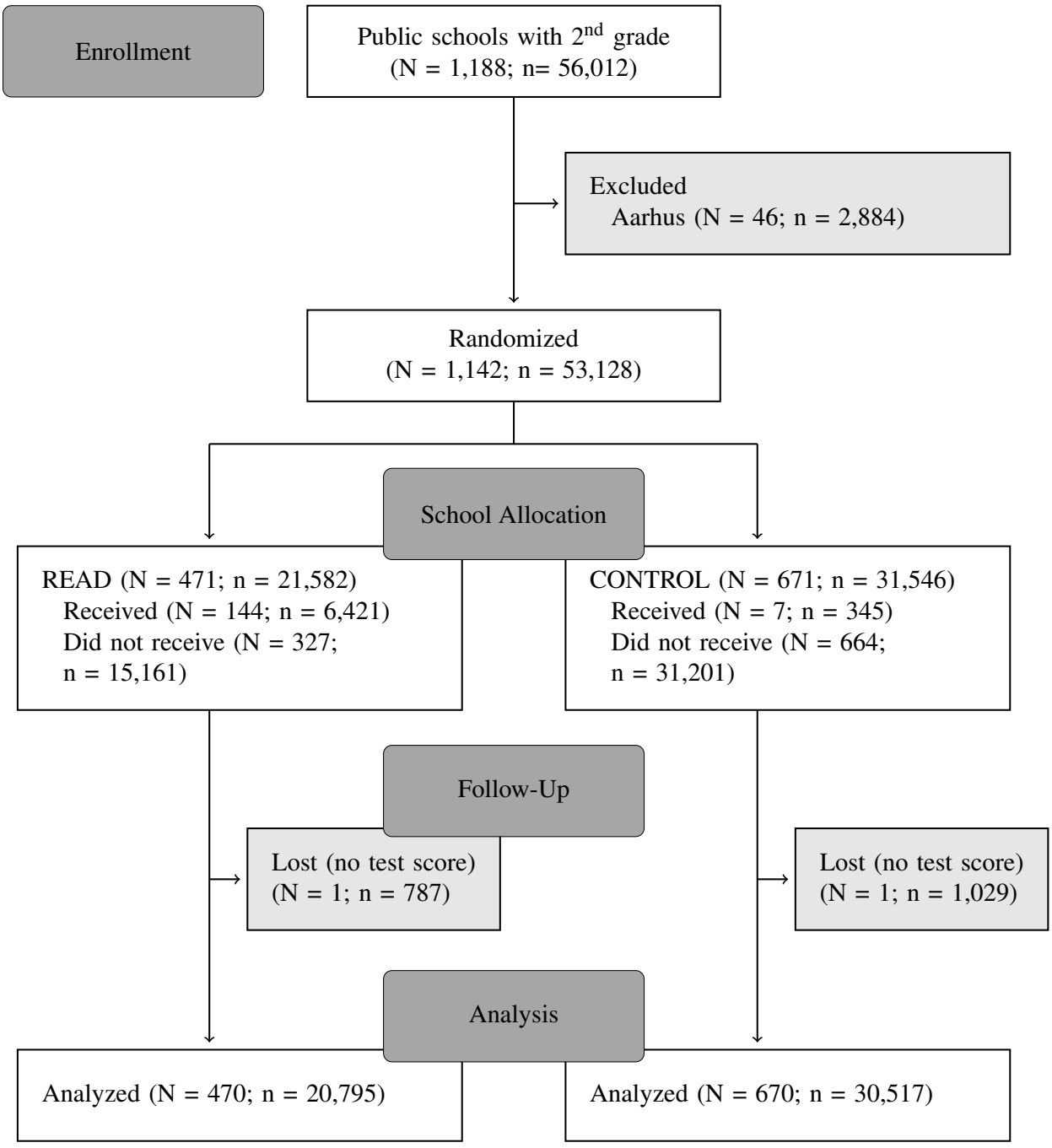


Figure 2: Participant Flow Diagram of the Experiment at Nationwide Scale

Note: N: Number of schools. n: Number of students. "Received": Schools that adopted READ. "Did not receive": Schools that did not adopt READ.

The experiment was set up as an encouragement design (see Angrist and Pischke, 2009) in which schools were invited to participate in the READ program. Using a cluster-randomized design, we randomly assigned students in 1,142 schools to two conditions. The first group (our control group) did not receive an invitation to receive the READ program. The second group (READ assignment schools) received an official e-mail from the Ministry of Education with an invitation to receive READ for all 2nd-graders at the school.

The invitations were sent out in three waves. This procedure ensured that we did not exceed the budget by attracting too many adopting schools. Specifically, we generated a random number for each school and sorted the list of schools accordingly. Using this randomly sorted list, the first 188 schools (16.5% of the total 1,142 schools) received invitations in the initial wave. Subsequently, the following 189 schools (16.5%) on the list were invited in the second wave. As we had additional budgetary resources after the first two waves were sent out, the subsequent 94 schools (8.2%) on the list were invited in a third wave. In total, 471 schools (41.2%) were randomly assigned to the program, while the remaining 671 schools on the list (58.8%) constitute our control group. In our main analysis, we pool the three waves and compare those randomly assigned to the program against the remaining schools. However, we also estimate the effect separately for each of the three waves and find that the results are very similar across waves (see Appendix Table C.2).

Invitations were sent out during the period from September 6 to September 27, 2016. Figure 3 provides a timeline of the implementation of the READ program. The median time from the beginning of the intervention to the test of the students was 94 days (average 84 days). The READ program was provided to all schools that accepted the invitation from the Ministry of Education.

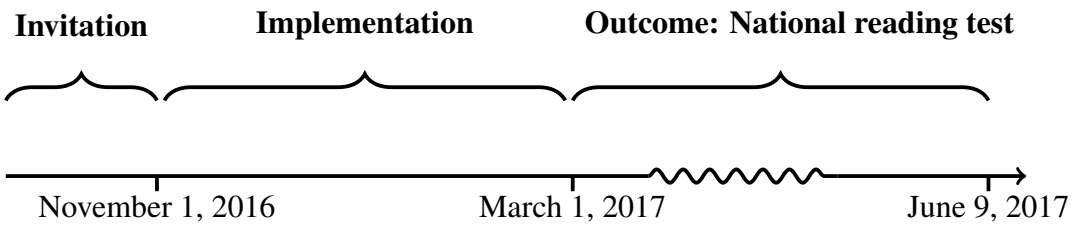


Figure 3: Timeline

The READ Program

Similar to the first READ trial, the schools were responsible for the distribution of the bags with the READ material to the families. Thus, the role of teachers was to distribute the READ material in the classroom and encourage students and their families to engage in the program.

The bags contained the original READ material, including four books to get parents started as well as information on how to find other reading material at the library. They also received a booklet and access to a video that underpin the three learning components of READ. In addition to the original READ program, parents at participating schools were also provided log-in information to a mobile application. The READ app contained video material with information on the importance of reading with children, and families could also use the app to register their reading activities and track their development (Appendix D, Figures D.3a and D.3b show screen shots from the app).

One advantage of the encouragement design using the full population of schools is that it allows us to implement the program in a way that mimics a real-world setting. As a token of appreciation for their effort, the Ministry of Education offered schools DKK 2,500 (USD 375) for participating. Schools were reminded by e-mail about the program approximately one week after the invitations were sent out. A consultant also contacted the schools to inform them about the program and remind them about the decision to participate.The local governments were informed about the project and that some of their schools had been invited. However, in contrast to the first trial—which was initiated by Aarhus Municipality—the local governments were not an active part of the implementation.

To study the implementation of the program, two experiments were embedded in the trial. First, one experiment tested two slightly different versions of the invitation letter. Second, the implementation in the original READ program included a social reward component. Specifically, to encourage the child's effort, parents and children could use a logbook to note every time they read. The logbook thereby endorsed the child's effort rather than performance or results (i.e., speed and accuracy). When the child had read ten times, she could bring the logbook to her teacher, and the class would get a sticker. The class with the most stickers received a reward. To

test the importance of this social reward component, participating schools in the nationwide trail were randomized to one of two versions of the program (i.e., one set of schools participated in a competition for a reward based on their reading, whereas the other set of schools did not receive this incentive). We found no strong effect of either of these variants. In the following analyses, we pool these subconditions (for further details, see Appendix D).[8]

Outcome

The main outcome is standardized test scores in 2nd-grade national reading tests. As of 2010, all students in public schools are tested in ten mandatory tests during basic education in grades 2 through 8. As the test result for each student is confidential and only known by the student's subject-specific teacher, tests are relatively low stakes. From 2015, however, test results at the school level are used as a soft accountability tool with potentially higher stakes for schools.

The 2nd-grade reading test constitutes a good outcome measure for three reasons. First, the tests are IT-based and performed by the students in class on a computer, and the scoring procedure is standardized (i.e., the score is automatically generated within the test system). The standardized procedure ensures that teachers and students cannot manipulate the test result, and the scoring of the tests is thereby blinded to the schools and students' treatment status in the trial. Second, tests have been shown to be a strong predictor of later-stage educational outcomes (Beuchert and Nandrup, 2018). Third, the reading test is divided into three subtests: "Language Comprehension," "Decoding," and "Reading Comprehension," which allows us to study different parts of literacy.[9]

[8] A power analysis conducted prior to the experiment showed that a total of 65 schools in each treatment arm would allow us to detect an effect size of 0.15 standard deviations in test scores. This power calculation is conditional on a power of 0.80, a significance level of 0.05, an intra-school correlation of 0.10, and baseline data accounting for 0.20 of the variation in the outcome. To provide sufficient power to be able to detect an effect size of 0.15 standard deviations in the embedded experiment that tested two versions of the program, we aimed for 130 schools to accept the invitation.

[9] The underlying psychometric model for the test is a Rasch model. The test score for each of the three subtests is measured on a logic scale (for further details, see Beuchert and Nandrup, 2018). We standardize each of the logit scores with mean zero and a standard deviation of one. To compute the overall score, our main outcome, we take the mean of the three standardized subtest

Test scores were measured about three months (i.e., 84 days on average) after the beginning of the intervention.

Data

The Danish administrative registers allow us to track the full population of all 2nd-grade students in a public school. We match each student to the school that they were enrolled in when the program was implemented (September 6, 2016). Our main data consists of 53,128 students in our population of 1,142 schools.

We merge our main data with additional data sources. First, the Ministry of Education provided student-level test data on the 2nd-grade tests in reading. Second, the data are linked with records from Statistics Denmark containing detailed information on the children, including the child's ethnicity, gender, and age as well as on their families (e.g., the parents' length of education). The parental characteristics are measured in 2014 (two years prior to the intervention). Third, we are able to track whether parents download and register the READ application that was part of the program to measure parents' adoption of the program.

Figure 2 illustrates the experimental design and presents data on compliance with the experimental protocol and attrition. Attrition, which is usually a main threat to the internal validity in experiments, is limited because the 2nd-grade reading tests are compulsory. We observe test scores on 96.6 % of the students, and importantly, there is no evidence of systematic differences in attrition between treated and controls. We define our analytical sample as the 51,312 students for which we observe the test scores.

Descriptive Statistics and Balance

In Table 1, we provide descriptive statistics for our analytical sample across experimental conditions. We include the same variables as in the original study by Andersen and Nielsen (2016). Because of the random assignment of the READ invitations, there should be no systematic differ-

scores and, subsequently, standardize this average score.

ences in the distribution of covariates between schools assigned to the control—Column (1)—and treatment—Column (2)—conditions. Column (3) compares invited and non-invited schools. All differences are substantially small with no tests significant at the 5-% significance level. Table B.1 in the Appendix shows the balance on the full set of variables.

Table 1: Differences in Mean between Invited and Non-Invited on Background Characteristics

	(1) Non-invited	(2) Invited	(3) 1-2	*p*-values
Child is a boy	0.52	0.52	-0.00	0.68
Child's age in 2016	8.09	8.09	0.00	0.41
Child immigrant	0.10	0.12	0.01	0.07
Child living with both parents	0.72	0.71	-0.01	0.25
Child living with single parent	0.20	0.21	0.01	0.10
Child living with parent in new relationship or not living with own parents	0.08	0.08	-0.00	0.49
No. of children in family	2.31	2.33	0.02	0.17
Mother compulsory education	0.14	0.14	0.01	0.16
Mother upper seconday education	0.05	0.05	-0.00	0.58
Mother vocational education	0.31	0.31	-0.00	0.94
Mother short-cycle education	0.05	0.05	-0.00	0.09
Mother medium-cycle education	0.27	0.27	0.00	0.97
Mother long-cycle education	0.13	0.13	-0.00	0.66
Missing on Mother's education (6 categories)	0.06	0.07	0.01	0.22
Mother employed	0.79	0.78	-0.01	0.06
Mother unemployed	0.04	0.04	0.00	0.25
Mother outside labor market	0.17	0.18	0.01	0.07
Missing on Mother's employment status (3 categories)	0.02	0.03	0.00	0.08
Mother's total income (1000 DKK)	263.12	256.89	-6.23	0.19
Missing on Mother's total income (1000 DKK)	0.02	0.02	0.00	0.10
Mother's age in 2014 (years)	38.76	38.63	-0.14	0.11
Missing on Mother's age in 2014 (years)	0.02	0.02	0.00	0.08
Mother is teenager at date of birth	0.01	0.01	0.00	0.67
Father compulsory education	0.16	0.17	0.01	0.16
Father upper seconday education	0.06	0.05	-0.00	0.08
Father vocational education	0.41	0.41	0.00	0.97
Father short-cycle education	0.08	0.08	-0.00	0.70
Father medium-cycle education	0.14	0.13	-0.00	0.45
Father long-cycle education	0.13	0.13	-0.00	0.77
Missing on Father's education (6 categories)	0.08	0.08	0.01	0.10
Father employed	0.88	0.87	-0.01	0.06
Father unemployed	0.03	0.03	0.00	0.40
Father outside labor market	0.09	0.10	0.01	0.05
Missing on Father's employment status (3 categories)	0.04	0.04	0.00	0.12
Father's total income (1000 DKK)	387.25	375.20	-12.05	0.13
Missing on Father's total income (1000 DKK)	0.03	0.04	0.00	0.09
Father's age in 2014 (years)	41.31	41.21	-0.10	0.20
Missing on Father's age in 2014 (years)	0.04	0.04	0.00	0.10
Father is teenager at date of birth	0.00	0.00	0.00	0.96

Notes: *p*-values based on standard errors clustered at the school level.

Estimation

The random assignment of invitations allows us to recover an unbiased estimate of the intent-to-treat effect of inviting schools to participate in READ by comparing the test scores among experimental groups (i.e., comparing the outcome between the "control group" and the "READ assignment group"). Consider the following equation:

$$y_{is} = \beta_0 + \delta INVITE_s + \boldsymbol{X}'_{is}\boldsymbol{\beta_1} + u_{is} \tag{1}$$

where y_{is} is the standardized test score for student i in school s, $INVITE_s$ is an indicator that equals one for schools assigned to the READ program, $\boldsymbol{X}_{is}$ is a vector of covariates. We use the double Lasso procedure to select the covariates in order to reduce researcher degrees of freedom. The double Lasso procedure uses machine learning to select variables that either predict the outcome, y_{is}, or the treatment indicator, $INVITE_s$ (Urminsky, Hansen and Chernozhukov, 2016).[10] u_{is} is a student-level error term. δ captures the intention-to-treat effect—that is, the average difference in test scores between schools assigned to READ and the control group.

Not all schools that were assigned to the READ group accepted the invitation (see Figure 2). Out of the 471 invited schools, 144 schools chose to participate (corresponding to 30.6 %). Moreover, seven of the 671 schools in the control group ended up receiving the READ program for various reasons.[11] To recover the effect of receiving the READ program, we use an instrumental variable (IV) approach in which we use the randomly assigned invitation to READ as an instrument for the school adopting READ. As invitations are randomly assigned, the instrument should be unrelated to unobserved outcome-relevant factors. Moreover, given that the invitation itself does not affect test scores, the instrument also satisfies the exclusion restriction. Under these assumptions, and given that there is a first stage, the IV approach allows us to estimate the local average treat-

[10] The selected variables can be seen in Table 3

[11] One reason for this non-compliance was that some schools are nested in administrative partnerships and, therefore, share an e-mail address. Thus, some invitations were forwarded to schools in the control group. If interested in participating, these schools were allowed to receive the READ program.

ment effect (LATE) of receiving the READ treatment as opposed to not receiving any treatment for those schools that complied with the assignment to treatment. To be specific, the first-stage equation can be written as follows:

$$READ_s = \alpha_0 + \lambda INVITE_s + \boldsymbol{X}'_{is}\boldsymbol{\alpha_1} + e_{is} \tag{2}$$

where $READ_s$ is a dummy for the school adopting the READ program. In a model without covariates, the LATE is the ratio of the reduced form estimate to the first-stage estimate ($\gamma = \frac{\delta}{\lambda}$). We use 2SLS to estimate the LATE effect. To take the nested structure of students in schools into account, we estimate all models with cluster-robust standard errors at the school level.

2.4 The Effect of READ on Student Achievement

Table 2 presents the main results on students' reading skills. Model 1 shows the intention-to-treat effects of assigning schools to the READ program on student test scores. Model 2 shows the model with covariates, $\boldsymbol{X}_{is}$ (see Equation 1). In both models, the effect is small in magnitude and not statistically distinguishable from zero.[12]

Model 3 presents first-stage estimates from Equation 2 and shows that the invitation increased the probability of participation by 28.6 percentage points compared to the non-invited group. Models 4 and 5 present two-stage least squares (2SLS) estimates. As the intention-to-treat estimates are close to zero, the 2SLS estimates are also rather small and insignificant. Table C.1 in the Appendix presents intention-to-treat estimates for the three subdomains "Language comprehension,"

[12] As a supplementary analysis, we apply a randomization inference procedure that resembles the assignment mechanism with the three waves. Specifically, we randomize the list of the 1,142 schools 1,000 times. Each time, we assign the first 471 schools to a hypothetical invitation group, and estimate the Intent-to-Treat effect using the Lasso procedure. This process generates a two-sided p-value of 0,60, indicating that in the absence of any effect there is a 60% chance of obtaining an estimate that is at least as large as our estimate presented in Table 2, Model 2. Moreover, Table C.2 presents a model in which we estimate the treatment effect separately for each of the three waves in which invitations were sent out. As expected, the standard errors are larger compared to our main model. However, the estimated treatment effects are very similar across the waves ranging from -0.014 to -0.013 standard deviations.

"Decoding," and "Text comprehension," separately. For all domains, the estimates are small and statistically insignificant.[13]

Table 2: Main Results on Students' Reading Skills

	(1) ITT	(2) ITT	(3) First stage	(4) LATE	(5) LATE
Invited	-0.014	0.011	0.286**		
	(0.024)	(0.019)	(0.024)		
Participating				-0.048	0.039
				(0.084)	(0.068)
Observations	51312	51030	51312	51312	51030
Schools (clusters)	1140	1130	1140	1140	1130
Mean control group	0.006	0.006	0.011	0.006	0.006
Adjusted R-squared	0.000	0.149	0.178	0.000	0.149
LASSO Covariates	No	Yes	No	No	Yes

Notes: Columns (1)-(3) are estimated with OLS. Columns (4)-(5) are estimated with 2SLS. Standard errors clustered at the school level in parentheses.

3 Scalability

To understand the small effects in the national trial compared to the original, local trial, we examine three types of scale-up challenges: selection into the trial (Heckman, 1992, 2020; Heckman and Smith, 1995), replicability (Gelman and Carlin, 2014), and implementation (Duflo, Glennerster and Kremer, 2008).

[13]The 2SLS estimates may be statistically insignificant because of lack of power. To complement the analysis, Table C.3 in the Appendix uses a DiD model to compare schools that adopted the program to schools that did not. The DiD estimates are more precise. However, they are also statistically insignificant and suggest that we can rule out effects larger than 0.11 standard deviations.

3.1 Selection into the Trial

We begin by examining which schools adopted the READ program in the national study and in the original, local study. The detailed Danish register data provide a unique opportunity to study characteristics of the schools that opted into the program in the two experiments. Table 3 compares the schools that decided to participate to the non-participants on observed characteristics of both the families and schools (using the variables selected by the Double Lasso procedure). Column 1 presents descriptive statistics on the participants in the national trial, whereas column 2 shows descriptive statistics on those who did not participate nationally (excluding the City of Aarhus, in which we ran a second, local trial—see section 3.2). Column 3 shows the difference between these two groups, and column 4 presents *p*-values for these differences. Although there are some differences between the READ participants and the non-participants, these are not large in magnitude.

Column 5 presents descriptive statistics on the schools that participated in the initial test of READ in the local government, and column 6 compares characteristics of these schools to those participating in the national trial. Column 7 tests for differences between the population in 2016 and participants in the original, local trial. Although many of these differences are statistically significant, again, most of them are small in absolute terms. However, there are some differences deserving attention. The initial READ trial had 21 % immigrants compared to 12 % among the participants in the national trial. Moreover, 26 % of the fathers had a vocational education in the original trial, compared to 43 % among the national participants.

Table 3: Differences in Mean between Participants and Non-Participants on Background Characteristics

	(1) National participating	(2) National not participating	(3) 2-1	(4) *p*	(5) READ 1.0	(6) 5-1	(7) *p*
Student level							
Child is a boy	0.53	0.52	-0.01	0.16	0.51	-0.02	0.16
Child's age	8.09	8.09	0.00	0.81	8.12	0.03	0.00
Child immigrant	0.12	0.11	-0.02	0.19	0.21	0.09	0.02
Child living with both parents	0.70	0.72	0.01	0.21	0.76	0.06	0.00
Child living with single parent	0.21	0.20	-0.01	0.45	0.20	-0.01	0.61
No. of children in family	2.37	2.31	-0.05	0.00	2.37	0.01	0.87
Mother compulsory education	0.15	0.14	-0.02	0.07	0.18	0.03	0.30
Mother upper seconday education	0.05	0.06	0.01	0.10	0.08	0.03	0.00
Mother short-cycle education	0.05	0.05	0.00	0.18	0.04	-0.01	0.09
Mother medium-cycle education	0.27	0.27	0.00	0.96	0.23	-0.04	0.02
Mother long-cycle education	0.11	0.13	0.02	0.09	0.20	0.09	0.00
Mother employed	0.77	0.79	0.02	0.05	0.69	-0.08	0.03
Mother unemployed	0.04	0.04	0.00	0.43	0.04	-0.00	0.58
Mother's total income (1000 DKK)	244.13	262.97	18.84	0.00	222.64	-21.49	0.16
Mother's age in	38.33	38.76	0.44	0.00	37.23	-1.10	0.00
Mother is teenager at date of birth	0.01	0.01	-0.00	0.11	0.02	0.00	0.66
Father compulsory education	0.17	0.16	-0.01	0.21	0.15	-0.02	0.21
Father upper seconday education	0.05	0.05	0.01	0.03	0.07	0.02	0.01
Father vocational education	0.43	0.40	-0.03	0.03	0.26	-0.17	0.00
Father medium-cycle education	0.13	0.14	0.01	0.10	0.17	0.04	0.00
Father long-cycle education	0.11	0.13	0.03	0.03	0.21	0.10	0.00
Father employed	0.87	0.88	0.01	0.31	0.76	-0.11	0.00
Father outside labor market	0.10	0.09	-0.01	0.19	0.15	0.05	0.01
Father's total income (1000 DKK)	357.79	385.91	28.12	0.00	329.67	-28.12	0.23
Father's age in	40.94	41.32	0.38	0.00	39.63	-1.31	0.00
School level							
School size	44.81	46.78	1.98	0.38	58.78	13.97	0.00
School average test score 2016	-0.09	-0.02	0.07	0.04	-0.09	0.00	0.99
Students	6766	49246	53128		1587	8353	
Schools	151	991	1142		27	178	

Notes: *p*-values based on standard errors clustered at the school level. Variables selected by the Double Lasso procedure.

These differences in participation could potentially explain differences in the results when the program is scaled up. However, differences in participation would only affect the estimated effects if there are also heterogeneous treatment effects for the groups that are over- or underrepresented. Based on previous evidence (Bergman and Chan, 2021), there are reasons to expect that parent-aimed interventions may have a particularly high potential for specific groups of students (e.g., those who are performing poorly). Table 4 examines the same subgroups that were studied in the first READ study, that is, parental education and ethnicity (Andersen and Nielsen, 2016). Thus, we split the sample by ethnic background and whether the mother has a college education. There is no evidence that the effects are different across these subgroups in the nationwide trial.

Table 4: Treatment Effects for Subgroups (OLS)

	(1) Mother low education	(2) Mother high education	(3) Immigrant background	(4) Danish background
Invited	0.008	0.008	-0.016	0.013
	(0.023)	(0.020)	(0.043)	(0.019)
P-value interaction	.891		.559	
Observations	27635	20460	5104	45926
Schools (clusters)	1130	1119	865	1130
Mean control group	-0.164	0.301	-0.500	0.059
Adjusted R-squared	0.087	0.074	0.111	0.125
LASSO Covariates	Yes	Yes	Yes	Yes

Notes: *p*-values are based on the null hypothesis that the point estimates are the same for the two subsamples. Standard errors clustered at the school level in parentheses.

To further examine possible heterogeneous treatment effects, we use the Causal Forest analysis by Wager and Athey (2018). Results are presented in Appendix C.2. The omnibus test (proposed by Chernozhukov et al., 2018) generated large p-values, which suggests either that the forest does not capture heterogeneity well or that there is not much heterogeneity. Therefore, also the causal forest analysis does not support the notion that heterogeneous effects are causing the difference in results between the original trial and the large-scale implementation.

Given that there are no indications of effects for any subgroups in the nationwide trial, the scale-up challenge is unlikely a result of different types of students in the local and the national samples.

3.2 Replicability: An Observational Study

During the same school year as the nationwide implementation, the local government in the City of Aarhus implemented the READ program again among a new cohort of 2nd-grade students.[14] Putting the program into operation, a subgroup of schools chose to participate in the program. The participating schools received the same material as the participating schools in the nationwide implementation. We use this observational replication to examine whether effects replicate when implemented at the same scale and in the same environment as in the original trial, but put into operation by the local government.

Design and Estimation

Participating schools in Aarhus received the READ intervention in January 2017. As schools were not randomly assigned to the program, systematic differences may be seen between local READ schools and non-READ schools that could explain differences in outcomes even in the absence of the READ program. To identify the impact of the READ program in the local government, we apply a DiD design in which we compare trends in student learning before and after the implementation of the READ program in the local government schools to two control groups. First, we compare the local treatment schools to a national control group—that is, schools in the other municipalities that were not invited and did not participate in the national READ program. Second, as a local control group, we compare the local treatment schools to the remaining schools in the local government that did not participate in READ.

To be specific, we estimate the following DiD model:

$$y_{ist} = \delta_1 Post_t + \delta_2 (READ_s \times Post_t) + \phi_s + v_{ist},$$

where $Post_t$ is a dummy indicating the school year after the implementation of READ. The

[14] Whereas the logo and design of the materials differed, the content of the program was similar to the one implemented nationwide.

interaction term $READ_s \times Post_t$ indicates the control group (i.e., national or local control schools) in the year after READ implementation. ϕ_s is a school fixed effect, and v_{ist} is an individual specific error term. Under the assumption of common trends in the absence of treatment, the coefficients δ_2 captures the effect of being assigned to the READ intervention on the reading outcome, y_{ist}.

As the national reading tests were changed in 2015—and test scores therefore not comparable before/after 2015—we include data as of 2015. To not confound the analysis by changes in school composition, we use the balanced panel of schools for which we have test score data and consistent school identifiers for the years 2015, 2016, and 2017. To enable comparisons in effect size to the nationwide experiment, we standardize the test scores based on the national mean and standard deviation of the 2016 population.

Results

Table 5 presents results of the DiD estimation. The upper panel provides results when we compare the local treatment schools to the remaining schools in the local government that did not participate in READ ("Local control group"). The lower panel provides results where the control group consist of schools in other municipalities than Aarhus that were not invited and did not participate in the national READ program ("National control group"). Model 1 suggests that the program increased student test scores by 0.18 standard deviations compared to the local comparison group and 0.14 compared to the national control group (both statistically significant at a 10-% level). Although the difference in pre-trend slopes between the the treatment group and the national controls question the validity of the common-trend assumption for this specification (see Appendix Figure C.2), the similarity in the two DiD estimates across the two control groups is reassuring. Models 2-4 show that the coefficients are positive and substantially large in magnitude across all subdomains—but largest for "Text comprehension."[15] Although slightly smaller in magnitude, the pattern in the effect estimates across subdomains is rather similar to the findings in the original, local trial, in

[15]Table C.5-C.8 present robustness results and show that the estimates are rather similar across specifications with and without school fixed effects and with student and school covariates.

which effect sizes were estimated to be .19 for "Language Comprehension," .23 for "Decoding," and .27 for "Text comprehension" (see Andersen and Nielsen, 2016, Table 1). Table C.9 in the Appendix presents a placebo test that uses 2016 as the treatment year. All estimates from the placebo specification are statistically insignificant, and seven out of eight are smaller in magnitude than in the main specification.

Even though the identification of the causal effect is not as credible in the DiD design as in a randomized controlled trial—and despite the small sample size of the "local" treatment group—the fact that a second study finds effect sizes across subdomains that are consistent with the first study provides some evidence of the effectiveness of the program. Moreover, it is reassuring that the placebo tests show that the program does not have an impact on outcomes for non-treated cohorts. Thus, the positive effects in the observational study suggest that problems due to statistical inference may not be the main reason that the READ program was not effective at improving student learning in the nationwide trial.

Table 5: Difference-in-Difference Estimates on Students' Reading Skills. Total Score and the Three Subdomains (OLS)

LOCAL CONTROL GROUP				
	(1)	(2)	(3)	(4)
	Total score	Language comprehension	Decoding	Text comprehension
Post treatment	-0.043	-0.031	-0.041	-0.040
	(0.059)	(0.060)	(0.050)	(0.056)
READ X Post treatment	0.176^{+}	0.123	0.138	0.201*
	(0.092)	(0.081)	(0.089)	(0.086)
Constant	0.039*	0.047**	0.039*	0.016
	(0.015)	(0.014)	(0.015)	(0.014)
Observations	8284	8284	8284	8284
Schools (clusters)	46	46	46	46
Adjusted R-squared	0.002	0.001	0.001	0.003
School Fixed Effects	Yes	Yes	Yes	Yes
NATIONAL CONTROL GROUP				
	(1)	(2)	(3)	(4)
	Total score	Language comprehension	Decoding	Text comprehension
Post treatment	-0.006	-0.004	-0.008	-0.003
	(0.013)	(0.012)	(0.013)	(0.013)
READ X Post treatment	0.139^{+}	0.096^{+}	0.104	0.163*
	(0.071)	(0.056)	(0.074)	(0.066)
Constant	0.011*	0.011**	0.010*	0.006
	(0.004)	(0.004)	(0.004)	(0.004)
Observations	92862	92862	92862	92862
Schools (clusters)	676	676	676	676
Adjusted R-squared	0.000	0.000	0.000	0.000
School Fixed Effects	Yes	Yes	Yes	Yes

Notes: Standard errors clustered at the school level in parentheses.

3.3 Take-Up and Implementation Support

Our results rule out small effects of the intention of the national government to implement the program around the country. One explanation for not finding an average effect in the national, large-scale experiment of similar magnitude as in the original, local study could be low levels of take-up and poor program implementation. An effective intervention is more than just the core program. It also includes the delivery of the program through several steps of the implementation process. Thus, it is crucial to think about adoption at each step of the process, from school principals to individual families. As shown in Section 2.3, about one-third of the schools accepted the invitation. However, even if school principals accepted the program, teachers had to convey the message and program content to the families in order to reach the children. Thus, low levels of take-up at the school level and among parents at schools that adopted the program could both explain the lack of an overall effect.

Take-Up among Parents in the National and Local Settings

To analyze the extent to which the program was adopted among families within schools that participated, we use data generated by The READ smartphone app.[16] To compare the level of take-up in the local replication study to the national government study, Figure 4 presents data on the app use. Figure 4a (left panel) shows the proportion of pending users that had not signed up in the app over time. About 48 % of the children's families that were assigned to the program in the local government of Aarhus registered in the app, whereas only about 24 % registered among the families in the national study. Figure 4b (right panel) compares the distribution of the proportion of users across schools in the two studies. The distribution is moved to the right in the Aarhus study. Moreover, a relatively large proportion of participating schools in the nationwide setting

[16] Out of the 151 READ schools that chose to participate in the national trial, 109 schools provided access to student identifiers that enable us to match the app data at the student level with the administrative records provided by Statistics Denmark. In contrast, we were able to match app data for all students at all participating schools in the local replication. Given that the schools in the national trial that did not provide access were less likely to implement the program subsequently, a higher take-up in the local replication than in the national trial would be a lower bound.

have no students signing up. The large differences in take-up of the app between schools in the local setting and the national setting—and the fact that no parents signed up in the app in about 20% of the schools in the national setting—-could suggest that some schools did not invest the resources necessary to implement the program sufficiently.

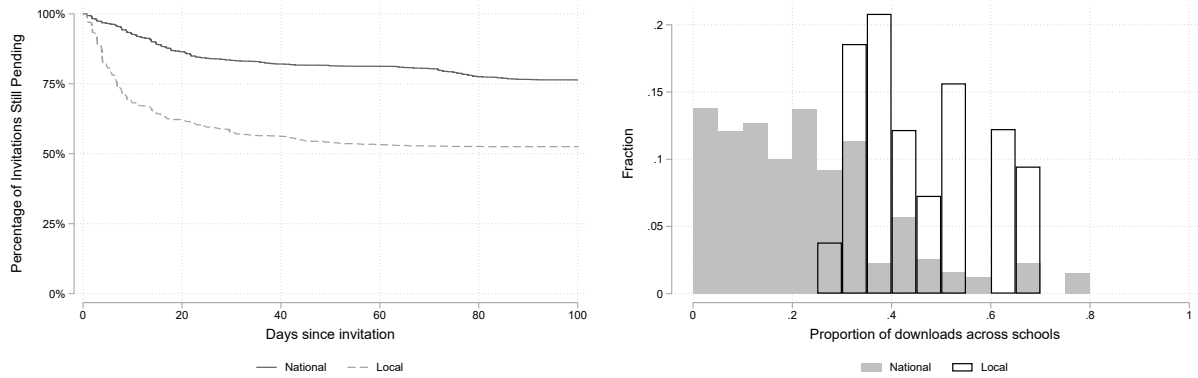
One explanation for the differences in take up between the local and national contexts could be differences in school capacity. The local government in Aarhus is not an outlier in terms of school funding, though. Municipality-level data show that the annual budget per student in Aarhus schools was DKK 80,615 in 2016. If we look at the national distribution, the median is DKK 74,106, and the 75th percentile is DKK 80,456. The average class size in Aarhus was 22.1 (median among all municipalities was 21.5), and the average school size in Aarhus was 607 compared to the median of 426, and the 75th percentile was 545. Thus, although Aarhus schools are larger on average and relatively well funded compared to other municipalities, the local government is not markedly different from the rest of the country.

Take-Up and Child Outcomes

To understand the importance of take-up for the effectiveness of the program, we study the extent to which take-up explains differences in reading test scores between treatment and control group students in the national trial.

Figure 5 reports differences in test scores conditional on the proportion of parents at the school who have downloaded the app. The graph plots the marginal effect based on a linear model and a histogram of the proportion of adopters at the school. To relax the linear functional form assumption, we also present results from a binning estimator that splits schools into tertiles based on percentage of parents who downloaded the app and estimates the difference in test scores (compared to control group students) within each bin separately.

Even though these analyses are exploratory, the linear model suggests that the effect of the READ program increases with take-up. The binning estimator shows a significant positive effect among the top tertile, whereas the effect is not significantly different from zero for the remaining



(a) Pending downloads across families (b) Downloads across schools

Figure 4: Take up: App users.

Notes: Panel (a) plots survival curves that show the distribution of families that downloaded the app in the national sample and the Aarhus sample. Each survival curve plots the percentage of app user names still pending versus the number of days elapsed since the families received the invitation to download the app. The Aarhus sample consists of all families at schools that participated in READ. Among these families, 47.6 percent downloaded the READ app within 100 days. The national sample consists of all families at schools that accepted to receive the READ program. Panel (b) shows the distribution of the proportion of downloads at the school level across setting.

groups. This finding is robust to several specification checks. First, the positive result for the top tertile is not sensitive to whether we include the full set of school or student covariates (see Appendix Tabel C.11).[17] Second, the statistically significant finding for top adopters holds whether we divide bins by the median (Appendix Tabel C.12, Model 1) or by quartiles (Appendix Tabel C.12, Model 2). Third, the estimates are very similar irrespective of whether we include the non-compliers among the non-invited schools (Appendix Tabel C.12, Model 3). Moreover, the three groups (as measured by the tertiles) are rather balanced on covariates with little evidence of the level of take up being systematically related to school and student characteristics (see Appendix Tabel C.10). Importantly, prior school performance does not explain the degree to which the program is adopted at the school level.

[17] Whereas the result for the top tertile is stable across specifications, the size of the coefficient for the lower tertile changes somewhat in magnitude across specification although the sign does not change.

Although schools may differ systematically on outcome-relevant factors that are unobserved, these findings suggest that the effect of the READ program is heterogeneous with respect to number of downloads of the app at the school level. Moreover, these supplementary analyses suggest that the program is effective once take-up is sufficiently high within schools.

Implementation Support

Given the evidence on the positive impact once the program was sufficiently adopted at the school, one key question is why we observe so profound differences in parents' take-up between the local and national setting—and why so many schools failed to implement the program in the national setting.

While our experiments do not answer these questions directly, we can use information on how the program was run in the national trial and in the local replication study. Interestingly, there were noticeable differences in the implementation procedures between the local program and the national program. First, more resources were allocated to the implementation of the program in the local setting than in the national setting. Specifically, in Aarhus, a team of internal consultants worked on the implementation of the intervention, whereas there was little administrative support at the national level. Moreover, municipalities were merely informed about the program in the national setting and were not involved in the implementation. If governments do not scale up the number of people working on implementing the intervention proportional to the number of clients, implementation support is spread out more thinly, which may lead to lower take-up and less perseverance at the front line.

Second, the people implementing the programs may vary in their skills and the effort that they invest. Even if implementation costs are held fixed per client served, scarcity in the supply of highly skilled and devoted professionals is likely to occur as programs scale (Davis et al., 2017). For example, the Aarhus team held information meetings with the schools and communicated continuously with the schools about the program, whereas the national implementations merely included information sent out regarding the program in the beginning. These differences in the

way the program was run locally and nationally may provide one explanation for the profound differences that we observe in take-up, which could drive the differences in the overall effectiveness of the program that we observe in the two studies. Unfortunately, we do not have information about the time spent by consultants implementing the program at the local and the national level. Nevertheless, the findings could suggest that scaling of the implementation support is particularly important to ensure take-up and implementation at the school level. Certainly, we see the importance of implementation support as a key area for future research.

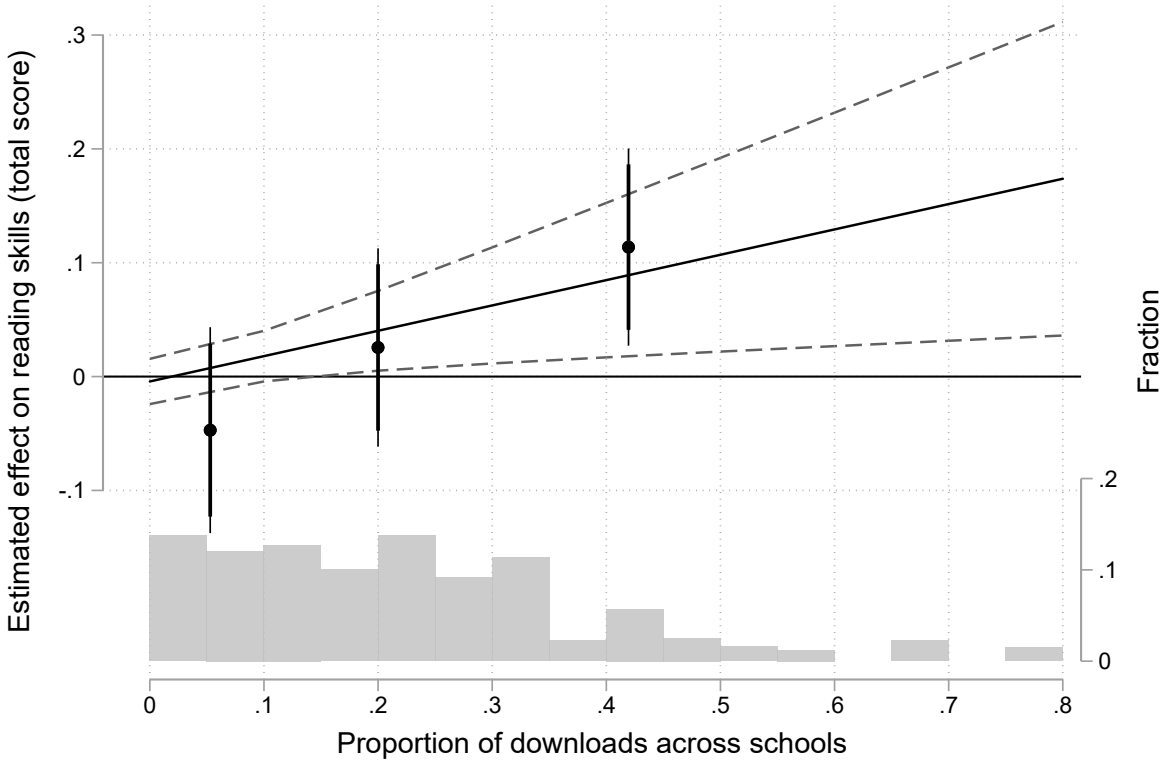


Figure 5: Estimated Differences in Test Scores Between Treatment and Control Group Students Conditional on Downloads

Notes: Dots and vertical lines show coefficients and 90 % and 95 % confidence intervals based three levels (tertiles) of downloads among invited schools that received the intervention. The black line and dotted lines show the estimated difference in test scores and 95 % confidence intervals conditional on a continuous measure of proportion of downloads across schools excluding non-compliers from the control group. The bars at the bottom show the fraction of families in schools with different proportions of downloads (similar to Figure 4b).

4 Conclusion

Various explanations for lack of scalability of experimental findings have been discussed in the literature. To study the scalability of parent-aimed interventions, we fielded an experiment as part of the scale-up of a reading program that had shown promising results in a local setting. The findings from the large-scale, national study suggest that the nationwide implementation of the program was not effective at improving student learning to the same extent as in the original, local setting. The intention-to-treat coefficients in the national setting were small in magnitude and not statistically significant.

Schools participating in the local READ program in the original, local trial differed from the participating schools in the national trial on some parameters, such as the share of immigrants. Yet, we did not find much evidence of heterogeneous effects across subgroups in the national trial, which suggests that representativeness of the original sample was not the main cause of the unsuccessful scaling. Moreover, the observational replication study run by the same local government as the original study found effects that were rather similar to those of the original study. The replication of effects from the original, local trial suggests that the challenge of scaling may not be driven by problems of replicability.

Behavioral data on how many families signed into the READ app provides one objective measure of differences in the level of take-up among parents at participating schools. Even though parents may have used the READ program without using the app, comparing data on app use between the local and the national studies gives an indication of differences in take-up at the family level. These data show that a much larger proportion of parents signed into the app in the local government program than in the national program. Moreover, at the national level, students at schools with relatively high levels of take-up among parents obtained relatively higher reading test scores—a difference that persists even after we control for a large set of covariates at both the family and school level.

Supplementary information on the process suggests that the effort to implement the program was substantially higher when the program was run on a small, local scale with very devoted

personnel than at a nationwide scale with less resources allocated to the implementation. These findings are a strong indication that the low levels of take-up among both schools and parents in the national trial explains why the effects from the smaller-scale trial did not replicate at scale. Moreover, we have indications that the way that the local government ran the program was not representative of the way the national government implemented it. Thus, the results of the analyses and the qualitative insights may suggest that implementation is crucial for scalability, and that implementation is not done by securing high take-up at the school level. Engagement of families is also a crucial part.

The potential scalability of programs constitutes an important criteria in judging their importance to policy and practice. Our findings suggest that home-based educational programs may be hard to scale because of low levels of take-up even though the behavioral change that is required to implement these programs should be relatively low compared to other educational interventions. Ultimately, the success of home-based programs depends on the motivation and capacity of school leaders, teachers, and parents to implement them.

One important question for policy is how to improve take-up rates at each stage of the implementation process. Previous research provides some input on how to improve take-up rates. For example, offering schools modest compensation for participating could be an effective tool to increase take-up rates and support (Andersen and Hvidman, 2021). While school support is important for take-up rates, parental behaviors and beliefs appears to be equally important. Fortunately, behavioral interventions at the family level have proven to have the potential to boost families' engagement (Mayer et al., 2018). Future research should study such implementation factors systematically, since they seem to be crucial for gaining the potential benefits of scaling parent-aimed education programs.

References

Allcott, Hunt. 2015. "Site Selection Bias in Program Evaluation." *The Quarterly Journal of Economics*, 130(3): 1117–1165.

Al-Ubaydli, Omar, John A. List, and Dana L. Suskind. 2017. "What Can We Learn from Experiments? Understanding the Threats to the Scalability of Experimental Results." *American Economic Review*, 107(5): 282–286.

Andersen, Simon Calmar, and Helena Skyt Nielsen. 2016. "Reading Intervention with a Growth Mindset Approach Improves Children's Skills." *Proceedings of the National Academy of Sciences*, 113(43): 12111–12113.

Andersen, Simon Calmar, and Ulrik Hvidman. 2021. "Can Reminders and Incentives Improve Implementation Within Government? Evidence from a Field Experiment." *Journal of Public Administration Research and Theory*, 31: 234–249.

Angrist, Joshua David, and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton:Princeton University Press.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, and Michael Walton. 2017. "From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application." *Journal of Economic Perspectives*, 31(4): 73–102.

Bergman, Peter. 2021. "Parent-Child Information Frictions and Human Capital Investment: Evidence from a Field Experiment." *Journal of Political Economy*, 129: 286–322.

Bergman, Peter, and Eric W. Chan. 2021. "Leveraging Parents: The Impact of High-Frequency Information on Student Achievement." *Journal of Human Resources*, 56: 125–158.

Bergman, Peter, and Todd Rogers. 2017. "The Impact of Defaults on Technology Adoption, and Its Underappreciation by Policymakers." CESifo Working Paper Series Working Paper 6721.

Berkowitz, Talia, Marjorie W. Schaeffer, Erin A. Maloney, Lori Peterson, Courtney Gregor, Susan C. Levine, and Sian L. Beilock. 2015. "Math at Home Adds up to Achievement in School." *Science*, 350(6257): 196–198.

Beuchert, Louise Voldby, and Anne Brink Nandrup. 2018. "The Danish National Tests at a Glance." *Nationaloekonomisk Tidsskrift*, 1: 1–37.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur. 2018. "Experimental Evidence on Scaling up Education Reforms in Kenya." *Journal of Public Economics*, 168: 1–20.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val. 2018. "Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, with an Application to Immunization in India."

Czibor, Eszter, David Jimenez-Gomez, and John A. List. 2019. "The Dozen Things Experimental Economists Should Do (More of)." *Southern Economic Journal*, 86(2): 371–432.

Davis, Jonathan M.V., Jonathan Guryan, Kelly Hallberg, and Jens Ludwig. 2017. "The Economics of Scale-Up." National Bureau of Economic Research Working Paper 23925.

DellaVigna, Stefano, and Elizabeth Linos. 2022. "RCTs to Scale: Comprehensive Evidence From Two Nudge Units." *Econometrica*, 90(1): 81–116.

Doss, Christopher, Erin M. Fahle, Susanna Loeb, and Benjamin N. York. 2019. "More Than Just a Nudge: Supporting Kindergarten Parents with Differentiated and Personalized Text Messages." *Journal of Human Resources*, 54(3): 567–603.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2008. "Using Randomization in Development Economics Research: A Toolkit." *T. Schultz and John Strauss, eds., Handbook of Development Economics* Vol. 4. Amsterdam and New York:North Holland.

Dweck, Carol S. 1999. *Self-Theories: Their Role in Motivation, Personality, and Development.* Psychology Press.

Dweck, Carol S. 2006. *Mindset: The New Psychology of Success.* . Reprint edition ed., New York:Random House.

Ganimian, Alejandro J. 2020. "Growth-Mindset Interventions at Scale: Experimental Evidence From Argentina." *Educational Evaluation and Policy Analysis*, 42(3): 417–438.

Gelman, Andrew, and John Carlin. 2014. "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors." *Perspectives on Psychological Science*, 9(6): 641–651. PMID: 26186114.

Guryan, Jonathan, Erik Hurst, and Melissa Kearney. 2008. "Parental Education and Parental Time with Children." *Journal of Economic Perspectives*, 22(3): 23–46.

Haimovitz, Kyla, and Carol S. Dweck. 2016. "Parents' Views of Failure Predict Children's Fixed and Growth Intelligence Mind-Sets." *Psychological Science*, 27(6): 859–869.

Harrison, Glenn W., and John A. List. 2004. "Field Experiments." *Journal of Economic Literature*, 42(4): 1009–1055.

Heckman, James J. 1992. "Randomization and Social Policy Evaluation." In *Evaluating Welfare Training Programs*. , ed. C.F. Manski and I. Garfinkel, 201–230. Cambridge, MA:Harvard University Press.

Heckman, James J. 2020. "Randomization and Social Policy Evaluation Revisited." Human Capital and Economic Opportunity Global Working Group Working Paper 2020-001.

Heckman, James J., and Jeffrey A. Smith. 1995. "Assessing the Case for Social Experiments." *Journal of Economic Perspectives*, 9(2): 85–110.

Jakobsen, Morten, and Simon Calmar Andersen. 2013. "Coproduction and Equity in Public Service Delivery." *Public Administration Review*, 73(5): 704–713.

Kizilcec, René F., Justin Reich, Michael Yeomans, Christoph Dann, Emma Brunskill, Glenn Lopez, Selen Turkay, Joseph Jay Williams, and Dustin Tingley. 2020. "Scaling up behavioral science interventions in online education." *Proceedings of the National Academy of Sciences*, 117(26): 14900–14905.

Koch, Alexander, Julia Nafziger, and Helena Skyt Nielsen. 2015. "Behavioral Economics of Education." *Journal of Economic Behavior & Organization*, 115: 3–17.

Kraft, Matthew A. 2020. "Interpreting Effect Sizes of Education Interventions." *Educational Researcher*, 49(4): 241–253.

Lavecchia, A. M., H. Liu, and P. Oreopoulos. 2016. "Chapter 1 - Behavioral Economics of Education: Progress and Possibilities." In *Handbook of the Economics of Education*. Vol. 5, , ed. Eric A. Hanushek, Stephen Machin and Ludger Woessmann, 1–74. Elsevier.

Lortie-Forgues, Hugues, and Matthew Inglis. 2019. "Rigorous Large-Scale Educational RCTs are Often Uninformative : Should We Be Concerned?" *Educational Researcher*.

Mayer, Susan E., Ariel Kalil, Philip Oreopoulos, and Sebastian Gallegos. 2018. "Using Behavioral Insights to Increase Parental Engagement: The Parents and Children Together Intervention." *Journal of Human Resources*, 0617–8835R.

Moorman, Elizabeth A., and Eva M. Pomerantz. 2010. "Ability Mindsets Influence the Quality of Mothers' Involvement in Children's Learning: An Experimental Investigation." *Developmental Psychology*, 46(5): 1354–1362.

Mullainathan, Sendhil, and Eldar Shafir. 2013. *Scarcity: Why Having Too Little Means So Much.* New York:Times books.

Noble, Claire, Giovanni Sala, Michelle Peter, Jamie Lingwood, Caroline Rowland, Fernand Gobet, and Julian Pine. 2019. "The impact of shared book reading on children's language skills: A meta-analysis." *Educational Research Review*, 28: 100290.

Pomerantz, Eva M., Elizabeth A. Moorman, and Scott D. Litwack. 2007. "The How, Whom, and Why of Parents' Involvement in Children's Academic Lives: More Is Not Always Better." *Review of Educational Research*, 77(3): 373–410.

Rogers, Todd, and Avi Feller. 2018. "Reducing Student Absences at Scale by Targeting Parents' Misbeliefs." *Nature Human Behaviour*, 2(5): 335.

Urminsky, Oleg, Christian Hansen, and Victor Chernozhukov. 2016. "Using Double-Lasso Regression for Principled Variable Selection." *SSRN Scholarly Paper*.

Vivalt, Eva. 2020. "How Much Can We Generalize from Impact Evaluations?" *Journal of the European Economic Association*, 18: 3045–3089.

Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests." *Journal of the American Statistical Association*, 113(523): 1228–1242.

York, Benjamin N., Susanna Loeb, and Christopher Doss. 2019. "One Step at a Time: The Effects of an Early Literacy Text-Messaging Program for Parents of Preschoolers." *Journal of Human Resources*, 54(3): 537–566.

Appendix

A Meta-analysis of Educational Interventions

The meta-analysis is based on data from two studies. First, studies of educational programs commissioned by EEF and NCEE are based on data from Lortie-Forgues and Inglis (2019). These are categorized as educational interventions. Second, studies specifically focusing on parent-aimed interventions are based on data from a systematic review of shared book reading by Noble et al. (2019). Following Noble et al. (2019), we exclude studies with effect sizes greater than three standard deviations, and we include only randomized controlled trials with standardized test outcomes in the meta-analysis.

We supplement the data on parent-aimed interventions with a systematic literature search that broadens the search criteria. The systematic search was based on the same search string as Noble et al. (2019) with the following addition: "caregiver*", "parent*", or "home" combined with "reading", "training", "education", "information", "implement*", "intervention", "achievement", "engagement", "text messag*", or "provid* knowledge". The systematic search was limited to studies with the words "random*", "causal", "experiment", or "impact" in the following journals: *Science, Proceedings of the National Academy of Sciences, Nature Human Behaviour, Journal of Economic Perspectives, Economics Letters, Journal of Human Resources, Economics of Education Review, Journal of Policy Analysis and Management, Journal of Public Administration Research and Theory, Public Administration Review, Annals of The American Academy of Political and Social Science, and Early Child Development and Care*.

A separate supplementary materials document lists all the data and references included in our meta-analysis. We will make this available as an online appendix upon publication.

B Baseline and balance

B.1 Baseline balance using full set of variables

Table B.1: Differences in mean between invited and non-invited on background characteristics. Full set of variables

	(1) Non-invited	(2) Invited	(3) 1-2	*p*-values
Child is a boy	0.52	0.52	-0.00	0.68
Child's age in 2016	8.09	8.09	0.00	0.41
Child immigrant	0.10	0.12	0.01	0.07
Child living with both parents	0.72	0.71	-0.01	0.25
Child living with single parent	0.20	0.21	0.01	0.10
Child living with parent in new relationship or not living with own parents	0.08	0.08	-0.00	0.49
No. of children in family	2.31	2.33	0.02	0.17
Mother compulsory education	0.14	0.14	0.01	0.16
Mother upper seconday education	0.05	0.05	-0.00	0.58
Mother vocational education	0.31	0.31	-0.00	0.94
Mother short-cycle education	0.05	0.05	-0.00	0.09
Mother medium-cycle education	0.27	0.27	0.00	0.97
Mother long-cycle education	0.13	0.13	-0.00	0.66
Mother employed	0.79	0.78	-0.01	0.06
Mother unemployed	0.04	0.04	0.00	0.25
Mother outside labor market	0.17	0.18	0.01	0.07
Mother's total income (1000 DKK)	263.12	256.89	-6.23	0.19
Mother's age in 2014 (years)	38.76	38.63	-0.14	0.11
Mother is teenager at date of birth	0.01	0.01	0.00	0.67
Father compulsory education	0.16	0.17	0.01	0.16
Father upper seconday education	0.06	0.05	-0.00	0.08
Father vocational education	0.41	0.41	0.00	0.97
Father short-cycle education	0.08	0.08	-0.00	0.70
Father medium-cycle education	0.14	0.13	-0.00	0.45
Father long-cycle education	0.13	0.13	-0.00	0.77
Father employed	0.88	0.87	-0.01	0.06
Father unemployed	0.03	0.03	0.00	0.40
Father outside labor market	0.09	0.10	0.01	0.05
Father's total income (1000 DKK)	387.25	375.20	-12.05	0.13
Father's age in 2014 (years)	41.31	41.21	-0.10	0.20
Father is teenager at date of birth	0.00	0.00	0.00	0.96
School size	61.11	60.49	-0.62	0.75
School average test score 2016	0.01	-0.02	-0.03	0.18

Notes: *p*-values based on standard errors clustered at the school level.

C Supplementary results

C.1 Additional results on the nationwide experiment

Table C.1: Effects on the three sub domains (Intention-to-treat)

	(1) Language comprehension	(2) Decoding	(3) Text comprehension
Invited	-0.003	0.010	0.004
	(0.017)	(0.020)	(0.019)
Constant	0.913**	1.783**	1.072**
	(0.148)	(0.147)	(0.137)
Observations	51030	51030	51030
Schools (clusters)	1130	1130	1130
Adjusted R-squared	0.123	0.117	0.117
Wave Indicators	Yes	Yes	Yes
LASSO Covariates	Yes	Yes	Yes

Notes: Models estimated with OLS. Standard errors clustered at the school level in parentheses.

Table C.2: Effects on the three recruitment waves (Intention-to-treat)

	(1) Danish reading - Total score
Wave=1	-0.014
	(0.031)
Wave=2	-0.013
	(0.034)
Wave=3	-0.014
	(0.041)
Observations	51312
Schools (clusters)	1140
Mean control group	0.006
Adjusted R-squared	-0.000

Notes: Models estimated with OLS. Standard errors clustered at the school level in parentheses.

Table C.3: Participants vs non-participants (excluding Aarhus). Difference-in-difference estimates on students' reading skills. Total score (OLS)

	(1) Total score	(2) Total score	(3) Total score
Post treatment	-0.009	-0.009	-0.012
	(0.011)	(0.011)	(0.010)
Participating X Post treatment	0.036	0.036	0.050
	(0.032)	(0.032)	(0.031)
Constant	-0.022*	0.003	1.734**
	(0.011)	(0.003)	(0.089)
Observations	150685	150685	150685
Schools (clusters)	1123	1123	1123
Adjusted R-squared		0.000	0.115
Fixed Effects	No	Yes	Yes
LASSO Covariates	No	No	Yes

Notes: Standard errors clustered at the school level in parentheses. Different specifications reported in Appendix C.

C.2 Causal Forest Analysis of heterogeneous effects

To further examine possible heterogeneous treatment effects we use the Causal Forest analysis by Wager and Athey (2018). Figure C.1 shows the Out-of-bag conditional average treatment effect (CATE). A few students have large either positive or negative predicted effects. However, when we use the omnibus test proposed by Chernozhukov et al. (2018) (see Table C.4), we obtain large p-values, which suggest either that the forest does not capture heterogeneity well, or that there is not much heterogeneity.

Figure C.1: Causal forests: out-of-bag CATE

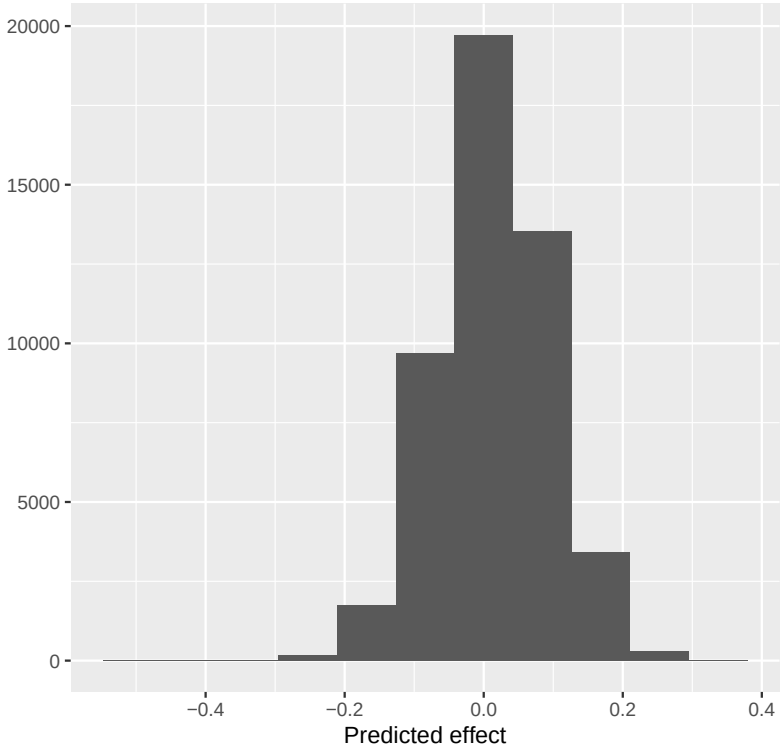


Table C.4: Omnibus test

	Estimate	Pr(>t)
Mean Forest Prediction	0.827	.286
	(1.463)	
Differential Forest Prediction	−0.373	. 981
	(0.180)	

Note: The table shows estimates for the omnibus test inspired by Chernozhukov et al. (2018) and implemented through the test_calibration function from the grf library in R. (Standard errors in parentheses)

C.3 Difference-in-Differences analyses of local replication experiment

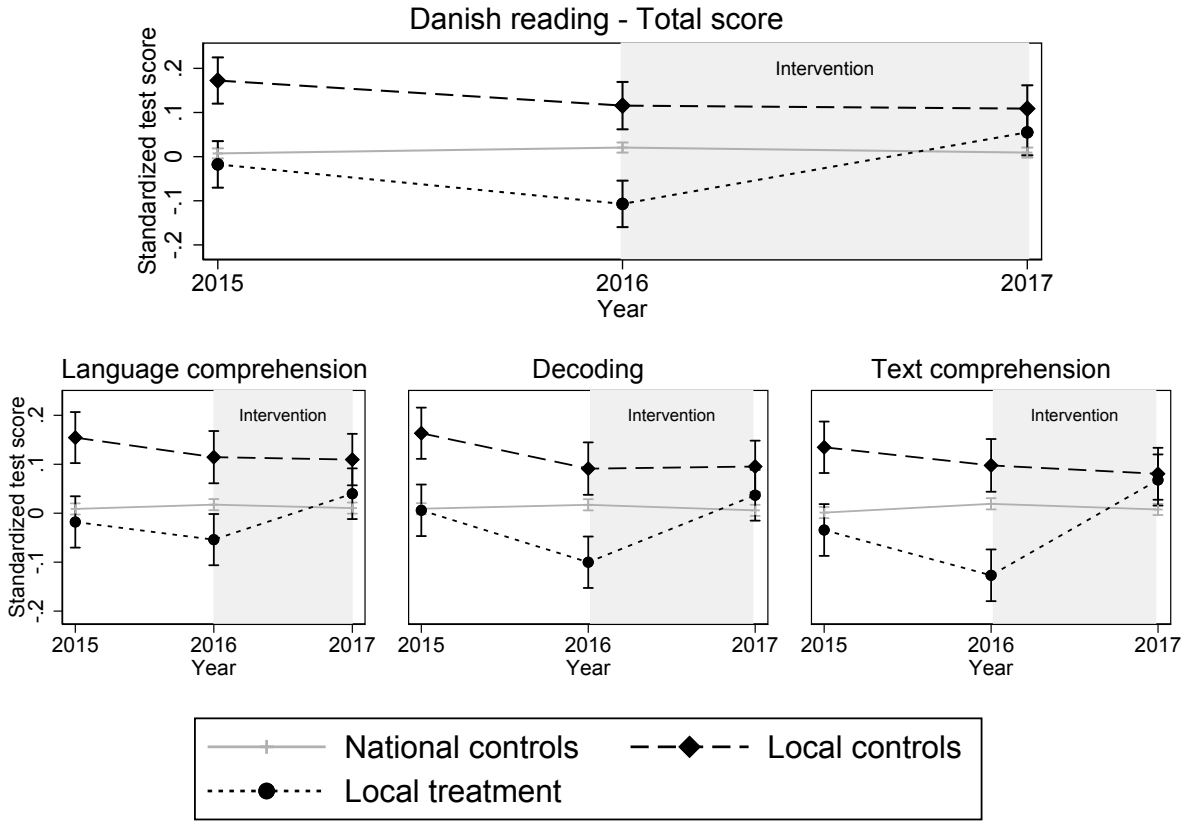


Figure C.2: Effect of READ in observational replication study on total, composite reading test score and three subscales

Table C.5: Total score

	(1)	(2)	(3)
Not READ	0.076		
	(0.055)		
Not READ Aarhus	0.207		
	(0.073)		
Post READ	0.117	0.133	0.103
	(0.073)	(0.070)	(0.071)
Not READ X Post READ	-0.122	-0.139	-0.113
	(0.074)	(0.071)	(0.072)
Not READ Aarhus X Post READ	-0.153	-0.176	-0.153
	(0.095)	(0.091)	(0.091)
Mean of control	-0.062	-0.062	-0.062
Observations	96965	96965	96965
Schools (clusters)	699	699	699
Adjusted R-squared	0.001	0.000	0.116
Fixed Effects	No	Yes	Yes
Covariates	No	No	Yes

Notes: Standard errors clustered at the school level in parentheses. The full list of the included covariates is reported in Table 3.

Table C.6: Language comprehension

	(1)	(2)	(3)
Not READ	0.049		
	(0.048)		
Not READ Aarhus	0.171		
	(0.064)		
Post READ	0.076	0.092	0.068
	(0.057)	(0.054)	(0.053)
Not READ X Post READ	-0.078	-0.096	-0.071
	(0.058)	(0.056)	(0.054)
Not READ Aarhus X Post READ	-0.101	-0.123	-0.102
	(0.083)	(0.081)	(0.079)
Mean of control	-0.036	-0.036	-0.036
Observations	96965	96965	96965
Schools (clusters)	699	699	699
Adjusted R-squared	0.001	0.000	0.092
Fixed Effects	No	Yes	Yes
Covariates	No	No	Yes

Notes: Standard errors clustered at the school level in parentheses. The full list of the included covariates is reported in Table 3.

Table C.7: Decoding

	(1)	(2)	(3)
Not READ	0.060		
	(0.053)		
Not READ Aarhus	0.175		
	(0.072)		
Post READ	0.084	0.097	0.070
	(0.076)	(0.073)	(0.074)
Not READ X Post READ	-0.091	-0.104	-0.085
	(0.077)	(0.074)	(0.075)
Not READ Aarhus X Post READ	-0.117	-0.138	-0.119
	(0.092)	(0.088)	(0.088)
Mean of control	-0.047	-0.047	-0.047
Observations	96965	96965	96965
Schools (clusters)	699	699	699
Adjusted R-squared	0.001	0.000	0.092
Fixed Effects	No	Yes	Yes
Covariates	No	No	Yes

Notes: Standard errors clustered at the school level in parentheses. The full list of the included covariates is reported in Table 3.

Table C.8: Text comprehension

	(1)	(2)	(3)
Not READ	0.091		
	(0.050)		
Not READ Aarhus	0.197		
	(0.064)		
Post READ	0.149	0.161	0.133
	(0.067)	(0.064)	(0.067)
Not READ X Post READ	-0.151	-0.163	-0.142
	(0.068)	(0.066)	(0.068)
Not READ Aarhus X Post READ	-0.185	-0.201	-0.179
	(0.088)	(0.085)	(0.086)
Mean of control	-0.081	-0.081	-0.081
Observations	96965	96965	96965
Schools (clusters)	699	699	699
Adjusted R-squared	0.001	0.000	0.093
Fixed Effects	No	Yes	Yes
Covariates	No	No	Yes

Notes: Standard errors clustered at the school level in parentheses. The full list of the included covariates is reported in Table 3.

Table C.9: PLACEBO test of Difference-in-difference estimates on students' reading skills. Using 2016 as treatment year. Total score and the three subdomains (OLS)

LOCAL CONTROL GROUP				
	(1)	(2)	(3)	(4)
	Total score	Language comprehension	Decoding	Text comprehension
Placebo post treatment	-0.060	-0.046	-0.070	-0.040
	(0.054)	(0.047)	(0.052)	(0.057)
Local treatment X Placebo post treatment	-0.026	0.013	-0.033	-0.049
	(0.092)	(0.072)	(0.089)	(0.096)
Constant	0.077**	0.068**	0.083**	0.050*
	(0.023)	(0.018)	(0.022)	(0.024)
Observations	5494	5494	5494	5494
Schools (clusters)	46	46	46	46
Adjusted R-squared	0.001	0.000	0.002	0.001
Fixed Effects	Yes	Yes	Yes	Yes
NATIONAL CONTROL GROUP				
	(1)	(2)	(3)	(4)
	Total score	Language comprehension	Decoding	Text comprehension
Placebo post treatment	0.011	0.008	0.004	0.015
	(0.015)	(0.013)	(0.015)	(0.015)
Local treatment X Placebo post treatment	-0.096	-0.041	-0.107	-0.105
	(0.075)	(0.056)	(0.073)	(0.078)
Constant	0.007	0.008	0.011	0.001
	(0.007)	(0.006)	(0.007)	(0.007)
Observations	61734	61734	61734	61734
Schools (clusters)	676	676	676	676
Adjusted R-squared	0.000	-0.000	0.000	0.000
Fixed Effects	Yes	Yes	Yes	Yes

Notes: Standard errors clustered at the school level in parentheses.

C.4 Robustness of Implementation Analyses

Table C.10: Baseline balance

	(1) Invite, not part	(2) Lower third	(3) Midlle third	(4) Upper third	(5) 1-2	(6) 1-3	(7) 1-4
Student level							
Child is a boy	0.51	0.52	0.53	0.53	-0.01	-0.02	-0.02
					(0.69)	(0.09)	(0.15)
Child's age (2016)	8.09	8.08	8.09	8.09	0.01	0.00	0.01
					(0.31)	(0.92)	(0.69)
Child immigrant	0.11	0.17	0.10	0.09	-0.06	0.01	0.03
					(0.02)	(0.37)	(0.18)
Mother compulsory education (2014)	0.14	0.17	0.14	0.14	-0.03	-0.00	0.01
					(0.06)	(0.92)	(0.55)
Mother upper secondary education (2014)	0.06	0.05	0.05	0.05	0.01	0.01	0.00
					(0.21)	(0.21)	(0.55)
Mother vocational education (2014)	0.30	0.30	0.32	0.34	-0.00	-0.01	-0.04
					(0.95)	(0.48)	(0.08)
Mother short-cycle education (2014)	0.05	0.04	0.05	0.05	0.01	0.00	-0.00
					(0.10)	(0.60)	(0.87)
Mother medium-cycle education (2014)	0.27	0.25	0.28	0.29	0.02	-0.01	-0.01
					(0.19)	(0.42)	(0.25)
Mother long-cycle education (2014)	0.13	0.13	0.12	0.10	0.01	0.01	0.03
					(0.79)	(0.51)	(0.15)
Father compulsory education (2014)	0.17	0.18	0.17	0.15	-0.01	-0.00	0.01
					(0.34)	(0.92)	(0.39)
Father upper secondary education (2014)	0.05	0.05	0.05	0.05	0.00	0.01	0.00
					(0.40)	(0.29)	(0.54)
Father vocational education (2014)	0.40	0.41	0.42	0.46	-0.01	-0.03	-0.07
					(0.60)	(0.18)	(0.01)
Father short-cycle education (2014)	0.08	0.07	0.08	0.08	0.00	-0.01	-0.01
					(0.59)	(0.24)	(0.36)
Father medium-cycle education (2014)	0.14	0.12	0.13	0.13	0.02	0.01	0.00
					(0.22)	(0.59)	(0.80)
Father long-cycle education (2014)	0.14	0.12	0.11	0.10	0.01	0.03	0.04
					(0.59)	(0.17)	(0.04)
Missing test score (2017)	0.04	0.05	0.03	0.03	-0.02	0.01	0.00
					(0.10)	(0.29)	(0.74)
School level							
School size[1]	46.36	43.23	46.59	43.89	3.13	-0.23	2.47
					(0.45)	(0.96)	(0.55)
Average test score (2016)[2]	-0.04	-0.13	-0.03	-0.06	0.09	-0.01	0.02
					(0.20)	(0.90)	(0.76)
Students	15161	2075	2283	2063	17236	17444	17224
Schools	327	48	49	47	375	376	374

Notes: *p*-values in parentheses based on standard errors clustered at the school level. [1]Number of students in second grade. [2]Standardized using the mean and the standard deviation from the national sample in 2017.

Table C.11: Effect of READ by Level of Implementation (Downloads). Models with and without covariates included.

	Categorical variable			Continous variable		
	(1)	(2)	(3)	(4)	(5)	(6)
Invite, not part	-0.012	-0.004	0.002			
	(0.027)	(0.023)	(0.021)			
Lower third	-0.128	-0.072	-0.047			
	(0.060)	(0.051)	(0.046)			
Midlle third	-0.004	0.023	0.025			
	(0.052)	(0.046)	(0.044)			
Upper third	0.074	0.099	0.114			
	(0.043)	(0.043)	(0.044)			
Proportion of downloads across schools				0.121	0.201	0.222
				(0.092)	(0.087)	(0.091)
Girl		ref.	ref.		ref.	ref.
Child is a boy=1		-0.267	-0.268		-0.267	-0.268
		(0.009)	(0.009)		(0.009)	(0.009)
Child age 7 years		ref.	ref.		ref.	ref.
8		0.043	0.045		0.047	0.049
		(0.037)	(0.037)		(0.037)	(0.037)
9		-0.170	-0.179		-0.167	-0.175
		(0.041)	(0.041)		(0.041)	(0.041)
10		-0.358	-0.406		-0.358	-0.403
		(0.145)	(0.147)		(0.145)	(0.147)
Child immigrant=1		-0.279	-0.254		-0.280	-0.255
		(0.023)	(0.023)		(0.023)	(0.023)
Child lives with both parents		ref.	ref.		ref.	ref.
Child living with single parent		-0.078	-0.072		-0.078	-0.072
		(0.012)	(0.013)		(0.012)	(0.013)
Child living with parent in new relationship or not living with own parents		-0.051	-0.053		-0.049	-0.051
		(0.017)	(0.017)		(0.017)	(0.017)
No. of children in family		-0.033	-0.031		-0.034	-0.032
		(0.005)	(0.006)		(0.006)	(0.006)
Mother compulsory education		ref.	ref.		ref.	ref.
Mother upper seconday education (2014)		0.261	0.251		0.263	0.253
		(0.023)	(0.023)		(0.023)	(0.024)
Mother vocational education (2014)		0.112	0.110		0.115	0.112
		(0.016)	(0.016)		(0.016)	(0.016)
Mother short-cycle education (2014)		0.286	0.275		0.291	0.279
		(0.023)	(0.023)		(0.023)	(0.023)
Mother medium-cycle education (2014)		0.340	0.331		0.344	0.334
		(0.018)	(0.018)		(0.018)	(0.018)
Mother long-cycle education (2014)		0.478	0.457		0.480	0.459
		(0.021)	(0.021)		(0.021)	(0.021)
Mother outside labor market		ref.	ref.		ref.	ref.
Mother unemployed		-0.034	-0.030		-0.033	-0.028
		(0.027)	(0.027)		(0.027)	(0.027)
Mother employed		0.048	0.043		0.050	0.046
		(0.016)	(0.016)		(0.016)	(0.016)
Mother's total income (1000 kr.)		0.000	0.000		0.000	0.000
		(0.000)	(0.000)		(0.000)	(0.000)
Mother's age in 2014, y		0.003	0.002		0.003	0.003
		(0.001)	(0.001)		(0.001)	(0.001)

Continues on next page.

Table C.11 continued

Father compulsory education		ref.	ref.		ref.	ref.
Father upper seconday education (2014)		0.262	0.246		0.263	0.248
		(0.022)	(0.022)		(0.022)	(0.022)
Father vocational education (2014)		0.125	0.119		0.126	0.120
		(0.014)	(0.014)		(0.014)	(0.014)
Father short-cycle education (2014)		0.258	0.245		0.256	0.243
		(0.020)	(0.020)		(0.020)	(0.020)
Father medium-cycle education (2014)		0.316	0.300		0.317	0.302
		(0.017)	(0.017)		(0.017)	(0.017)
Father long-cycle education (2014)		0.409	0.382		0.409	0.384
		(0.019)	(0.019)		(0.019)	(0.019)
Father outside labor market		ref.	ref.		ref.	ref.
Father unemployed		0.011	0.014		0.018	0.021
		(0.030)	(0.030)		(0.030)	(0.030)
Father employed		0.063	0.058		0.065	0.059
		(0.017)	(0.017)		(0.017)	(0.017)
Father's total income (1000 kr.)		0.000	0.000		0.000	0.000
		(0.000)	(0.000)		(0.000)	(0.000)
Father's age in 2014, y		-0.001	-0.001		-0.001	-0.001
		(0.001)	(0.001)		(0.001)	(0.001)
Missing on Mother's education (6 categories)		0.138	0.134		0.141	0.138
		(0.029)	(0.029)		(0.029)	(0.029)
Missing on Mother's employment status (3 categories)		0.174	0.126		0.172	0.127
		(0.093)	(0.092)		(0.093)	(0.093)
Missing on Mother's total income (1000 kr.)		-0.237	-0.199		-0.230	-0.196
		(0.141)	(0.139)		(0.142)	(0.140)
Missing on Mother's age in 2014, y		-0.179	-0.192		-0.171	-0.183
		(0.136)	(0.136)		(0.138)	(0.137)
Missing on Father's education (6 categories)		0.125	0.104		0.127	0.106
		(0.029)	(0.029)		(0.029)	(0.029)
Missing on Father's employment status (3 categories)		0.173	0.180		0.172	0.179
		(0.088)	(0.086)		(0.089)	(0.086)
Missing on Father's total income (1000 kr.)		-0.056	-0.074		-0.053	-0.070
		(0.083)	(0.077)		(0.084)	(0.078)
Missing on Father's age in 2014, y		-0.093	-0.086		-0.095	-0.088
		(0.081)	(0.081)		(0.082)	(0.081)
School size			-0.000			-0.000
			(0.000)			(0.000)
School average test score 2016			0.285			0.283
			(0.028)			(0.027)
Constant	0.006	-0.387	-0.322	-0.002	-0.406	-0.339
	(0.016)	(0.060)	(0.064)	(0.013)	(0.060)	(0.064)
Observations	51312	51312	51030	50980	50980	50698
Clusters (Schools/Municipalities)	1140	1140	1130	1133	1133	1123
Adjusted R-squared	0.001	0.139	0.150	0.000	0.140	0.150

Notes: Standard errors clustered at the school level in parentheses.

Table C.12: Effect of READ by Level of Implementation (Downloads). Proportion of downloads in 2 and 4 categories, and continuous variable with non-compliers included.

	(1) 2 categories	(2) 4 categories	(3) Continous, incl. non-compliers	(4) Continous, excl. no student identifiers
Invite, not part	0.002			
	(0.021)			
Lower half	-0.038			
	(0.039)			
Upper half	0.099			
	(0.036)			
Control group, not invited		ref.		
Invite, not part		0.002		
		(0.021)		
1st quarter		-0.099		
		(0.051)		
2nd quarter		0.013		
		(0.054)		
3rd quarter		0.094		
		(0.054)		
4th quarter		0.104		
		(0.042)		
Proportion of downloads across schools			0.196	0.224
			(0.087)	(0.101)
Constant	-0.322	-0.321	-0.325	-0.343
	(0.064)	(0.064)	(0.064)	(0.065)
Observations	51030	51030	51030	49258
Schools (clusters)	1130	1130	1130	1087
Adjusted R-squared	0.150	0.150	0.150	0.149
Student covariats	YES	YES	YES	YES
School covariates	YES	YES	YES	YES

Notes: Standard errors clustered at the school level in parentheses. Model 3 for includes non-compliers, i.e. seven schools that were not invited but participated. Model 4 excludes schools that participated, but did not grant access to student identifiers on use of app. The full list of the included covariates is reported in Table C.11.

D Supplementary Materials

Two experiments were embedded in the trial. First, as illustrated in Figure D.1 two variants of the invitation letter were send to the schools. Second, among schools that accepted the invitations, two variance of the READ program was tested. Below we describe each of these embedded experiments. Figure D.1 contains less information than Figure 2. The purpose is to illustrate the two embedded experiments.

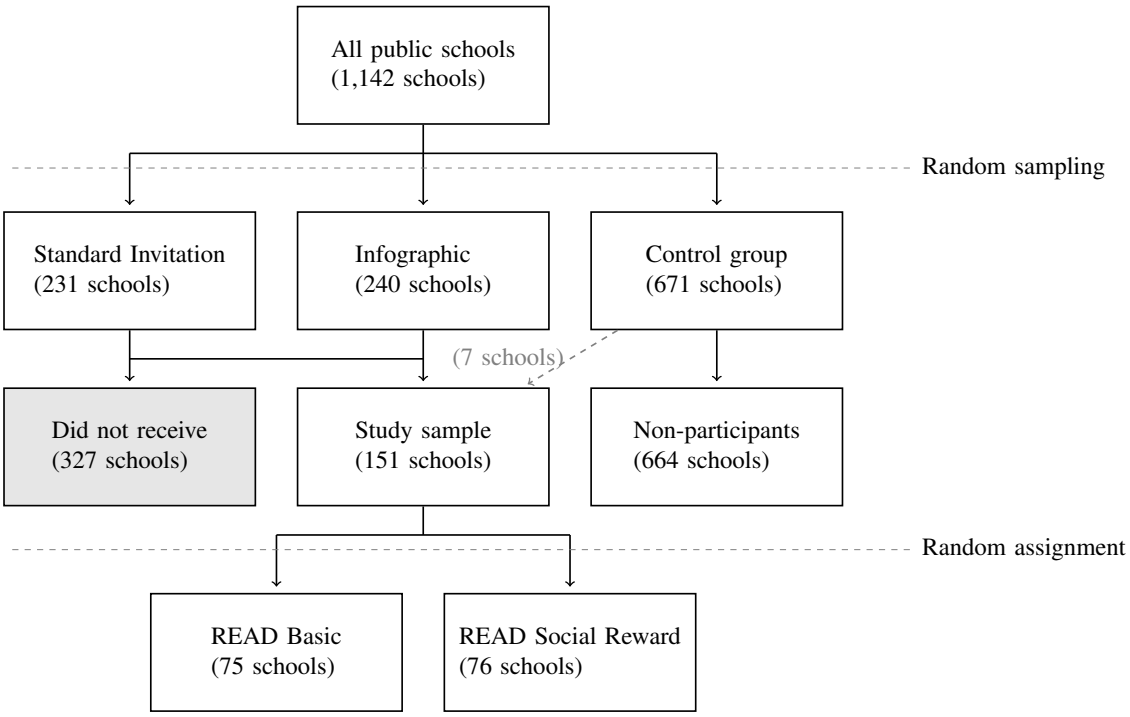
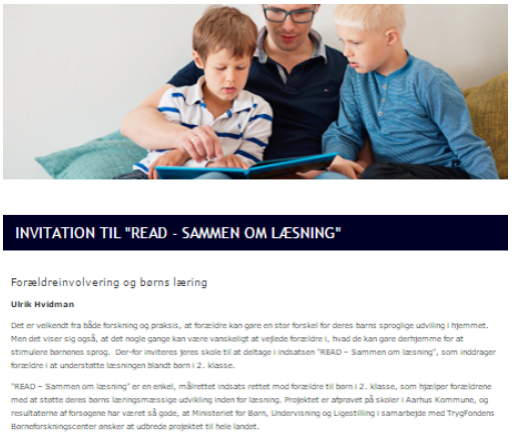


Figure D.1: Design of two embedded experiments

D.1 Invitations

To study how to encourage schools to adopt the program, we randomly assigned schools to one of two versions of the invitation letter. Both groups received an invitation from The Ministry of Education describing the program and its effects. However, since effects from randomized controlled trials may be difficult to convey to persons without a background in research, the one group of schools were assigned to an infographic illustrating the effect of the intervention as estimated in the first randomized controlled-trial in 2014. Apart from infographic, the invitations were identical.

Figure D.2 panel (a) shows the main invitation that all schools receive. Panel (b) shows the infographic that was randomly assigned to half of the schools in the invitation group.



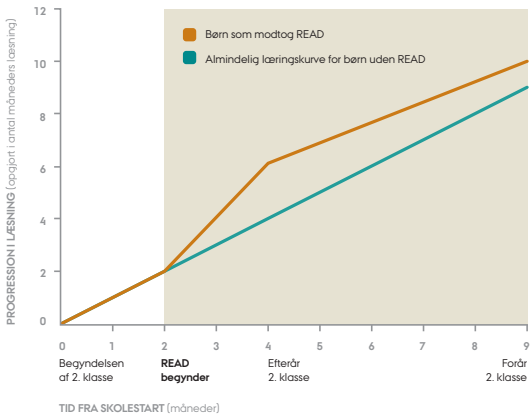
INVITATION TIL "READ - SAMMEN OM LÆSNING"

Forældreinvolvering og børns læring

Ulrik Hvidman

Det er velkendt fra både forskning og praksis, at forældre kan gøre en stor forskel for deres børns sproglige udvikling i hjemmet. Men det viser sig også, at det nogle gange kan være vanskeligt at vejlede forældre i, hvad de kan gøre derhjemme for at stimulere børnenes sprog. Der-for inviteres jeres skole til at deltage i indsatsen "READ – Sammen om læsning", som inddrager forældre i at understøtte læsningen blandt børn i 2. klasse.

"READ – Sammen om læsning" er en enkel, målrettet indsats rettet mod forældre til børn i 2. klasse, som hjælper forældrene med at støtte deres børns læringsmæssige udvikling inden for læsning. Projektet er afprøvet på skoler i Aarhus Kommune, og resultaterne af forsøgene har været så gode, at Ministeriet for Børn, Undervisning og Ligestilling i samarbejde med TrygFondens Børneforskningscenter ønsker at udbrede projektet til hele landet.



Note til figur: Effektevalueringen er gennemført af Aarhus Universitet i samarbejde med Børn og Unge i Aarhus og VIA University College. Den stiplede sorte linje viser den gennemsnitlige udvikling i læsefærdigheder for elever i løbet af de første 9 måneder i 2. klasse, som ikke modtog READ. Den røde linje viser den gennemsnitlige udvikling i tilsvarende periode for børn i 2. klasse, som modtog READ to måneder inde i skoleåret (indførelsen af READ er angivet med vertikale stiplede linje).

(a) Standard Invitation (b) Infographic

Figure D.2: Invitation email. All schools received the standard invitation. Half of the schools were randomly assigned to also receive the infographic.

Table D.1 shows that the two experimental groups in the embedded invitation experiment were balanced on major baseline characteristics.

Table D.1: Balance Invitation experiment

	(1) Standard Invitation	(2) Infographic	(3) 1-2	*p*-values
Average test score 2016	-0.05	-0.05	0.00	(0.93)
Mother high education	0.39	0.38	0.01	(0.34)
Child immigrant	0.11	0.09	0.02	(0.09)
School size	46.29	45.38	0.91	(0.70)
Observations (Schools)	231	240	471	

Table D.2 shows that the Infographic invitation did not increase participation in the READ program significantly.

Table D.2: Effect of Infographic on participation in the READ program

	(1)	(2)
Infographic	-0.063	-0.070
	(0.042)	(0.043)
Mean of control	0.338	0.338
Observations (Schools)	471	468
Adjusted R-squared	0.003	0.001
Covariates	No	Yes

Notes: Standard errors in parentheses. The full list of the included covariates is reported in Table 3.

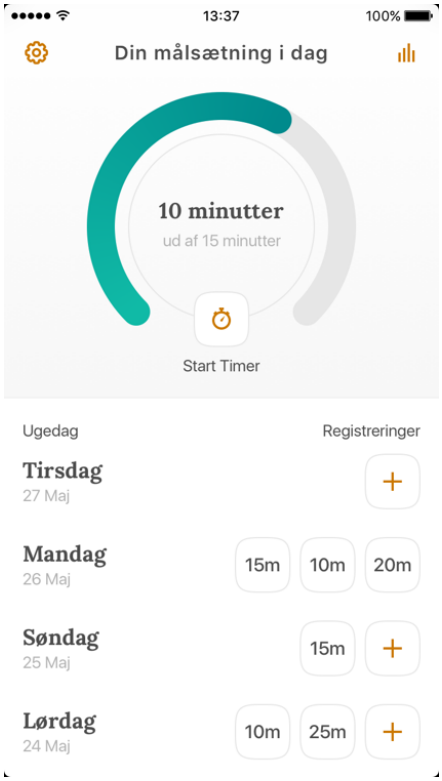
D.2 Social Rewards

Behavioral barriers may constitute a key challenge to the effectiveness of educational interventions that target families. The basic notion behind parent-aimed interventions is that parents will build a better learning environment at home. However, a rapidly growing research literature in behavioral social science has focused on understanding why people often fail to do things they know they should do. Even parents who know what steps to take to significantly improve their children's abilities may fail to take these steps because of behavioral factors (for reviews of behavioral economics of education, see Lavecchia, Liu and Oreopoulos, 2016; Koch, Nafziger and Nielsen, 2015). One explanation may be that parents experience a present bias or lack self-control. As people often discount future outcomes relative to immediate outcomes, it is hard for parents to invest time and effort today for a return on their child's human capital that might show up years later. Moreover, parents may lack self-control and perseverance in their busy everyday lives. As a result, many programs stop at the good intentions because of scarcity of time, energy, and persistence among participants (Mullainathan and Shafir, 2013). In a study of a school information system that provided information to parents, Bergman (2021) found that less than half of the families ever used the system—and that non-users were typically low-income families and families of low-achieving students. The same constraints may be true for interventions that provide resources to parents and try to encourage them to read with their children. Some have proposed approaches to mitigate these behavioral barriers. In a study of the use of a reading application, a treatment group was exposed to three different behavioral tools (i.e., a commitment device, text message reminders, and a social reward). The study suggests that behavioral tools were effective as they increased the usage of the reading application by 1 standard deviation (Mayer et al., 2018).

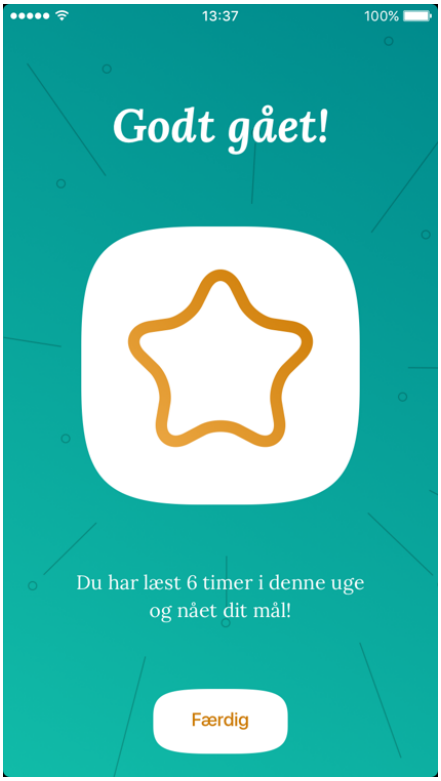
To examine the influence of such behavioral barriers, we randomly assigned participating schools to two versions of the READ program. By making small deviations from the basic READ program, we can test the effect of these modifications. We experimented with social rewards designed to shift preferences by increasing the utility of the current behavior. In the original READ program (Andersen and Nielsen, 2016), some teachers decided to use a logbook in which families

could note every time the child read (as previously described). The logbook endorsed child effort, not performance or results (not the speed or accuracy of the reading). When the children had read ten times, they could bring the logbook to their schoolteacher, and the class would receive a sticker. The class with the most stickers received a prize. In the original trial, use of the logbook was not randomized but selected by teachers. To test the additional effect of this social reward entailed by the logbook competition, we randomly assigned schools that accepted to receive READ to one of two conditions: READ Basic and READ Social Reward. Parents in the READ Social Reward group were provided with the same material as READ Basic, but also with the logbook. As in the original trial, when the children had read ten times, they could bring the logbooks to their schoolteacher, and the class would receive a sticker. One class—the class with most stickers at the school-level—received a prize: a gift card to a reading store worth 10,000 DKK (USD 1,500). This experiment enabled us to test the effect of the social rewards component.

Figure D.3 shows screenshots from the app, where parents could register every time they had read with their child.



(a) Screenshot (I) from READ app



(b) Screenshot (II) from READ app

Figure D.3: Smartphone app

Table D.3 shows that schools assigned to either READ Basic or READ Social Reward were balanced at baseline.

Table D.3: Baseline balance of READ Social Reward relative to READ Basic

	(1) Basic	(2) Social Reward	(3) 1-2	*p*-values
Student level				
Child is a boy	0.53	0.52	0.01	(0.35)
Child's age (2016)	8.08	8.10	-0.02	(0.09)
Child immigrant	0.12	0.12	-0.00	(0.84)
Mother compulsory education (2014)	0.15	0.15	0.00	(0.88)
Mother upper secondary education (2014)	0.05	0.05	0.00	(0.38)
Mother vocational education (2014)	0.32	0.33	-0.01	(0.64)
Mother short-cycle education (2014)	0.05	0.05	-0.01	(0.38)
Mother medium-cycle education (2014)	0.27	0.27	0.00	(0.98)
Mother long-cycle education (2014)	0.12	0.11	0.01	(0.48)
Father compulsory education (2014)	0.17	0.18	-0.01	(0.42)
Father upper secondary education (2014)	0.05	0.04	0.01	(0.28)
Father vocational education (2014)	0.43	0.44	-0.00	(0.86)
Father short-cycle education (2014)	0.08	0.08	-0.00	(0.90)
Father medium-cycle education (2014)	0.13	0.13	-0.00	(0.90)
Father long-cycle education (2014)	0.12	0.10	0.02	(0.27)
Missing test score (2017)	0.03	0.04	-0.01	(0.40)
School level				
School size[1]	45.00	44.62	0.38	(0.92)
Average test score (2016)[2]	-0.08	-0.10	0.02	(0.80)
Students	3375	3391	6766	
Schools	75	76	151	

Notes:*p*-values based on standard errors clustered at the school level. [1]Number of students in second grade. [2]Standardized using the mean and the standard deviation from the national sample in 2017.

Table D.4 shows that READ Social Reward did not change the number of app downloads compared to READ Basic.

Table D.4: Effect of READ Social Reward relative to READ Basic on the number of app downloads

	(1)	(2)
Social Reward	0.040	0.043
	(0.037)	(0.036)
Mean of control	0.228	0.228
Observations	4804	4804
Schools (clusters)	110	110
Adjusted R-squared	0.002	0.032
Covariates	No	Yes

Notes: Standard errors clustered at the school level in parentheses. The full list of the included covariates is reported in Table 3.