Helping the Many by Tutoring the Few: Evidence from a Large-Scale Randomized Field Experiment*

Paul Bingley Jens Dietrichson Vibeke Myrup Jensen
Mette Friis-Hansen

March 24, 2026

Abstract

Being able to read well is a fundamental skill. However, on average around 20 percent of 15-year-olds in OECD countries are not proficient readers. Thus, finding cost-effective methods to help students at risk of reading difficulties is a high priority for educational policymakers. Tutoring shows promise in this regard but in-person, small-group instruction is considered costly and the scaling-up potential uncertain. Using data from a large cluster-randomized field experiment including 81 Danish schools and more than 2,000 students, we estimate the effects of an early literacy tutoring intervention in the first year of primary school. The treatment schools received a program in which the students most at risk of reading difficulties received tutoring in 15-minute sessions, 3-4 times a week, for around 10 weeks. Importantly, tutors were school personnel and the main extra resource supplied to the treatment schools was the program material. The total school-level effect is positive and significant on standardized tests of decoding (0.27 standard deviations, SD) and letter knowledge (0.16 SD). We show that there are substantial effects also on non-tutored peers, which indicate that scaled-up versions of tutoring programs can be more cost-effective than previously thought.

JEL Classification: I21, I24, I26

Keywords: Tutoring, Literacy training, RCT, Kindergarten

*This RCT was registered as AEARCTR-0011007 (Bingley et al., 2024). The research was approved by VIVE's Institutional Review Board 2019-09-16 and, for the updated testing procedures, 2024-03-10 (protocol number 2024/8). We thank Astrid Kofoed Nielsen for excellent research assistance, and are grateful to Gabriel Pons Rotger, Mikkel Helding Vembye and seminar and conference participants at Copenhagen Education Network, IWAEE 2025, ESPE 2025, and VIVE for very helpful comments and discussions. Financial support from TrygFonden (grant nr. 128499) and Independent Research Fund Denmark (grant nr. 3168-00006B) is gratefully acknowledged. All authors are at VIVE–The Danish Center for Social Science Research. Contact: Paul Bingley: pab@vive.dk, Jens Dietrichson: jsd@vive.dk, Vibeke Myrup Jensen: vmj@vive.dk, Mette Friis-Hansen: mfh@vive.dk

1 Introduction

Being able to read well is a fundamental skill. On average, 20% of 15-year-olds in the OECD-countries are not proficient readers (OECD, 2023). Thus, finding cost-effective methods to help students with or at risk of reading difficulties is a high priority for researchers and educational policy makers. A large literature suggests that tutoring is one of the most effective ways of improving at-risk students' skills.[1] However, in-person, one-to-one or small-group instruction is considered expensive and scaling up tutoring programs may be difficult (White et al., 2021; Huffaker et al., 2025). The evidence-base is mainly built on small-scale studies and relatively few large-scale studies show smaller effects Kraft et al. (2024).

We present evidence that scaled-up tutoring programs can be more cost-effective than previously thought. The main reason is positive spillover effects on non-tutored peers of tutored students. While targeted interventions focus on improving outcomes for specific students, spillovers can occur through various channels, such as higher overall instruction levels as tutored students improve, more teacher attention and school resources available for non-tutored peers, smaller class sizes during tutoring, reduced classroom disruptions, adoption of intervention methods for broader use, and peer-to-peer learning effects.

We conducted a large-scale intervention in the spring of 2023 and 2024, in which 2,083 Danish kindergarten students from 81 schools were randomly assigned by their schools to the treatment or control group. The treatment was the literacy program *Læseklar* (meaning 'Reading ready'), a one-to-one or small-group tutoring program for the students most at risk of reading difficulties, which a few students per treatment school received (205 in total).[2]

When comparing all kindergarten students, we find that students in treated schools can, on average, read 2.4 more words on our primary outcome, a standardized test of decoding, than students in control schools (the effect size is 0.27 standard deviations, SD). The effect on

[1] See e.g., Baye et al. (2019), Dietrichson et al. 2017; 2020; 2021, Fryer Jr (2017), Neitzel et al. (2022), Pellegrini et al. (2021) and Slavin et al. (2011) for reviews comparing different types of interventions, including tutoring. See Kraft et al. (2024) and Nickow et al. (2024) for recent reviews of tutoring interventions.

[2] The program has been evaluated in two individually-randomized waiting-list designs: Bøg et al. (2021) found large effects on decoding and letter knowledge in 12 Swedish schools. Seerup et al. (2025) found positive but smaller and not statistically significant effects on similar measures in eight Danish schools. The latter trial was interrupted by the COVID-19 pandemic, which meant that five of the originally 13 schools could not implement the program, and there were additional implementation problems in the eight that could.

letter knowledge is also positive and statistically significant (0.16 SD), despite considerable ceiling effects on this measure. Estimating intention-to-treat (ITT) effects using the sample of students in the lowest fifth of the achievement distribution, in which slightly less than half received the program, we find statistically significant improvements in the treatment group relative to the control group for both literacy outcomes. Using the in-sample standard deviation, the effect sizes are 0.38 SD in decoding and 0.37 SD in letter knowledge.[3]

Our first contribution is to estimate the effect for all students, not just those who received tutoring. Previous research finds that disruptive or low-performing peers reduce academic achievement and longer-run outcomes (Lavy et al., 2012a,b; Kristoffersen et al., 2015; Carrell et al., 2018; Goulas et al., 2024; Carneiro et al., 2025), but the only study we know of that examines the peer effects of a targeted reading intervention similar to ours is Berlinski et al. (2023). They study a third-grade tutoring intervention in Columbia that targets literacy skills in low-achieving students. The intervention improved literacy scores by 0.36 SD among the targeted students and by 0.11 SD among their peers. To decompose the school-level treatment effect into effects on tutored and non-tutored peers, we conduct a series of analyzes under progressively stronger assumptions.[4] Using only information on the mean and minimum scores of the decoding test, we estimate a lower bound of 0.19 SD for the effect on non-tutored peers.

Our second contribution relates to the effects of scaled-up tutoring programs. In addition to involving comparatively many schools and randomizing treatment at the school level–allowing us to analyze spillover effects–our study has two further features that help answer this question.

First, we aimed to approximate how schools would implement the program in practice, outside a research setting. For example, we did not interfere with the selection of students for tutoring. We show that schools selected students differently than they would, had a pre-test cutoff been used (as in e.g., Gersten et al., 2015; Berlinski et al., 2023; Marinelli et al., 2024). Moreover, although we recommended that tutoring did not take place during

[3]Using the full sample SD, the effect sizes are 0.26 and 0.49 SD. However, because effect sizes from tutoring programs are typically calculated using only the tutored students' and/or their control group's standard deviation (Bøg et al., 2021), using the full sample SD likely results in less comparable effect sizes.

[4]In our case, selection into treatment is endogenous, as schools could choose the students who received tutoring.

literacy lessons, schools ultimately decided how to organize tutoring and who would serve as tutors. In practice, schools adopted different approaches: some relied on classroom teachers, some on reading specialists, and others a combination of both.

Second, in contrast to other studies, we provided schools with only limited additional resources. Treatment schools received program materials, including a video course on conducting tutoring sessions, and a follow-up phone call a few weeks into the program. Other large-scale randomized controlled trials (RCTs) of tutoring programs have provided schools directly with tutors (e.g., Gersten et al., 2015; Berlinski et al., 2023; Guryan et al., 2023), with funding to hire tutors/extra teachers or cover costs of tutoring time (Bonesrønning et al., 2022; Gunnes et al., 2025), or ongoing support, coaching, and feedback (Coyne et al., 2019). To our knowledge, only one other large-scale tutoring study both refrained from providing additional tutoring resources and allowed schools to select tutored students: Nunes et al. (2018) implemented a math program using teacher's aides. Although schools could choose which students to tutor, they were required to tutor exactly four students.[5]

Together with the relatively large school-level effect, the limited additional resources make the program comparatively cost-effective. We estimate a cost per unit increase in effect size (CES) of USD 163, which compares favorably with most interventions reported in Rosholm et al. (2021) and Hollands et al. (2013). Among 25 interventions, one preschool language-literacy intervention had a similar CES (Bleses et al., 2018), while a same-age peer-tutoring kindergarten literacy intervention had a lower CES (Stein et al., 2008).

The remainder of the paper is structured as follows: Section 2 provides background and a program description. Section 3 describes the experimental design, sample construction, outcome measures, implementation, and additional data. Section 4 outlines the estimation strategy. Section 5 presents the results. Section 6 discusses potential mechanisms, and section 7concludes.

[5]Nunes et al. (2018) found positive (0.18 SD) but not statistically significant effects for tutored students and did not estimate effects on non-tutored peers.

2 Background

We implemented the tutoring program in kindergarten classes, i.e., the first year of Danish primary school. Children typically start kindergarten the year they turn six (exceptions are possible). Instructions at this stage focus on developing phonological awareness and skills related to conversation, storytelling, writing, and reading. National guidelines issued by the Ministry of Education (2019) also require kindergarten students to recognize all letters of the alphabet.[6] More specific instructional materials and the organization of regular classroom literacy instruction vary between schools (Seerup, 2023).

In the following, we discuss the causes of reading difficulties, the program, and how it aims to help students with these difficulties.

2.1 Reading Difficulties

The program targets students with or at risk of reading difficulties. Virtually all children need help learning the alphabetic principle: the insight that the visual symbols of the writing system (graphemes) represent the distinct units of sound in the language (phonemes; Castles et al., 2018). However, for various reasons, some children are at greater risk of reading difficulties and therefore require additional support. Examples include dyslexia, language and cognitive deficits, language minority and recent immigrant status (Bøg et al., 2021), and having attention and working memory deficits (e.g., Dion et al., 2011; Follmer, 2018; Castles et al., 2018).

Regardless of the underlying causes of reading difficulties, deficits in phonological awareness and letter-sound knowledge are among the most common challenges for affected children (Share, 1995). Phonological awareness enables children to distinguish phonemes and combine them into larger units such as syllables, while phonics training develops knowledge of letters and the ability to map letters or letter patterns to sounds. Explicit instruction in these areas may be particularly important for children at risk of reading difficulties (Lundberg and Høien, 1997). Consistent with this view, a large literature finds that interventions including instruction in phonological awareness, phonics, and decoding can improve literacy

[6]Except q, w, x, and z, which are not commonly used in Danish

skills among at-risk children (e.g., Snow et al., 1998; Ehri et al., 2001; Snowling and Hulme, 2011).

2.2 Description of the Program

2.2.1 Program Aims and Content

The program pursues three primary objectives. First, it develops students' awareness that words consist of both semantic content (for example, "lion" denotes an animal) and a phonological form composed of individual sounds. Second, it strengthens students' understanding of letter–sound correspondences. Third, and most importantly, it fosters students' ability to decode simple words by translating written letter combinations into spoken sounds.

Instruction is delivered through three types of sessions, sequenced to support progressive skill development. The first introduces letter–sound correspondences using clay figurines representing 24 Danish letter sounds, together with a box containing letter-specific compartments and a chart of the finger-spelling alphabet. Each session begins with the selection of three figurines. Students then identify a two-part code to place each figurine in its designated compartment, or "house," labeled with the corresponding letter. The code consists of the initial phoneme of the figurine's name (for example, /l/ in "lion") and the finger-spelling of the associated letter. Students produce the phoneme while simultaneously finger-spelling the letter, before placing the figurine in its compartment. After all three figurines have been placed correctly, the procedure is repeated once, and the session end.

Subsequent sessions begin with a review of previously learned figurines, followed by the introduction of new ones. For students who have difficulty retaining letter–sound associations, tutors may use a mirror to provide visual feedback on articulatory movements when producing letter names and sounds. This multisensory approach reinforces phonological learning and can also be incorporated into the second and third session types.

The second type of session is introduced once students have acquired approximately 12–15 letter sounds, a milestone typically reached within the first month of instruction. At this stage, tutors use images representing short words, compiled in a small booklet. Students articulate the phonemes corresponding to each image, while the tutor transcribes the word on

the facing page. Finger-spelling may be used as a supplementary strategy to support recall of letter–sound correspondences. The booklet also enables students to practice independently or with parental support, as reading accuracy can be verified by comparing the written word with the image.

The third type of session focuses on improving decoding speed and fluency. Students practice decoding short syllables using flashcards and are instructed to decode as many syllables as possible within a one-minute interval. Performance is recorded on a progress chart, and the same set of syllables is then repeated twice more in the same order. This repeated practice typically leads to measurable improvements in decoding performance. Visualizing progress on the chart is intended to enhance student motivation and engagement.

To maintain instructional variety and sustain engagement, the second and third session types are frequently alternated. Tutors may also return to the first type when students encounter words containing letter sounds beyond the initial 12–15 sounds. Together, these three session types provide a structured yet flexible instructional framework for developing early decoding skills.

2.2.2 Program Dosage and Organization

The program included 15-minute sessions per student or group, conducted three to four times per week over a ten-week period, for a total of 30-35 sessions. Schools had flexibility in organizing the sessions. In some cases, classroom teachers conducted the sessions while a teaching assistant or pedagogue supervised the remaining students; in others, sessions occurred during joint-instruction periods across classes. Several schools relied on reading specialists ("*læsevejledere*") as tutors, while others used a combination of classroom teachers and specialists. Across all schools, tutors were classroom teachers, reading specialists, or both.

Program materials included a written manual and three instructional videos illustrating the program's three core components. Monitoring was intentionally kept minimal, consisting of a single follow-up phone call three to four weeks after implementation began. Tutors also completed a brief diary documenting each session's content.

2.3 How the Program Works

The first type of session develops students' awareness that the auditory stream of spoken language can be segmented into distinct phonemes, supporting phonological differentiation. Phonological awareness and letter–sound knowledge are central to early literacy and are widely regarded prerequisites for successful reading acquisition (Bus et al., 1995; Snow et al., 1998; Ehri et al., 2001; Snowling and Hulme, 2011; Share, 1995). Moreover, phonemic awareness training may be more effective when letters are explicitly incorporated into instruction (Ehri et al., 2001).

The second and third session types focus on combining letter sounds into words and improving decoding fluency. Decoding is a prerequisite for reading comprehension (Hoover and Gough, 1990; Hoover and Tunmer, 2018), yet many children struggle to acquire it independently (Castles et al., 2018). Proficient decoding may function as a self-teaching mechanism, supporting the formation of stable word representations and increasing print exposure, which in turn fosters vocabulary growth and comprehension (Share, 1995; Blachman et al., 2014). By strengthening early mastery of the alphabetic principle, the program establishes a foundation for later reading comprehension.

Instruction is provided primarily in small groups, allowing individualized instruction, frequent feedback, and interaction with a supportive tutor. These features align with established learning theories (Vygotsky, 1978; Bandura et al., 1986; Hattie and Timperley, 2007) and likely contribute to the generally strong effectiveness of tutoring interventions (Juel, 1996; Neitzel et al., 2022).

A distinctive feature of the Læseklar program is its use of play-based and multi-sensory instructional methods. Activities such as placing figurines in designated "houses" and using finger-spelling engage kinesthetic, visual, tactile, and auditory modalities. Several mechanisms may explain why play and multi-sensory stimulation supports memory and retrieval. First, figurine-based activities create a narrative structure that may activate episodic memory, which involves recollection of personally experienced events (Tulving, 2002). Second, dual-coding theory suggests that presenting information across multiple sensory channels enables parallel encoding, which may facilitate chunking in short-term memory and reduce

cognitive load (Clark and Paivio, 1991; Shams and Seitz, 2008; Shams et al., 2011).

A related hypothesis is that multi-sensory input may mitigate limitations in working memory capacity. Performance on tasks requiring maintenance of multiple items typically declines as set size increases or under concurrent distractor tasks, particularly when items or distractors are similar (Cocchini et al., 2002; Conlin et al., 2005; Conlin and Gathercole, 2006; Cowan et al., 2014). These limitations are often attributed to interference between memory representations or depletion of domain-specific resources (Oberauer et al., 2016). In Læseklar, moving figurines into their houses may create visual and tactile representations that complement auditory representations of letter sounds, supporting maintenance of letter–sound associations (Bøg et al., 2021). Finger-spelling may further expand effective working memory capacity, allowing students to hold multiple letter–sound combinations while blending them into words. Although modest, this may be consequential given that working memory capacity is limited in early primary school and particularly low among children at risk of reading difficulties (Demetriou et al., 2014; Castles et al., 2018).

The program may also enhance motivation and self-efficacy. Helping figurines "find their homes" is likely more engaging than abstract phonemic tasks. Interaction with a supportive tutor can reinforce students' beliefs in their abilities (Juel, 1996). By intervening in kindergarten, before reading difficulties become entrenched in students' self-perceptions, the program may reduce the risk of motivational and emotional vulnerabilities associated with early reading failure (Nicholls, 1978; Butler, 1999; Muenks and Miele, 2017; Walgermo et al., 2018). Consistent with this, Bøg et al. (2021) found short-term effects of the Swedish version of Læseklar on a self-efficacy-related measure.

3 Data

We use data collected by the research team (pre-tests, outcomes, diaries) alongside administrative registers (descriptive statistics, covariates). We first describe sample construction, including randomization, descriptive statistics, balance tests, and student selection for tutoring, followed by pre- and post-test measures and implementation.

3.1 Sample Construction

For administrative reasons, the trial was conducted over two years (spring 2023 and 2024). Of 792 eligible schools invited, 123 schools accepted. By time of randomization, 40 schools (32 percent) had dropped out, mainly due to changes in school management or teacher leave and turnover. Appendix Figure A1 shows the recruitment and randomization process including attrition at each stage.

A total of 83 school units, nested within 81 schools, were randomly assigned to treatment or control. Units sharing management were treated as one school, even if geographically separated; thus, randomization occured at the school level.

We used stratified randomization to improve balance and statistical power (Athey and Imbens, 2017). Schools were stratified by recruitment year, as first-year schools received treatment before second-year recruitment. Within each year stratum, we applied a matching procedure similar to Bai (2022). Schools were ranked by the proportion of students at risk of reading difficulties based on pre-tests of literacy and language skills (Section 3.2), and then matched into pairs and one triple according to adjacent ranks.[7]

One treatment school dropped out during the 2023 intervention period, and three schools (two treatment, one control) dropped out in 2024. Despite these losses, there was no statistically significant difference in the proportion of at-risk kindergarten students between the treatment and control schools (11.33 vs. 12.55 students on average).

At randomization, 1,293 kindergarten students were enrolled in treatment schools and 1,291 students enrolled in control schools. Table 1 shows the number of students with pre- and post-tests, and the proportion considered at risk. Decoding post-test data are available for 2,014 students (22% attrition).[8] Of 205 students who began tutoring, 180 had a decoding test (12% attrition).

There are three main threats to balance: improper randomization, chance bias, and attrition. Appendix Table A1 reports school-level statistics based on register data with complete information at randomization (i.e., before attrition). We find no evidence of randomization

[7] We used Stata 17's random number generator to assign one school per pair to treatment (StataCorp, 2021). In the triple, randomization assigned one school to treatment, resulting in 40 treatment and 41 control schools.

[8] 18% treatment and 26% control.

Table 1: Observations by Treatment and Control Schools

	Tutored Students		Treatment Schools				Control Schools			
	# Students	%	# School units	%	# Students	%	# School units	%	# Students	%
Pre-test	197	96	40	98	1245	96	41	98	1255	97
Post-test (decoding)	180	88	39	95	1055	82	39	93	958	74
Pre- and post-test	174	85	38	93	1024	79	38	90	934	72
% at risk	145	71			453	35			526	41
Total	205	100	41	100	1293	100	42	100	1291	100

Note. The first two columns display statistics for tutored students. The next columns contain statistics for treatment and control schools, respectively. The total number of schools and students randomized to treatment and control schools is shown in the bottom row (labelled Total). The pre-test row displays the number of students and schools with pre-test scores; the post-test (decoding) row displays the number with information on our primary outcome; and the pre- and post-test row displays the number with both. The % at risk row indicates the share of students at risk of reading difficulties according to the pre-test.

failure: differences are small, and only one is statistically significant on the 5-percent level (henceforth, "significant" refers solely to statistical significance).

Table 2 presents summary statistics and balance tests for students with a primary outcome where attrition at both the student and school-level may cause imbalances. The table includes register data on mother and child characteristics, pre-test variables, and the two post-test outcomes for three groups: tutored students (column 1), students in treatment schools (including the tutored students; column 2), and students in control schools (column 3). Column 4 reports the difference in means between treatment and control school students, with standard errors (in parentheses) clustered at the pair/triple level, from a regression including only an intercept and the treatment indicator.

Table 2: Summary Statistics and Balance Tests

Variable	(1) Tutored Mean (SD)	(2) Treatment Mean (SD)	(3) Control Mean (SD)	(4) Diff (SE)
Register data				
Mother Danish	0.656 (0.477)	0.707 (0.455)	0.736 (0.441)	-0.029
Disposable income q1	0.411 (0.493)	0.329 (0.470)	0.373 (0.484)	-0.044
Disposable income q2	0.317 (0.466)	0.324 (0.468)	0.340 (0.474)	-0.016
Disposable income q3	0.183 (0.388)	0.201 (0.401)	0.175 (0.380)	0.026
Disposable income q4	0.050 (0.219)	0.100 (0.300)	0.078 (0.269)	0.021
Disposable income q5	0.006 (0.075)	0.032 (0.177)	0.014 (0.116)	0.019
Single	0.183 (0.388)	0.182 (0.386)	0.191 (0.393)	-0.009
Missing population register	0.022 (0.148)	0.009 (0.092)	0.001 (0.032)	0.007**
Basic education	0.322 (0.469)	0.234 (0.424)	0.244 (0.430)	-0.010
College education	0.117 (0.322)	0.190 (0.392)	0.141 (0.348)	0.049**
Missing education register	0.033 (0.180)	0.020 (0.140)	0.009 (0.096)	0.011
Employed	0.628 (0.485)	0.691 (0.462)	0.648 (0.478)	0.043
Unemployed	0.044 (0.207)	0.036 (0.187)	0.030 (0.171)	0.006
Outside labor force	0.278 (0.449)	0.250 (0.433)	0.294 (0.456)	-0.045
Missing labor force register	0.039 (0.194)	0.017 (0.130)	0.008 (0.091)	0.009
Parents live together	0.661 (0.475)	0.699 (0.459)	0.668 (0.471)	0.031
Child Danish	0.606 (0.490)	0.657 (0.475)	0.696 (0.460)	-0.038
Boy	0.522 (0.501)	0.503 (0.500)	0.475 (0.500)	0.027
Age in years	7.026 (0.381)	6.992 (0.338)	6.966 (0.339)	0.026
Pre-intervention tests				
At risk language	0.477 (0.501)	0.245 (0.430)	0.281 (0.450)	-0.036
At risk literacy	0.601 (0.491)	0.220 (0.414)	0.268 (0.443)	-0.048*
Within-school language rank	12.744 (13.194)	20.704 (15.961)	20.750 (17.773)	-0.046
Within-school literacy rank	9.772 (9.863)	20.519 (15.813)	21.041 (17.780)	-0.522
Post-intervention tests				
Decoding	6.161 (6.806)	9.479 (9.094)	7.107 (8.520)	2.372***
Letter knowledge	24.817 (4.019)	26.246 (3.132)	25.639 (3.737)	0.607***

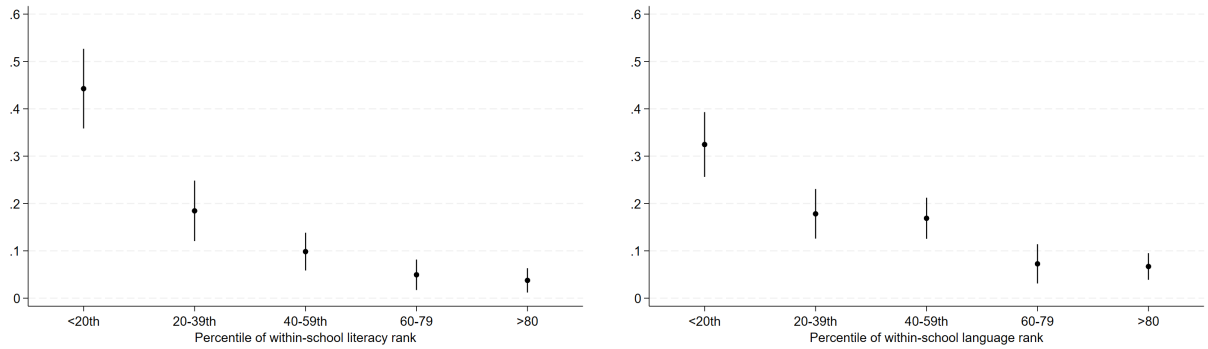
Note. This table reports means and standard deviations for (1) tutored students, (2) all students in treatment schools, and (3) all students in control schools. Column (4) report the difference in means between treatment and control school students. The parentheses contain the standard error, clustered on the pair/triple level, from a regression with an intercept and the treatment indicator.

The table indicates no problematic imbalances between the treatment and control school students, and, as expected, tutored students differ from other students. Aside from the outcome variables, most differences between the students in the treatment and control schools are small. Two differences are significant at the 5-percent level: the treatment group has

more college-educated mothers and fewer mothers with missing labor market information. To further assess balance, we follow Kerwin et al. (2024) and conduct a randomization inference-based omnibus test, examining whether pre-randomization variables jointly predict the treatment indicator.[9] The p-value is 0.299.[10] Thus we cannot reject that the treatment and control groups are balanced, even after attrition.

Tutored students are clearly more at risk than the average student. They more often come from ethnic minority families and where the mother is less likely to be employed, has lower income,and lower levels of education. The students are more frequently identified as at-risk of reading and language difficulties based on pre-tests (Table 2).

Figure 1: Tutored Students Across the Pre-Test Rank Distribution



Note. The figure displays the share and 95-percent confidence intervals (based on standard errors clustered by school) in five groups defined by percentiles of the within-school ranking of students, which are based on the preliteracy (right panel) and language pre-tests (left panel). The shares are estimated using the sample in treatment schools without missing observations on the respective pre-test.

Figure 1 shows the share of tutored students across the pre-test achievement distribution with the left panel using within-school ranks on a preliteracy test, and the right panel using language-test ranks, grouped into five percentile bins. Tutored students are overrepresented below the 20th percentile. Substantial shares in the 20-39th and 40-59th percentiles, and even some above, suggests schools used additional information beyond the pre-tests to select tutored students.

[9]Kerwin et al. (2024) show that conventional omnibus tests using heteroskedasticity or cluster-robust standard errors over-reject the null hypothesis. We implement the test via linear probability models using the *ritest* Stata (19) command (Heß, 2017), re-randomizing the treatment indicator within pairs/triple 1,000 times, and using the F-value to calculate the randomization inference p-value. Missing values are imputed with mean (continuous) or 0 (binary), with indicators included in regressions.

[10]The p-value is 0.703 if we instead use father characteristics

Contacts with schools confirm this, identifying selection criteria not captured by pre-tests, including concerns about progression, grade retention, non-native language background, dual language use, and low performance on other cognitive measures.

3.2 Pre-Tests and Post-Tests

As a pre-test, we primarily used the Danish national kindergarten reading test. The test is developed by Danish and international researchers and inspired by several language proficiency tests, e.g., the MacArthur-Bates Communicative Development Inventory test (Bleses et al., 2017). The test assesses vocabulary, language comprehension, rhyming, and letter knowledge. For each child and sub-test, a total score is calculated and age- and gender-adjusted using a large random sample of children of the same gender and age (in months).

Most schools administered this test, with teachers conducting the assessment. However, schools were eligible as long as they used a norm-referenced test. Because scores are not directly comparable across tests, we use within-school ranks and test-based at-risk indicators rather than raw scores as covariates. For students with missing pre-test data, we impute the mean for ranks and set the at-risk indicator to 0.

Our primary post-intervention outcome is a measure of decoding skills. The measure is a subtest of a standardized dyslexia assessment widely used in Denmark (the 30 real words in "*Elbros ordlister*" from "*Ordblinderisikotesten*"; Gellert and Elbro, 2016). The score equals the number of correctly read words (0–30). The test takes approximately five to eight minutes per student and the test-retest reliability is 0.78 (Gellert and Elbro, 2016). The test is similar in design to the Word Reading Efficiency subtest of the Test of Word Reading Efficiency (Torgesen et al., 1999). The main difference is that the this test imposes a 45-second limit for the total number of words read, whereas the Danish test allows up to 10 seconds per word. The maximum score is also higher in the Word Reading Efficiency test.[11]

As a secondary literacy outcome, we used a letter knowledge test from the standardized

[11]The lower maximum is unlikely to affect our results, only 26 students (1.3%) reached the maximum in our data. Seerup et al. (2025) used the same test in the efficacy trial of Læseklar. The decoding test used in the Swedish trial (Bøg et al., 2021) was modeled after the Test of Word Reading Efficiency (Elwér et al., 2016), and is therefore also comparable.

battery "Reading evaluation at entry-level" (Borstrøm and Klint Petersen, 2004). The score equals the number of correctly named letters of the Danish alphabet (0-29). Reported reliability is high, with Cronbach's $\alpha = 0.88$ (Petersen, 2005; Alinea, 2023).

The post-tests were conducted by the project at the end of kindergarten. In treatment schools, testing occurred on average 23 days after the last tutoring session.[12] Treatment schools were tested, on average, five days later than control schools (difference not significant, $p = 0.16$). Testers were hired by the project and were predominantly third- and fourth year teacher students, master's or bachelor's students in psychology, and a few former teachers. To mitigate ascertainment bias, arising when testers are aware of students' treatment status (Torgerson et al., 2011), we aimed to keep testers blind to group assignment. However, some schools inadvertently revealed their treatment status.

As noted in our protocol, we also collected measures of executive function–consistency of attention, inattention, inhibitory control, and working memory–for a subset of schools in the second recruitment round. Funding for these assessments was secured after the first recruitment round, and not all second-round schools agreed to participate in the additional testing. This resulted in a different sample and necessitated alternative analytical approaches. We therefore report these results in a separate paper.

3.3 Implementation

School-level compliance with treatment assignment was perfect: no control school implemented any component of the program, and all treatment schools conducted some tutoring. However, adherence to the program manual and its recommendations was imperfect.

To approximate a real-world school setting, we allowed considerable flexibility in program implementation. During the intervention period, tutors kept brief diaries documenting each student's progress. These records provide information on treatment dosage and indicators of compliance with the program manual (see Table 3).

The average session lasted 15 minutes, which is in line with the program manual's recommendations. Sessions were held twice a week, below the recommended three to four sessions

[12]This average is based on tutor diaries from 133 tutored students. Some tutored students also received sessions after the post-tests, their values were set to zero in the calculation.

per week, while duration of the intervention (8-10 weeks) was close to the recommended 10-12 weeks. Consequently, the total number of sessions (22) was lower than recommended (32-35). The program has three sessions types; 92% of students progressed to type 2 and 69% to type 3.

Table 3: Expected and Actual Treatment Dosage

Treatment	Recommendation	Actual	SD	N
Time per session (minutes)	15	15.50	5.32	159
# Sessions per week	3–4	1.92	0.62	179
# Weeks	10–12	11.99	3.91	183
Total # sessions	32–35	21.67	6.98	182
Total # hours	8–10	5.72	2.47	157
% Children pass step 1	100	90.63%		192
% Children pass step 2	100	68.75%		192
Taught by teacher		44.37%		151
Taught by reading specialist		48.34%		151
Taught by both		7.28%		151

Note. The table shows the recommended and actual treatment dosage in the Recommendation and Actual-columns, respectively. The SD-column displays the standard deviation of the continuous dosage variables. N defines the number of observations for each variable.

Interviews further indicate that 44% of students received tutoring from their teacher, 48% from a reading specialist; in 7% of cases, tutoring was shared between them, so a single provider cannot be identified. Anecdotal evidence also suggests that the program was implemented during school hours, though at various times throughout the day, depending on the daily schedule. As a result, students typically missed various activities, including joint sessions with other classes, math, literacy, or play sessions.

4 Empirical Framework

We pre-registered a protocol with primary and secondary outcomes in the American Economic Association's Social Science Registry (AEARCTR-0011007 Bingley et al., 2024) and note deviations from it below.

We estimate program effects using a cluster randomized controlled trial (RCT) in which schools are randomly assigned to treatment or control. Simultaneous assignment under the

same temporal conditions mitigates time trends and regression-to-the-mean (Torgerson et al., 2011).

School-level randomization also reduces crossover and spillover risks relative to class- or student-level randomization. Although schools may wish to adopt programs they perceive as successful, the risk is much lower than in designs where schools have direct access to the program through the project. Furthermore, the program is unavailable on the open market in Denmark, and trial schools were prohibited from sharing materials, making spillovers between schools highly unlikely (Seerup et al., 2025).

Randomization ensures that, in expectation, the treatment is independent of school and the student characteristics, so a comparison of school means would be unbiased. To increase statistical power and to estimate effects for different student groups, we use student-level data and include covariates to inproveprecision. In our primary analysis, we estimate the average treatment effects (ATE) using the following equation:

$$y_{i,post} = \alpha + \beta D_i + X_i\gamma + \varepsilon_i \tag{1}$$

In our main equation, $y_{i,post}$ is the outcome variable for student i, and D_i is an indicator equal to 1 for students in treatment schools and 0 for students in control schools (the treatment indicator, henceforth). X_i is a vector of the within-school pre-test rankings, indicators for being at risk of language or literacy difficulties derived from the pre-tests, and indicators for missing pre-test observations,[13] α is the intercept, β is a parameter to be estimated that measures the ATE, and ε is the error term. As D_i is based on the assignment to treatment and not the actual implementation of the program, this is the intention-to-treat (ITT) estimate of the effect.

Following de Chaisemartin and Ramirez-Cuellar (2024), we cluster standard errors at the pair/triple level for pair-matched and clustered randomizations, without including pair/triple indicators (i.e., strata fixed effects) in the primary analysis. Standard asymptotic methods can be sensitive to a small number of clusters or unequal cluster sizes (Cameron and

[13] In all regressions with covariates, we impute the mean for continuous covariates and zero for binary covariates. Imputing the mean or zero and including indicators for missing observations performs well in Zhao and Ding (2024).

Miller, 2015; MacKinnon and Webb, 2017), potentially misestimating standard errors in pair-matched designs (Bai et al., 2022; de Chaisemartin and Ramirez-Cuellar, 2024). While de Chaisemartin and Ramirez-Cuellar (2024) caution that their procedure is less reliable with fewer than 20 pairs, our 40 pairs exceed this threshold. We also report p-values from randomization tests, following recommendations from Athey and Imbens (2017), Bind and Rubin (2020), and Young (2019).

We pre-registered the decoding test as our primary outcome, focusing on students below the 20th percentile in each school's pre-test distribution. We also estimate effects for all students and separately for higher-achieving students (above the 20th percentile). By defining samples based on pre-test scores, we avoid bias from potentially endogenous selection into tutoring within treatment schools.

To contextualize our results, we calculate effect sizes. To account for clustering of students within schools, we estimate the standard deviation using an intercept-only multilevel model. The effect size is the treatment effect divided by this SD, corresponding to Hedges' g (Hedges, 1981). We compute g using the in-sample SD for subgroups, which we consider most comparable to other tutoring studies that typically exclude non-tutored peers.

4.1 Heterogeneity

To explore how the program affected students across the outcome distribution, we run separate analyses for all students, as well as for those below and above the 20th percentile of the pre-test distribution. We use distribution regressions (e.g., Chernozhukov et al., 2013; Goodman-Bacon, 2016), estimating a series of regressions with binary outcomes equal to one if a student's score exceeds each possible threshold up to the maximum.

The regressions are implemented as linear probability models (LPM).[14] To avoid predictions outside [0,1] and because primary results are similar without covariates, we exclude covariates.[15]

We additionally examine effects across the pre-test distribution by interacting the treat-

[14] See Freedman (2008b) and Gomila (2021) for arguments against logit or probit in RCTs.

[15] Distribution regressions also reduce the influence of outliers and respect the categorical nature of the outcome. For comparability with other tutoring studies, we treat outcomes as continuous in the primary analysis.

ment indicator with five pre-literacy percentile groups. The pre-literacy test is used because its ranks are highly correlated with overall pre-test scores ($\rho = 0.67$) and it better predicts the outcomes than the language test.

4.2 Effects on Tutored Students and Non-Tutored Peers

We examine effects on tutored students and non-tutored peers using analyses with progressively stronger assumptions, focusing on the decoding test as our primary outcome due to ceiling effects in the letter knowledge test. The instrumental variables (IV) analysis—but not the bounds analysis—was pre-specified in our protocol.

To assess effects on both tutored and non-tutored students within treatment schools, we first note that subgroup means and the test's minimum score (0) impose restrictions that inform possible effects on non-tutored peers. We then calculate bounds for both groups, assuming the share of would-be tutored and non-tutored students in the control group matches that in the treatment group. We also consider how imposing non-negative treatment effects (Kang and Keele, 2018) or positive effects for tutored students affects the bounds for non-tutored peers.

For point estimates, we use a two-stage least squares (2SLS) IV strategy to estimate effects on students adhering to the program and their non-tutored peers. Adherence at the school level was high: only one school discontinued the program entirely and all other schools are defined as adherent.

The first-stage equation is:

$$T_i = \alpha + \beta_1 D_i \times 1(i \geq 20) + \beta_2 D_i \times 1(i < 20) + X_i \delta_1 + \epsilon_i \tag{2}$$

where the outcome variable is either adhering to tutoring or being a non-tutored peer in a school where at least one student adhered to the tutoring program. The instruments are interactions between the treatment indicator D_i and indicators for being above or below the 20th percentile of the joint pre-test distribution. In addition, X_i contains the same variables

as in the primary analysis. The second stage is:

$$y_{iPost} = \alpha + \rho_1 \hat{T}_i + \rho_2 \hat{T}_i^{peer} + X_i \delta_2 + \epsilon_i \tag{3}$$

where y_{is} is a post-test score, $\hat{T}_i$ is the predicted variable for students who adhere to the program, and $\hat{T}_i^{peer}$ is the corresponding variable for being a non-tutored peer. We cluster standard errors at the pair/triple level in both stages (and in the reduced form regression).

A major problem for IV in our setting is interference between students within schools. While our design guards against interference between schools, one student being tutored may decrease the probability that other students are tutored, and being tutored may affect the outcomes of other students. That is, there is a risk of interference due to instrument-to-treatment and due to treatment-to-outcomes spillovers (e.g., Nibbering and Oosterveen, 2025).

Even in the absence of interference, 2SLS with two endogenous variables requires strong assumptions. Beyond the standard IV conditions for a single endogenous variable (relevance, exclusion restrictions), Bhuller and Sigstad (2024) show that two additional assumptions are needed. First, a form of monotonicity, which, implies that, on average, instrument 1 has a non-negative partial effect on treatment 1 for all agents. Second, there can be no "cross effects" meaning that after controlling for the instruments, instrument 1 has no partial effect on treatment 2 for any agent.[16] Furthermore, Keane and Neal (2023) argue that the conventional 2SLS t-test have serious limitations and recommend using the Anderson-Rubin test (in the one-endogenous-variable, one-instrument case) or a conditional likelihood-ratio test (in the overidentified case), instead.

Despite these strong assumptions, 2SLS may still produce less biased estimates than, for example, regression or matching-methods, providing a better–but not a entirely correct–decomposition of effects.

[16] An alternative method by Fusejima (2024) relies on rank similarity, requiring that each student's outcome rank is invariant across treatment conditions except for non-systematic, identically distributed "slippages", which is also a strong assumption (Chernozhukov et al., 2017).

4.3 Testing Multiple Hypotheses

Our empirical strategy involves testing multiple hypotheses. Testing multiple hypotheses risks increasing the family-wise error rate–the risk of rejecting at least one true null hypothesis above the desired level of statistical significance (Romano and Wolf, 2005)–and the false discovery rate–the expected proportion of falsely rejected null hypotheses (Benjamini and Hochberg, 1995). To mitigate these problems, we clearly designated one primary outcome in our pre-registration (i.e., the decoding test) and one primary specification (i.e., containing the students below the 20th percentile in the pre-test distribution in each school). Furthermore, the statistical significance of the school-level treatment effect on our primary outcome is not sensitive to adjusting for multiple hypothesis testing (e.g., by a Bonferroni adjustment; Dunn, 1961).

4.4 Cost-Effectiveness

We measure cost-effectiveness as the costs per unit increase in effect size (CES) per student at the school level (Hollands et al., 2013). As the program's end goal is to teach students decoding skills, we focus on this outcome and the school-level treatment effect in the cost-effectiveness analysis. The effect size is Hedges' g, calculated as described in Section 4.

We use the ingredients method to calculate the costs (Hollands et al., 2013; Levin and Belfield, 2015). The method accounts for the opportunity cost of all of the resources required to implement the intervention. Consistent with our focus on school-level treatment effects, we calculate the program implementation costs for schools (excluding program development costs). Because we measure the effect for all participating students in the schools, not just the tutored students, and because school personnel implemented the program, the only extra resources used by treatment schools during the intervention are the cost of the program materials and the time the research team spent on follow-up phone calls.[17] All other costs (e.g., tutor preparation, session time) are re-allocations of personnel resources within the school. That is, these costs would have also occurred if the school had not implemented the program.

[17] We reimbursed control schools for participating in the post-intervention tests. As this cost would not have occurred outside of the study, we did not include it in the cost calculation.

Setting the opportunity costs to zero for re-allocations of school resources rests on the assumption that there are no unmeasured effects of these re-allocations. The potential effects on the decoding skills of non-tutored, same-grade peers are captured by our outcome tests. To examine effects on non-targeted cohorts, we use national reading tests administered in second, third, fourth, sixth, and eighth grades. The youngest of these cohorts were in grade 1 during the implementation of our study. We could obtain test information from 77 treatment and control schools from the Danish administrative registers.

It is also conceivable that literacy tutoring has spillover effects on other skills, which we do not measure. Earlier studies indicate spillover effects that are positive or null on social-emotional skills (Morgan et al., 2008; Fives et al., 2013; Bøg et al., 2021), executive functions (Dietrichson et al., 2025; Seerup et al., 2025), and math (Bailey et al., 2020; Lindström-Sandahl, 2024). That is, we seem more likely to underestimate than overestimate the cost-effectiveness when we do not include measures of other skills.

A remaining concern is the effects on kindergarten classes in treated schools that did not participate in the study. We will first be able to address this concern when data from the national reading tests in second grade become available for our study cohorts.

5 Results

This section presents our results. We first report ITT estimates of the program's effect based on school-level random assignment to treatment. We then examine heterogeneity across the outcome and pre-test distributions, as well as effects on tutored students and their non-tutored peers.

5.1 Intention to Treat Estimates of the Program Effect

Table 4 reports the ITT estimates from our primary analysis. Odd-numbered columns use the decoding test as the outcome, while even-numbered columns use the letter knowledge test. Columns 1-2 use the full sample, 3-4 restrict the sample to students scoring at or below the 20th percentile of the joint pre-test distribution, and columns 5-6 include students above the 20th percentile.

Table 4: ITT Estimates of the Average Treatment Effects

Variables	(1) All Decoding	(2) All Letters	(3) $\leq$ 20th Decoding	(4) $\leq$ 20th Letters	(5) $>$ 20th Decoding	(6) $>$ 20th Letters
Treatment	2.43***	0.55***	2.30***	1.70***	2.49***	0.27
	(0.71)	(0.20)	(0.68)	(0.49)	(0.82)	(0.17)
Constant	6.00***	26.12***	3.77***	24.70***	6.41***	26.44***
	(0.82)	(0.29)	(0.83)	(0.72)	(0.85)	(0.26)
Covariate set	Pretest	Pretest	Pretest	Pretest	Pretest	Pretest
Students	2014	2044	426	426	1533	1563
R^2	0.19	0.17	0.12	0.12	0.15	0.11
g (in-sample SD)	0.27	0.16	0.38	0.37	0.27	0.10
Randomization p	0.004	0.008	0.001	0.001	0.005	0.135

Note. All specifications include the following pre-test covariates: the within-school pre-literacy and language score rankings, the indicators for literacy and language at-risk status, and indicators for missing observations. The exception is the indicator for missing the language rank in columns 3-5 and indicator for missing literacy at-risk status in column 6, which were omitted because there were too few with missing observations. All covariates are omitted from the table for brevity. Average treatment effects are given by the coefficient on the Treatment variable. The constant represent the average for students at the mean of the rank variables who also had non-missing pre-tests. Standard errors clustered on the 40 pairs/triples are shown in parentheses below the coefficient. Effect sizes (Hedges' g), calculated as the effect estimate divided by the estimated SD for the sample at hand (i.e., the "in-sample" SD separately for each outcome/sample). We estimate the SD using an intercept-only multilevel model to account for clustering of students within schools. *** $p < 0.01$, ** $p < 0.05$, $p < 0.1$.

The estimates indicate a positive and significant effect on decoding skills, our primary outcome, in the full sample.[18] Students in treated schools read, on average, read 2.43 more words than students in control schools ($g = 0.27$). The effect on letter knowledge is also positive and significant ($g = 0.16$). As indicated by the constant–representing the control group mean for students at the mean of the rank variables and with non-missing pre-tests, many children score at or close to the maximum on this test. That is, there are considerable ceiling effects.

The effects on decoding are positive and significant for both lower and higher-achieving students. Although the point estimates are similar in terms of the number of words (2.30 and 2.49), effect sizes calculated using the in-sample SD are larger for students in the lower

[18]The p-value for the school-level treatment effect on decoding in column 1 is 0.004. Applying a Bonferroni-adjustment and dividing the 5% significance level by three–corresponding to the three hypothesis tests on the primary outcome variable–yields 0.017. The coefficient remains significant under this conservative adjustment.

part of the distribution–where a larger share received tutoring ($g = 0.38$ vs. 0.27). Effects on letter knowledge are also larger for lower-achieving students ($g = 0.37$ vs. $g = 0.10$), and are not significant for higher-achieving students. The larger difference for this measure is likely reflects more severe ceiling effects among higher-achieving students. However, as indicated by the constant, the ceiling effects likely affected both groups.

Given the small share of tutored students and that we estimate ITT effects, the effect sizes at the bottom of Table 4 are substantial. A similar impression emerges when comparing the estimates to the control-group mean: 34% for the full sample and 82% for low-achieving students on the decoding outcome.

Inferences are robust to randomization test-based p-values. The last row of Table 4 shows that all randomization p-values are below 0.01, except–consistent with the primary analysis–letter knowledge among higher-achieving students.

The results are also robust to alternative specifications. As shown in Table 2, school-level treatment effect for both decoding and letter knowledge are very similar to the primary analysis estimates in Table 4 when we exclude covariates and compare means directly. In the appendix, we report results from several alternative specifications: we follow Lin (2013) and include interactions between the treatment indicator and covariates,[19], we include fixed effects for randomization strata, use only register-based covariates with minimal missing data, and incorporate both pre-test and register-based covariates. The sign and significance of the estimates remain unchanged, and point estimates are broadly similar–sometimes slightly larger and sometimes slightly smaller than in the primary analysis.

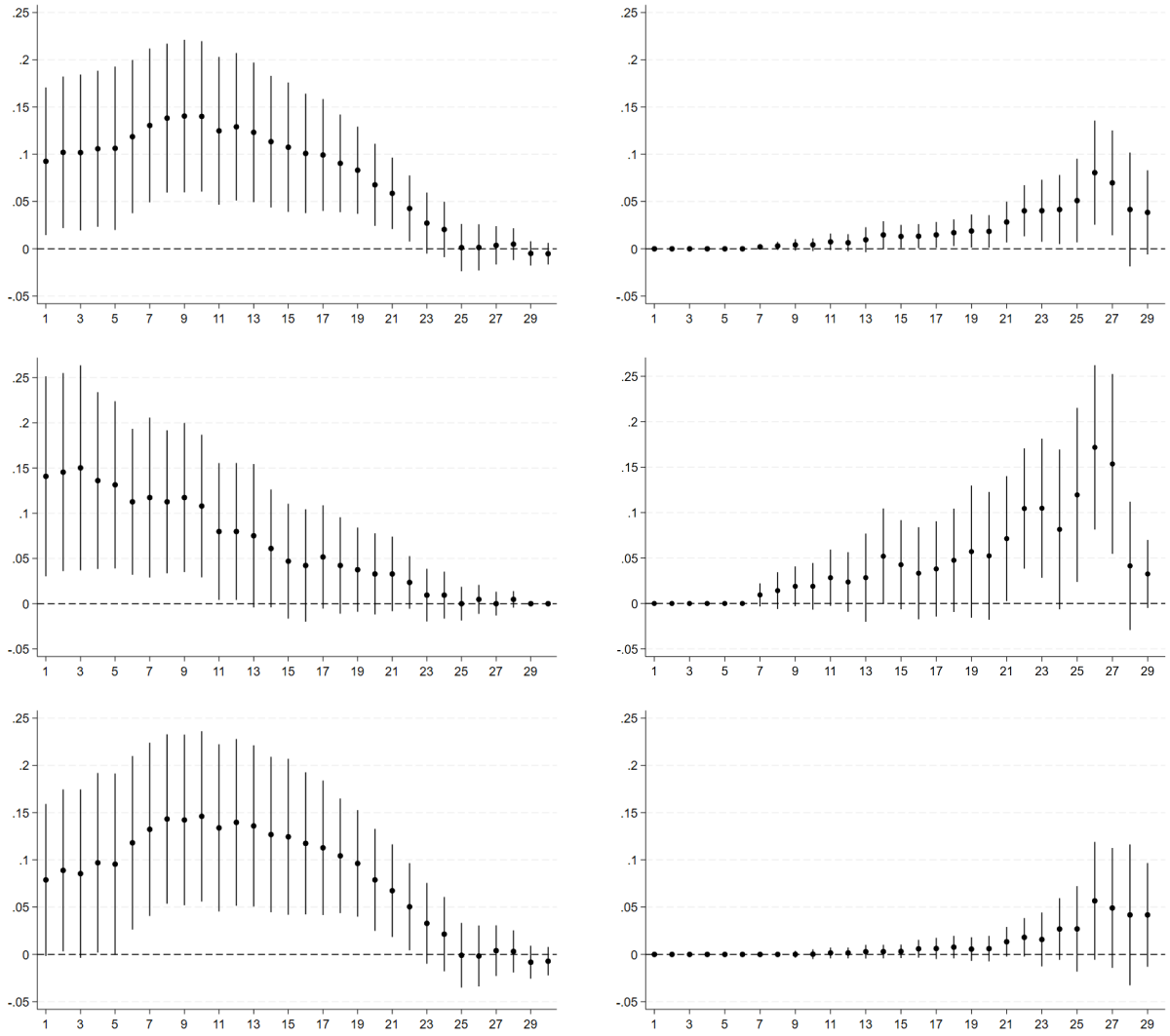
5.2 Heterogeneity

This section examines the heterogeneity in the ITT estimates. Figure 2 reports coefficients and 95% confidence intervals from distribution regressions. The outcome is an indicator equal to one if a student can decode at least the number of words (or letters) on the x-axis; thus, positive coefficients imply higher probabilities of reaching each threshold in treatment

[19] As our randomization produced treatment and control groups of unequal size, including covariates can introduce bias (Freedman, 2008a). Lin (2013) suggests that interacting the treatment variable with all covariates removes most of this bias (continuous variables are standardized with mean zero and standard deviation of one in each sample).

schools. The top panel presents estimates for the full sample, the middle panel for students at or below the 20th percentile of the pre-test distribution, and the bottom panel for those above. The left-hand panels use decoding as the outcome, while the right-hand panels use letter knowledge.

Figure 2: Average Treatment Effects Across the Outcome Distributions



Note. The figure reports coefficients on the treatment indicator and 95% confidence intervals (clustered by pair/triple) from a series of LPMs. Outcomes are binary indicators of decoding and letter knowledge post-test performance. The y-axis shows treatment effects, the x-axis indexes thresholds. For threshold k, the outcome equals one if the students decodes (or knows) at least k words (letters), and zero otherwise, for k=1,...,30 (decoding) and k=1,...,29 (letter knowledge).

In the upper-left panel, treatment effects are positive over most of the decoding distri-

bution. Estimates are statistically significant from 1 to 22 words, with magnitudes of 5-15 percentage points. Relative effects are sizable: the share able to read at least one word is 14% higher in treatment schools, while the shares able to read at least 10 and 22 words are 46% and 45% higher, respectively. Effects are negligible beyond 25 words, consistent with the sparse upper tail (the 95th percentile starts at 26 words).

Patterns differ between the middle-left and bottom-left panels. In the middle panel, effects are about 15 percentage points at low thresholds, decline approximately linearly, and become statistically non-significant at 12 words and above. Relative effects remain large: the share able to read at least one word is 34% higher in the treatment group, and the share able to read at least 12 words is 94% higher.

In the bottom panel, effects are around 10 percentage points up to five words, then increase in magnitude and statistical significance through 22 words. Relative effects are broadly similar to those in the full sample.

The right panels of Figure 2 report results for letter knowledge. Patterns are similar for the full sample and the students at or below the 20th percentile: Effects become statistically significant around 21 letters, increase through 26, and then decline, becoming insignificant by 28. Magnitudes are larger for low-achieving students, with a peak of 17 percentage points (41%), compared to 8 percentage points (12%) in the full sample. For students above the 20th percentile, the pattern is similar, but estimates are not statistically significant.

Table 5 reports estimates across five quintiles of the pre-literacy (rank) distribution. Consistent with Figure 1, most tutored students are below the 20th percentile, with relatively few above the 40th percentile.

In Column (1), decoding effects are statistically significant for all groups below the 60th percentile, and for the top quintile. Effects are positive but imprecisely estimated between the 60th and 80th percentiles. In column (2), letter knowledge effects are largest for the bottom quintile and remain positive up to the 60th percentile, but they are small and not significant above, consistent with ceiling effects in higher-performing groups.

Table 5: Heterogeneity Across the Literacy Pre-Test Quantiles

Variables	(1) Decoding	(2) Letters
Treatment×1($i < 20th$)	2.48***	1.80****
	(0.57)	(0.50)
Treatment×1($20th \geq i < 40th$)	2.12**	0.65*
	(0.83)	(0.34)
Treatment×1($40th \geq i < 60th$)	3.88***	0.78**
	(0.96)	(0.33)
Treatment×1($60th \geq i < 80th$)	1.27	0.19
	(1.23)	(0.31)
Treatment×1($i \geq 80th$)	3.02**	-0.03
	(1.13)	(0.21)
1($i < 20th$)	1.48***	22.46***
	(0.34)	(0.61)
1($20th \geq i < 40th$)	4.35***	25.31***
	(0.56)	(0.31)
1($40th \geq i < 60th$	5.44***	25.97***
	(0.65)	(0.32)
1($60th \geq i < 80th$)	9.15***	26.62***
	(0.85)	(0.23)
1($i \geq 80th$)	13.31***	27.56***
	(0.70)	(0.15)
Missing pretest	-1.85	-2.05***
	(1.44)	(0.42)
Students	2,014	2,044
R^2	0.59	0.99

Notes: Treatment effects by percentile group x are given by the coefficients on Treatment $\times 1(x)$. The intercept is omitted. Standard errors clustered at the pair/triple level (40 clusters) are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

5.3 Effects on Tutored Students and Non-Tutored Peers

This section decomposes the effects into impacts on tutored students and their non-tutored peers. We first demonstrate that non-tutored peers must be affected for the estimates to be consistent with known test properties and observed treatment and control group means. We then derive bounds on these effects. Finally, under stronger assumptions, we decompose

the total effect using an IV approach. Throughout, we focus on our primary outcome, the decoding test.

5.3.1 Existence of Effects on Non-Tutored Peers and Bounds

We first argue that non-tutored peers must be affected, since the effect on tutored students alone cannot explain the total school-level effect. To demonstrate the existence of effects on non-tutored peers, we make the following assumption:

Assumption 1. *The school-level effect is a weighted average of the effect on tutored students and the effect on non-tutored peers where each student has equal weight.*

The estimated school-level effect is 2.43 words (see column (1) in Table 4). Because the minimum score on the decoding test is zero, the maximum effect possible for the tutored students is equal to the mean score among tutored students. This mean is 6.16 words. The sample of treatment school students with post-intervention decoding scores comprises 1,055 students, of whom 180 are tutored and 875 are non-tutored peers. If the effect on non-tutored peers was zero, the implied effect for tutored students would be 14.24 words (2.43/(180/1,055)). That is, under Assumption 1, the effect on tutored students would have to be unrealistically large, indicating that some of the total effect must come from gains among non-tutored peers.

To calculate bounds on the effect for tutored students and non-tutored peers, we introduce an additional assumption:

Assumption 2. *The proportion of tutored students and non-tutored peers would have been the same in the control group as in the treatment group, had the control group been assigned to treatment.*

Because the treatment assignment was random, it is reasonable to assume that the proportion of tutored and non-tutored students in the control schools would have mirrored that in treatment schools–approximately 17% of students would receive tutoring (180 out of 1055) and 83% would be non-tutored peers (875 out of 1055).

We calculate the lower bound for both tutored students and their non-tutored peers by assigning the maximum effect to the other group. Under assumptions 1 and 2, and with an effect on tutored students of 6.16 words, the effect for non-tutored peers cannot exceed 1.66

words if the overall weighted mean is to remain 2.43 words.[20] In addition, summary statistics show that there are more students scoring zero in the control schools (335 students score zero) than the share that would have been tutored by our assumptions (0.17 out of 958 control group students equals 163), so this maximum of 6.16 words is theoretically attainable for the tutored students. Consequently, we obtain a lower bound of 1.66 words for non-tutored peers and an upper bound of 6.16 words for tutored students. Notably, the lower bound for non-tutored peers, derived under mild assumptions, corresponds to 0.19 SD, which is larger than the 0.11 SD effect for non-tutored peers reported in Berlinski et al. (2023).

The maximum effect for non-tutored peers is 10.15 words. With and effect of this magnitude, the effect for tutored students would have to be -33.93 words to maintain the overall weighted mean of 2.43 words. As the minimum score is zero and the maximum score is 30, this effect is clearly impossible. Instead, we first follow the approach for calculating bounds in experiments with spillovers proposed by Kang and Keele (2018), and assume:

Assumption 3. *The treatment effect is non-negative for all students.*

Assumption 3 implies a lower bound for tutored students of 0. Under assumptions 1-3, the corresponding upper bound for non-tutored peers is 2.93. However, assuming a zero effect for tutored students may be overly conservative, given that the effect for students in the lowest 20th percentile is 2.30 words. A more plausible assumption might be:

Assumption 4. *The effect on tutored students is the same as for the students in the lowest 20th percentile.*

Under assumptions 1, 2 and 4, the upper bound for non-tutored students is 2.46.

In sum, by relying only on assumptions about the minimum and mean scores for tutored students and their non-tutored peers, and assuming that the proportions of these groups would have been the same in the control schools, we can derive an informative lower bound for non-tutored peers of 1.66 words and an upper bound for tutored students of 6.16 words. To establish a reasonable upper bound for non-tutored peers, additional assumptions about the effect on tutored students are required. Considering either a zero effect or an effect equal to that of the lowest performing students, the resulting bounds for non-tutored peers

[20]The total effect is a weighted mean of the effects, i.e., 2.43 = 0.17*6.16 + 0.83*X. Solving for X yields (2.43-6.16*0.17)*(0.83) = 1.66.

are relatively narrow: [1.66, 2.93] with an assumption of zero effects on tutored students and [1.66, 2.46] with an assumption of an effect on tutored students equal to that of the students in the lowest 20th percentile.[21] In the next section, we decompose the total effect and quantify the effects on tutored and non-tutored peers using an IV approach.

5.3.2 Decomposing the Total Effect using Instrumental Variables

Table 6 reports the IV estimates, with linear regression results shown for comparison in column (1). The linear regression estimates indicate positive and statistically effects for both groups, with somewhat larger estimates for non-tutored peers.

The IV results in column (5) suggests that OLS does not fully eliminate confounding. The reduced form coefficients in Column (2) are similar in magnitude and statistically significant. Columns (3) and (4) confirm instrument relevance : first stage coefficients are highly significant with F-statistics of 90 for tutored students and 2,418 for non-tutored peers. The second stage estimates in Column (5) suggests a slightly larger effect for tutored students than for non-tutored peers (3.56 words vs. 2.24), although only the latter effect is statistically significant ($p < 0.05$).

The IV approach relies on strong assumptions. However, the estimated effects for both tutored and non-tutored students lie within the bounds derived in section 5.3.1. Under one set of assumptions, the OLS estimates for non-tutored peers fall outside these bounds and imply a smaller effect for tutored students than for their non-tutored peers, which is difficult to justify theoretically. Moreover, if selection into tutoring depends upon unobserved factors not fully captured by the pre-tests, as seems likely, OLS estimates would be downward biased. Taken together, this suggests that despite its strong assumptions, the IV approach may be less biased than OLS.

Expressed as effect sizes, the IV estimates in Table 6 correspond to $g = 0.40$ for tutored students and $g = 0.25$ for non-tutored peers using the full sample SD. As this SD is likely

[21] Applying the method proposed by Kang and Keele (2018), which uses the treatment-on-the-treated effect rather than the weighted mean, and again exploits properties of the test and the means in the treatment group to refine the bounds, yields similar results: [0, 6.16] and [1.31, 3.22]. However, the Kang and Keele (2018) approach assumes that each student's potential outcomes depend only on the number of peers who received the treatment, not on the student's identity, which appears to be a strong assumption in our context.

larger than in most previous tutoring studies, we also use the SD among students below the 20th percentile of the pre-test distribution, yielding $g = 0.59$ for tutored students This exceeds average effects in recent meta-analyzes (Kraft et al., 2024; Nickow et al., 2024), is larger than the decoding effect in the Danish efficacy trial ($g = 0.15$ in Seerup et al., 2025), but smaller than the estimate from the Swedish trial ($g = 1.07$ in Bøg et al., 2021). The Danish trial had similar dosage but more implementation challenges, whereas the Swedish trial aligned with the recommended dosage.[22] Overall, the IV estimate for tutored students appears plausible relative to prior evidence.

Table 6: Effects on Decoding for Tutored Students and Non-Tutored Peers

	(1) OLS	(2) RF	(3) 1st: Tutored	(4) 1st: Peers	(5) IV
Tutored students	2.31***				3.56
	(0.65)				(2.36)
Non-tutored peers	2.60***				2.24**
	(0.77)				(1.10)
Treatment $\geq$ 20th		2.37***	0.11***	0.88***	
		(0.82)	(0.01)	(0.01)	
Treatment $<$ 20th		2.77***	0.41***	0.59***	
		(0.74)	(0.04)	(0.04)	
Covariate set	Pretest	Pretest	Pretest	Pretest	Pretest
Students	2014	2014	2014	2014	2014
R^2	0.19	0.19	0.22	0.73	0.19
F-statistic			90.17	2417.51	

Note. All specifications include an intercept and the set of pre-test covariates used in our primary specification, which are omitted from the table for brevity. Standard errors clustered on the pair/triple are shown in parentheses below the coefficient. *** $p < 0.01$, ** $p < 0.05$, $p < 0.1$.

5.4 Effects on Older Cohorts and Cost-Effectiveness

In this section, we present a simple cost-effectiveness analysis. In school settings, increases in resources for some grades may come at the expense of others, a margin that should be accounted for when assessing cost-effectiveness. We therefore begin by examining outcomes

[22] In Seerup et al. (2025), tutors overused session type 1, reducing time for decoding practice in session types 2 and 3.

for non-targeted cohorts in the 77 treatment and control schools for which administrative test data are available. Outcomes are national literacy tests in grade 2,3,4,6 and 8; the youngest of these cohorts was in grade 1 during program implementation.

Column (1) in Table 7 shows a positive but statistically insignificant average treatment effect (0.037 SD). Consistent with a reallocation of resources toward students needing additional support, the proportion of students in the two lowest performance categories (out of five) is 3.0 percentage points lower in treatment schools compared to control schools. The proportion in the top two categories is 2.8 percentage points higher, but not statistically significant.

There is thus no indications of *negative* spillovers to older students but some indications of gains among lower achievers. Given the lack of consistent statistical significance across outcomes, we treat these estimates as null findings in the cost-effectiveness analysis. In practice, we assume that the primary outcome captures the intervention's effects and set the opportunity costs of within-school reallocation to zero. If the positive patterns in Table 7 reflect true effects, our estimates understate program cost-effectiveness.

Table 7: Treatment Effects on Older Cohorts' Test Scores

	(1)	(2)	(3)
	Score (standardized)	Low score	High score
Treated schools	0.037	-0.030**	0.028
	(0.047)	(0.013)	(0.021)
Year	-0.022	0.022	-0.013
	(0.098)	(0.031)	(0.049)
Grade	0.018**	0.008**	-0.007*
	(0.008)	(0.003)	(0.004)
Constant	43.960	-43.464	26.072
	(198.515)	(63.573)	(98.431)
Number of observations	11,067	11,067	11,067

Note. The table tests the treatment effects on older cohorts' test results. We use the Danish national reading tests in 2nd, 3rd, 4th, 6th, and 8th grade and present three different outcome specifications. Column (1) shows the point score, standardized within each grade to have a mean of zero and a standard deviation of one, while Column (2) defines a dummy equal to one if the student receives a test score in the two lowest categories (out of five). Column (3) defines a dummy for a test score in the two highest categories. We observe test scores for 77 out of the 83 school units. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

The Swedish market price of the program materials at the time of the intervention was

USD 946 (EUR 818) for the first set purchased by a school.[23] Additional sets were available at a reduced price (USD 240/EUR 207).[24] We provided treatment schools with one set per participating class. In total, 69 treated classes (1.9 per treated school; 16 schools contributed only one class) had post-test decoding outcomes, implying material costs of USD 43,391/EUR 37,498. Follow-up phone calls required approximately one hour per treated school; at a wage of USD 80/EUR 69 per hour, this corresponds to USD 2,943/EUR 2,752. Including all inputs, total costs were USD 46,334/EUR 43,337. Dividing by the 1,055 students in treatment schools with post-test decoding outcomes, yields a per-student cost of USD 44/EUR 41. With a school-level effect size is 0.27 SD, the implied cost-effectiveness statistics (CES) is USD 163/EUR 152.

Comparing CES across interventions is challenging. Both effect sizes and cost estimates depend on contextual and methodological factors that vary across studies. Accordingly, the following should be interpreted as indicating that our intervention compares favorably in cost-effectiveness, not as a fully harmonized comparison across interventions.

Closest to our context, Rosholm et al. (2021) report short-term CES (outcomes measured within a year of intervention end) for 18 intervention arms in Danish preschools, schools and families. All interventions were implemented by TrygFonden's Centre for Child Research. Only one arm reports a lower CES (inflation-adjusted CES around 140 USD): LEAP-OPEN, a 20-week universal preschool program using play-based activities to support language and literacy (for more information, see Bleses et al., 2018).

Hollands et al. (2013) examine seven U.S. programs, five involving tutoring.[25] Again, only one study reports a lower CES than ours: K-PALS, a same-age peer-tutoring program targeting alphabetics in kindergarten (see Stein et al., 2008).[26]

While same-age peer-tutoring interventions typically benefit at-risk students (e.g., Slavin

[23] Currency conversions use the 2022-2023 average exchange rate from https://nationalbanken.statbank.dk/nbf/100249.

[24] Personal communication with Anna Aldenius-Isaksson, the Swedish developer.

[25] None of the tutoring studies account for effects on non-tutored peers; excluding spillovers may underestimate cost-effectiveness.

[26] Our end-of-intervention estimates may appear more cost-effective than the tutoring intervention reported in Guryan et al. (2023), which has a benefit-cost ratio comparable to Project STAR class size reductions (Krueger, 2003), and early childhood programs such as the Abecedarian Project (Barnett and Masse, 2007) and the HighScope Perry Preschool Program (Heckman et al., 2010). However, these interventions target broader skill sets and show long-term impacts, making them difficult to compare to our short-term CES.

et al., 2011; Dietrichson et al., 2017, 2020, 2021), it remains unclear whether they suffice for students with more severe reading difficulties (McMaster et al., 2005; Slavin et al., 2011). In such cases, one-to-one or small-group tutoring may be necessary (Gellert et al., 2024).

6 Mechanisms

We find statistically significant and substantial school-level treatment effects, as well as high cost-effectiveness, for an early literacy tutoring program. The effects, and in particular the cost-effectiveness, are partly driven by the effects on non-tutored peers of tutored students. Below, we discuss mechanisms that might explain these findings. First, we outline four common mechanisms to tutoring interventions, then consider study-specific features that may account for larger spill-overs than those reported in Berlinski et al. (2023).

First, as discussed by Berlinski et al. (2023), tutoring programs that pull tutored students out of the classroom temporarily reduce class size for non-tutored peers. Smaller classes may, in turn, improve skill acquisition.[27] In our setting, the class size reduction was minimal–typically one student–and tutoring usually occurred outside literacy instruction (we recommended scheduling during other subjects or free time). Any class-size effect is therefore likely negligible.

Second, tutoring may free up school resources that could be redirected toward non-tutored peers. Such resources include teaching assistants, reading specialists, or special educators. Reallocation may occur as tutored students' skills improve, and they require less support. It may also arise without skill gains–for example, if the classroom teacher provides tutoring, the time a reading specialist would otherwise spend on tutoring can be reallocated to non-tutored peers.

Third, improved literacy skills among tutored students may directly benefit non-tutored peers–for example, through peer learning, by allowing teachers to maintain classroom instruction at a higher level, or by freeing time to support other students.

Fourth, if tutoring improves students' social-emotional skills, classroom disruptions may

[27] See e.g., Krueger (2003) and Fredriksson et al. (2013). Effects are heterogeneous, however, and not all class-size reductions yield positive impacts (see Filges et al., 2018, for a review).

decline, and the overall climate may improve, thereby increasing learning opportunities for non-tutored peers.

Turning to why our effects on non-tutored peers exceed those in Berlinski et al. (2023), one possibility is larger impacts on tutored students. For three out of four mechanisms discussed above, stronger effects on tutored students should translate into larger spillovers. Although estimating impacts on tutored students–and comparing standardized effect sizes across quite different tests–is difficult, our estimates do not appear smaller than those in Berlinski et al. (2023). In addition, three features of our intervention may have contributed to the larger effects.

First, our primary outcome differs from the reading comprehension measure used by Berlinski et al. (2023). Decoding is a narrower skill, which may make large short-term effects easier to achieve. Moreover, while Danish kindergarten guidelines require students to learn the alphabet, they do not impose a similar requirement for decoding (Ministry of Education, 2019). Many students therefore score zero, suggesting that some may need only a small push to grasp the alphabetic principle. Because we observe effects across much of the decoding distribution (see Section 4.1), including among students likely to have learned decoding without the intervention, moving students from zero to reading a few words is unlikely to fully explain the spillovers to non-tutored peers. It may, however, contribute.

In addition, such floor effects might compress the standard deviation relative to Berlinski et al. (2023), inflating standardized effect sizes. Our effects remain sizable when we use a metric less sensitive to this issue (i.e., distribution regression estimates relative to the control group mean; see Section 4.1), though this measure has no direct counterpart in Berlinski et al. (2023).

Second, unlike in Berlinski et al. (2023), and many other large-scale tutoring RCTs, tutors in our trial were drawn from school personnel rather than brought in from the outside. Staff who learn new pedagogical methods while tutoring may also apply them more broadly in the classroom.[28] A related mechanism is that school personnel in treated schools may

[28]In interviews after the Swedish trial (Bøg et al., 2021), several teachers reported using or planning to use the program methods with all students. While we have not systematically surveyed teachers and reading specialists in this trial, similar practices have been mentioned in informal conversations (e.g., using finger-spelling or "hand phonemes" with all students). During the Danish efficacy trial, teachers expressed strong appreciation for the pedagogical methods (Seerup et al., 2025), suggesting a comparable transfer may have

become more motivated, dedicating additional time and attention to literacy instruction in the classrooms. This possibility emerged in discussions with participating schools.

Third, unlike many other tutoring studies, including Berlinski et al. (2023), schools in our study could choose which and how many students to tutor. This flexibility may have led to assignments closer to optimal than a strict pre-test-based cutoff. Schools may have considered a broader set of student characteristics than literacy skills alone (as Figure 1 suggests), or made choices to benefit the entire class.[29]

6.1 Testing Mechanisms

Testing the extent to which these mechanisms operate are challenging. The study was not designed for mechanisms analysis, and we lack both experimental variation and, in some cases, data on potential mediators. Estimating linear-in-means peer effects models relies on strong assumptions (e.g., the reflection problem; Manski, 1993), which features of our study exacerbate. For instance, having school personnel serve as tutors makes it less likely that tutored students' skills are the sole channel affecting peers, weakening the exclusion restriction and making treatment an invalid instrument for skills.

More generally, some mechanisms assume the effects on non-tutored peers operate solely through tutored students' literacy skills, while others allow for direct effects on peers or impacts on other skills. Direct effects may arise from class-size reductions, changes in school-level resource allocation, improvements in social-emotional skills, or adoption of the program's pedagogical methods in the classroom.

Because we have data on literacy skills, which also vary in the control group, we apply the method of Kwon and Roth (2025) to assess whether treatment effects are fully mediated by literacy skills among the bottom 20 percent at the pre-test. Most tutored students fall in this group, so the method provides an indirect test of which mechanisms may explain the effects, while relying on much weaker assumptions than linear-in-means peer effects models (Manski, 1993) or traditional mediator analysis (e.g., Imai et al., 2011; Huber, 2020).[30]

occurred here.

[29] In the interviews conducted alongside the efficacy trial, many teachers reported balancing the needs of tutored students with those of the rest of the class (Seerup, 2023).

[30] The main assumptions in Kwon and Roth (2025) are treatment independence of potential outcomes–likely

We discuss the assumptions and implementation of this method in more detail in the appendix. We find that in some cases, we can reject the sharp null hypothesis of full mediation. Treatment has a particularly strong effect on reducing the number of students unable to decode at all. Using the number of students below the 20th percentile who score zero as the mediator, we can reject the hypothesis that all effects on non-tutored peers above the 20th percentile are mediated through this variable. This suggests that improved literacy among the lowest performing students is not the sole channel driving school-level treatment effects.

7 Concluding Remarks

We evaluate an early literacy tutoring program using a large-scale RCT involving 81 Danish schools and more than 2,000 kindergarten students. We find significant and substantial ITT effects on decoding skills–the program's main target, and our primary outcome. Given that schools received only program materials and minimal additional support, these results are promising for the cost-effectiveness and the scale-up potential of tutoring in realistic school settings.

Effects are positive and significant both below the 20th pre-test percentile, our primary target group, and above it. Gains for the lowest-achieving students are larger in standard deviation units, but higher-achieving students also benefit meaningfully, suggesting the intervention shifts the full school-level achievement distribution rather than only lifting a small subset at the bottom. This pattern indicates that even highly targeted early interventions can generate broad benefits when implemented at scale.

Decomposing school-level effects, we find robust impacts on both tutored students and their non-tutored peers. Effects on tutored students are sizable relative to prior tutoring studies, and spillovers to non-tutored peers are non-trivial. These findings align with Berlinski et al. (2023), who document significant peer effects from a third-grade literacy tutoring intervention in Colombia, reinforcing that peer spillovers are a key channel through which targeted programs can affect aggregate achievement.

satisfied by random assignment–and monotonicity in mediator effects.

The positive spillovers imply that tutoring–whether one-to-one or small-groups and implemented by existing school personnel–can be more cost-effective than previously recognized. Measuring effects at the school level allows our cost-effectiveness calculations to explicitly include benefits for non-tutored peers. Comparing with short-term CES estimates in Rosholm et al. (2021) and Hollands et al. (2013), only 2 of 27 interventions show lower CES than ours. This low CES suggests that low-cost tutoring models relying on local school staff and structured materials may be highly attractive for policymakers with constrained budgets.

Our findings also inform the design and evaluation of future tutoring interventions. First, school-level random assignment and outcome measurement for all students are crucial for capturing spillovers and assessing true social returns. Second, allowing schools flexibility in selecting tutored students and organizing sessions–often viewed as a threat to internal validity–may be advantageous, better reflecting real-world implementation while still producing large gains. Third, the presence of meaningful peer effects highlights the importance of modeling and testing mechanisms linked to classroom composition, instructional level, and teacher time allocation.

Further research should extend these results along several dimensions. Priorities include studying the medium- and long-run consequences of early tutoring on reading comprehension, broader academic outcomes, school trajectories, and social-emotional skills and executive function. Investigating how design features–tutor identity, session frequency, and targeting rules–affect direct and spillover effects and cost-effectiveness is also important. Finally, additional large-scale studies school-level randomized studies that report outcomes for the full student body would help corroborate our finding that scaled-up, low-resource tutoring can be a highly cost-effective strategy for improving early literacy and potentially reducing reading difficulties.

References

Alinea (2023). Baggrundsoplysninger om prøverne i læseevaluering på begyndertrinnet. Technical report, https://www.alinea.dk/laeseevaluering-laerervejledning(open in a new window) (last accessed 2023 March 22).

Athey, S. and Imbens, G. W. (2017). The econometrics of randomized experiments. In *Handbook of economic field experiments*, volume 1, pages 73–140. Elsevier.

Bai, Y. (2022). Optimality of matched-pair designs in randomized controlled trials. *American Economic Review*, 112(12):3911–3940.

Bai, Y., Romano, J. P., and Shaikh, A. M. (2022). Inference in experiments with matched pairs. *Journal of the American Statistical Association*, 117(540):1726–1737.

Bailey, D. H., Oh, Y., Farkas, G., Morgan, P., and Hillemeier, M. (2020). Reciprocal effects of reading and mathematics? beyond the cross-lagged panel model. *Developmental psychology*, 56(5):912.

Bandura, A. et al. (1986). Social foundations of thought and action. *Englewood Cliffs, NJ*, 1986(23-28):2.

Barnett, W. S. and Masse, L. N. (2007). Comparative benefit–cost analysis of the abecedarian program and its policy implications. *Economics of education review*, 26(1):113–125.

Baye, A., Inns, A., Lake, C., and Slavin, R. E. (2019). A synthesis of quantitative research on reading programs for secondary students. *Reading Research Quarterly*, 54(2):133–166.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Berlinski, S., Busso, M., and Giannola, M. (2023). Helping struggling students and benefiting all: Peer effects in primary education. *Journal of Public Economics*, 224:104925.

Bhuller, M. and Sigstad, H. (2024). 2sls with multiple treatments. *Journal of Econometrics*, 242(1):105785.

Bind, M.-A. and Rubin, D. (2020). When possible, report a fisher-exact p value and display its underlying null randomization distribution. *Proceedings of the National Academy of Sciences*, 117(32):19151–19158.

Bingley, P., Dietrichson, J., Friis-Hansen, M., Jensen, V. M., and Seerup, J. (2024). The short- and medium-term effects of an early literacy tutoring intervention. *AEA RCT Registry*.

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

Bleses, D., Højen, A., Dale, P. S., Justice, L. M., Dybdal, L., Piasta, S., Markussen-Brown, J., Kjærbæk, L., and Haghish, E. (2018). Effective language and literacy instruction: Evaluating the importance of scripting and group size components. *Early Childhood Research Quarterly*, 42:256–269.

Bleses, D., Jensen, P., and Højen, A. (2017). Cdi-sprogvurdering. *Valideret redskab til vurdering af sprogudvikling hos børn på*, pages 18–34.

Bøg, M., Dietrichson, J., and Isaksson, A. A. (2021). A multi-sensory tutoring program for students at risk of reading difficulties: Evidence from a randomized field experiment. *The Journal of Educational Research*, 114(3):233–251.

Bonesrønning, H., Finseraas, H., Hardoy, I., Iversen, J. M. V., Nyhus, O. H., Opheim, V., Salvanes, K. V., Sandsør, A. M. J., and Schøne, P. (2022). Small-group instruction to improve student performance in mathematics in early grades: Results from a randomized field experiment. *Journal of Public Economics*, 216:104765.

Borstrøm, I. and Klint Petersen, D. (2004). *Læsetræning på begyndertrinnet*. Alinea.

Bus, A. G., Van Ijzendoorn, M. H., and Pellegrini, A. D. (1995). Joint book reading makes for success in learning to read: A meta-analysis on intergenerational transmission of literacy. *Review of educational research*, 65(1):1–21.

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

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

Carneiro, P., Cruz-Aguayo, Y., Salvati, F., and Schady, N. (2025). Disruption in the classroom: Experimental evidence from ecuador. Technical Report IDB-WP-1688, Inter-American Development Bank.

Carrell, S. E., Hoekstra, M., and Kuka, E. (2018). The long-run effects of disruptive peers. *American Economic Review*, 108(11):3377–3415.

Castles, A., Rastle, K., and Nation, K. (2018). Ending the reading wars: Reading acquisition from novice to expert. *Psychological science in the public interest*, 19(1):5–51.

Chernozhukov, V., Fernández-Val, I., and Melly, B. (2013). Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268.

Chernozhukov, V., Hansen, C., and Wüthrich, K. (2017). Instrumental variable quantile regression. *Handbook of quantile regression*, pages 119–143.

Clark, J. M. and Paivio, A. (1991). Dual coding theory and education. *Educational psychology review*, 3(3):149–210.

Cocchini, G., Logie, R. H., Sala, S. D., MacPherson, S. E., and Baddeley, A. D. (2002). Concurrent performance of two memory tasks: Evidence for domain-specific working memory systems. *Memory & Cognition*, 30(7):1086–1095.

Conlin, J. A. and Gathercole, S. E. (2006). Lexicality and interference in working memory in children and in adults. *Journal of Memory and Language*, 55(3):363–380.

Conlin, J. A., Gathercole, S. E., and Adams, J. W. (2005). Children's working memory: Investigating performance limitations in complex span tasks. *Journal of Experimental Child Psychology*, 90(4):303–317.

Cowan, N., Saults, J. S., and Blume, C. L. (2014). Central and peripheral components of working memory storage. *Journal of Experimental Psychology: General*, 143(5):1806.

Coyne, M. D., McCoach, D. B., Ware, S., Austin, C. R., Loftus-Rattan, S. M., and Baker, D. L. (2019). Racing against the vocabulary gap: Matthew effects in early vocabulary instruction and intervention. *Exceptional children*, 85(2):163–179.

de Chaisemartin, C. and Ramirez-Cuellar, J. (2024). At what level should one cluster standard errors in paired and small-strata experiments? *American Economic Journal: Applied Economics*, 16(1):193–212.

Demetriou, A., Spanoudis, G., and Shayer, M. (2014). Inference, reconceptualization, insight, and efficiency along intellectual growth: A general theory. *Enfance*, 3(3):365–396.

Dietrichson, J., Bøg, M., Filges, T., and Klint Jørgensen, A.-M. (2017). Academic interventions for elementary and middle school students with low socioeconomic status: A

systematic review and meta-analysis. *Review of educational research*, 87(2):243–282.

Dietrichson, J., Filges, T., Klokker, R. H., Viinholt, B. C., Bøg, M., and Jensen, U. H. (2020). Targeted school-based interventions for improving reading and mathematics for students with, or at risk of, academic difficulties in grades 7–12: A systematic review. *Campbell Systematic Reviews*, 16(2):e1081.

Dietrichson, J., Filges, T., Seerup, J. K., Klokker, R. H., Viinholt, B. C. A., Bøg, M., and Eiberg, M. (2021). Targeted school-based interventions for improving reading and mathematics for students with or at risk of academic difficulties in grades k-6: A systematic review. *Campbell Systematic Reviews*, 17(2):e1152.

Dietrichson, J., Seerup, J. K., Jensen, S. E. I., Klejs, J., Bengtsen, E., Strandby, M. W., and Thomsen, M. K. (2025). School-based language, math, and reading interventions for executive functions in children and adolescents: A systematic review. edworkingpaper no. 25-1198. Technical report, Annenberg Institute for School Reform at Brown University.

Dion, E., Roux, C., Landry, D., Fuchs, D., Wehby, J., and Dupéré, V. (2011). Improving attention and preventing reading difficulties among low-income first-graders: A randomized study. *Prevention Science*, 12(1):70–79.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American statistical association*, 56(293):52–64.

Ehri, L. C., Nunes, S. R., Willows, D. M., Schuster, B. V., Yaghoub-Zadeh, Z., and Shanahan, T. (2001). Phonemic awareness instruction helps children learn to read: Evidence from the national reading panel's meta-analysis. *Reading research quarterly*, 36(3):250–287.

Elwér, Å., Fridolfsson, I., Samuelsson, S., and Wiklund, C. (2016). *Läst-test i läsförståelse, läsning och stavning för årskurs 1–6*. Stockholm: Hogrefe.

Filges, T., Sonne-Schmidt, C. S., and Nielsen, B. C. V. (2018). Small class sizes for improving student achievement in primary and secondary schools: A systematic review. *Campbell Systematic Reviews*, 14(1):1–107.

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

Follmer, D. J. (2018). Executive function and reading comprehension: A meta-analytic

review. *Educational Psychologist*, 53(1):42–60.

Fredriksson, P., Öckert, B., and Oosterbeek, H. (2013). Long-term effects of class size. *The Quarterly journal of economics*, 128(1):249–285.

Freedman, D. A. (2008a). On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2):180–193.

Freedman, D. A. (2008b). Randomization does not justify logistic regression. *Statistical Science*, pages 237–249.

Fryer Jr, R. G. (2017). The production of human capital in developed countries: Evidence from 196 randomized field experiments. In *Handbook of economic field experiments*, volume 2, pages 95–322. Elsevier.

Fusejima, K. (2024). Identification of multi-valued treatment effects with unobserved heterogeneity. *Journal of Econometrics*, 238(1):105563.

Gellert, A. and Elbro, C. (2016). Rapport om projekt vedrørende tidlig identifikation af elever i risiko for udvikling af alvorlige afkodningsvanskeligheder (herunder ordblindhed). Technical report, Center for Læseforskning, Københavns Universitet og Undervisningsministeriet.

Gellert, A. S., Petersen, D. K., Borstrøm, I., Poulsen, M., and Elbro, C. (2024). Tidlig indsats mod ordblindhed–en effektundersøgelse med elever i 1. klasse. *NyS, Nydanske Sprogstudier*, (66):36–65.

Gersten, R., Rolfhus, E., Clarke, B., Decker, L. E., Wilkins, C., and Dimino, J. (2015). Intervention for first graders with limited number knowledge: Large-scale replication of a randomized controlled trial. *American Educational Research Journal*, 52(3):516–546.

Gomila, R. (2021). Logistic or linear? estimating causal effects of experimental treatments on binary outcomes using regression analysis. *Journal of Experimental Psychology: General*, 150(4):700.

Goodman-Bacon, A. (2016). The long-run effects of childhood insurance coverage: Medicaid implementation, adult health, and labor market outcomes. Technical report, National Bureau of Economic Research.

Goulas, S., Griselda, S., Megalokonomou, R., and Zenou, Y. (2024). Disruptive peers and academic performance: Short-and long-term outcomes. Technical Report 17539, IZA Dis-

cussion Paper Series.

Gunnes, T., Kirkebøen, L. J., Lindenskov, L., and Rønning, M. (2025). Small-and large-group instruction and a didactic method in mathematics for low-performing adolescents: results from a randomized field experiment. *The Scandinavian Journal of Economics*, 127(3):613–652.

Guryan, J., Ludwig, J., Bhatt, M. P., Cook, P. J., Davis, J. M., Dodge, K., Farkas, G., Fryer Jr, R. G., Mayer, S., Pollack, H., et al. (2023). Not too late: Improving academic outcomes among adolescents. *American Economic Review*, 113(3):738–765.

Hattie, J. and Timperley, H. (2007). The power of feedback. *Review of educational research*, 77(1):81–112.

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P. A., and Yavitz, A. (2010). The rate of return to the highscope perry preschool program. *Journal of public Economics*, 94(1-2):114–128.

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

Heß, S. (2017). Randomization inference with stata: A guide and software. *The Stata Journal*, 17(3):630–651.

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

Hoover, W. A. and Gough, P. B. (1990). The simple view of reading. *Reading and writing*, 2(2):127–160.

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

Huber, M. (2020). Mediation analysis. *Handbook of labor, human resources and population economics*, pages 1–38.

Huffaker, E., Robinson, C. D., Bardelli, E., White, S., and Loeb, S. (2025). When interventions don't move the needle: Insights from null results in education research. Technical report, Annenberg Insitute at Brown University.

Imai, K., Keele, L., Tingley, D., and Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, 105(4):765–789.

Juel, C. (1996). What makes literacy tutoring effective? *Reading Research Quarterly*, 31(3):268–289.

Kang, H. and Keele, L. (2018). Spillover effects in cluster randomized trials with noncompliance. Technical Report 1808.06418v2, ArXiv.

Keane, M. and Neal, T. (2023). Instrument strength in iv estimation and inference: A guide to theory and practice. *Journal of Econometrics*, 235(2):1625–1653.

Kerwin, J., Rostom, N., and Sterck, O. (2024). Striking the right balance: Why standard balance tests over-reject the null, and how to fix it (no. 17217). Technical Report IZA DP No. 17217, Institute of Labor Economics (IZA).

Kraft, M. A., Schueler, B. E., and Falken, G. (2024). What impacts should we expect from tutoring at scale? exploring meta-analytic generalizability. edworkingpaper no. 24-1031. *Annenberg Institute for School Reform at Brown University*.

Kristoffersen, J. H. G., Krægpøth, M. V., Nielsen, H. S., and Simonsen, M. (2015). Disruptive school peers and student outcomes. *Economics of Education Review*, 45:1–13.

Krueger, A. B. (2003). Economic considerations and class size. *The economic journal*, 113(485):F34–F63.

Kwon, S. and Roth, J. (2025). Testing mechanisms. Technical report.

Lavy, V., Paserman, M. D., and Schlosser, A. (2012a). Inside the black box of ability peer effects: Evidence from variation in the proportion of low achievers in the classroom. *The Economic Journal*, 122(559):208–237.

Lavy, V., Silva, O., and Weinhardt, F. (2012b). The good, the bad, and the average: Evidence on ability peer effects in schools. *Journal of Labor Economics*, 30(2):367–414.

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

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *Annals of Applied Statistics*, 7.

Lindström-Sandahl, H. (2024). *Early elementary school interventions in reading and math-*

ematics. Linkopings Universitet (Sweden).

Lundberg, I. and Høien, T. (1997). Levels of approaching reading and its difficulties. *Reading disability and its treatment*, pages 11–36.

MacKinnon, J. G. and Webb, M. D. (2017). Wild bootstrap inference for wildly different cluster sizes. *Journal of Applied Econometrics*, 32(2):233–254.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *Review of Economic Studies*, 60(3):531–542.

Marinelli, H. A., Berlinski, S., and Busso, M. (2024). Remedial education: Evidence from a sequence of experiments in colombia. *Journal of Human Resources*, 59(1):141–174.

McMaster, K. L., Fuchs, D., Fuchs, L. S., and Compton, D. L. (2005). Responding to nonresponders: An experimental field trial of identification and intervention methods. *Exceptional children*, 71(4):445–463.

Ministry of Education (2019). Børnehaveklassen – faghæfte 2019. Technical report, https://emu.dk/sites/default/files/2020-09/GSK$_f agh$

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

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

Neitzel, A. J., Lake, C., Pellegrini, M., and Slavin, R. E. (2022). A synthesis of quantitative research on programs for struggling readers in elementary schools. *Reading Research Quarterly*, 57(1):149–179.

Nibbering, D. and Oosterveen, M. (2025). Policy-relevant causal effect estimation using instrumental variables with interference. Technical report, arXiv preprint arXiv:2509.12538.

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

Nickow, A., Oreopoulos, P., and Quan, V. (2024). The promise of tutoring for prek–12 learning: A systematic review and meta-analysis of the experimental evidence. *American Educational*

Research Journal, 61(1):74–107.

Nunes, T., Barros, R., Evangelou, M., Strand, S., Mathers, S., and Sanders-Ellis, D. (2018). 1stclass number evaluation report and executive summary. Technical report, Education Endowment Foundation.

Oberauer, K., Farrell, S., Jarrold, C., and Lewandowsky, S. (2016). What limits working memory capacity? *Psychological bulletin*, 142(7):758.

OECD (2023). Pisa 2022 results (volume i): The state of learning and equity in education. *OECD Pub.*, pages 1–10.

Pellegrini, M., Lake, C., Neitzel, A., and Slavin, R. E. (2021). Effective programs in elementary mathematics: A meta-analysis. *Aera Open*, 7:2332858420986211.

Petersen, D. K. (2005). Distinkthedstræning i børnehaveklassen. *Dyslexi*, (3):16–19.

Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108.

Rosholm, M., Paul, A., Bleses, D., Højen, A., S. Dale, P., Jensen, P., M. Justice, L., Svarer, M., and Calmar Andersen, S. (2021). Are impacts of early interventions in the scandinavian welfare state consistent with a heckman curve? a meta-analysis. *Journal of Economic Surveys*, 35(1):106–140.

Seerup, J. K. (2023). *Ready to Read: Exploring School-based Tutoring Interventions for Struggling Readers: PhD Thesis*. b University of Copenhagen, Faculty of Social Sciences, Department of sociology.

Seerup, J. K., Dietrichson, J., Jensen, V. M., and Bingley, P. (2025). Tutoring young children at risk of developing reading difficulties: Effects of an early tutoring intervention on literacy skills and executive functions. *Journal of Education for Students Placed at Risk (JESPAR)*, pages 1–31.

Shams, L. and Seitz, A. R. (2008). Benefits of multisensory learning. *Trends in cognitive sciences*, 12(11):411–417.

Shams, L., Wozny, D. R., Kim, R., and Seitz, A. (2011). Influences of multisensory experience on subsequent unisensory processing. *Frontiers in psychology*, 2:264.

Share, D. L. (1995). Phonological recoding and self-teaching: Sine qua non of reading acquisition. *Cognition*, 55(2):151–218.

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

Snow, C. E., Burns, M. S., and Snow, C. E. (1998). *Preventing reading difficulties in young children*. National Academies Press.

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

StataCorp (2021). *Stata: Release 17*. StataCorp LLC.

Stein, M. L., Berends, M., Fuchs, D., McMaster, K., Sáenz, L., Yen, L., Fuchs, L. S., and Compton, D. L. (2008). Scaling up an early reading program: Relationships among teacher support, fidelity of implementation, and student performance across different sites and years. *Educational evaluation and policy analysis*, 30(4):368–388.

Torgerson, C. J., Wiggins, A., Torgerson, D. J., Ainsworth, H., Barmby, P., Hewitt, C., Jones, K., Hendry, V., Askew, M., Bland, M., Coe, R., Higgins, S., Hulme, C., Tymms, P., and Department for Education (DFE), c. c. (2011). Every child counts: The independent evaluation. *London: Department for Education (DfE)*.

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

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

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*, volume 86. Harvard university press.

Walgermo, B. R., Frijters, J. C., and Solheim, O. J. (2018). Literacy interest and reader self-concept when formal reading instruction begins. *Early Childhood Research Quarterly*, 44:90–100.

White, S., Carey, M., O'Donnell, A., and Loeb, S. (2021). Early lessons from implementing high-impact tutoring at scale. *National Student Support Accelerator*.

Young, A. (2019). Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *The quarterly journal of economics*, 134(2):557–598.

Zhao, A. and Ding, P. (2024). To adjust or not to adjust? estimating the average treatment

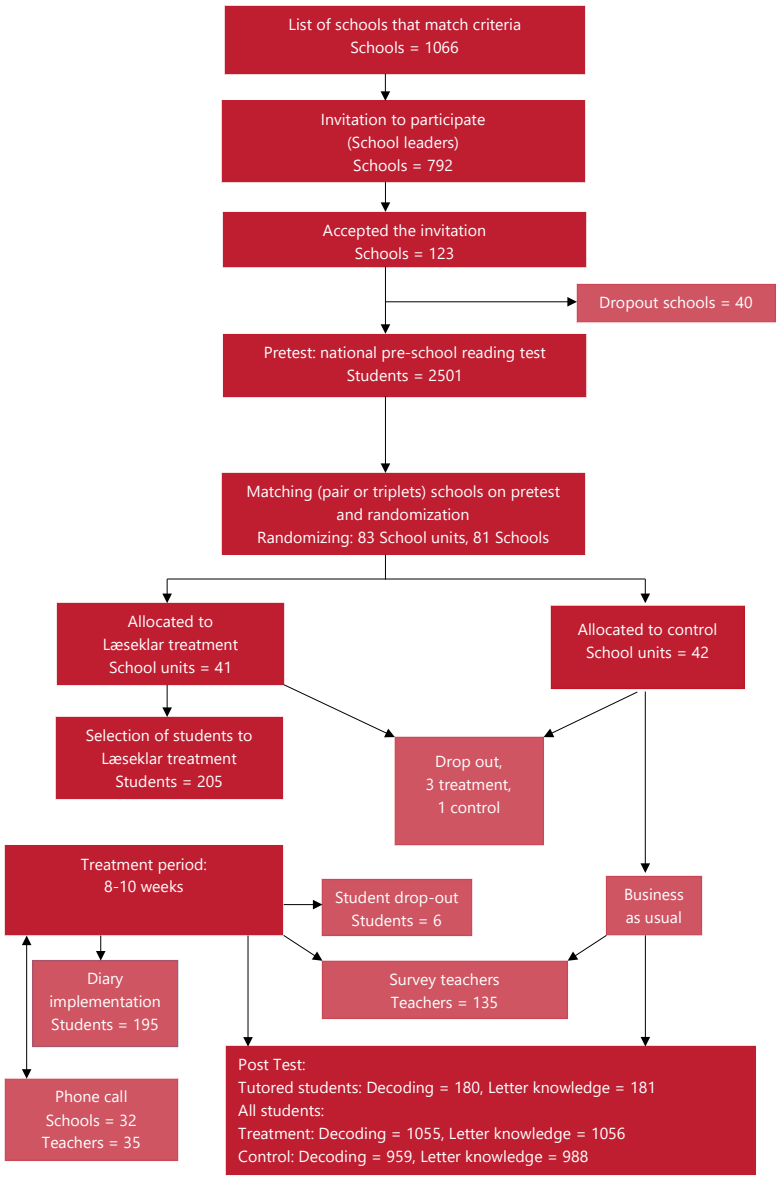
effect in randomized experiments with missing covariates. *Journal of the American Statistical Association*, 119(545):450–460.

8 Appendix

8.1 The recruitment, randomization, and testing process

Figure A1 displays the recruitment, randomization, and testing process, indicating sample size and attrition in each step.

Figure A1: Flowchart



8.2 School-Level Balance Test

Table A1 shows the school-level means and standard deviations of the treatment and control schools for a set of register data based variables for mothers and participating children. We have information regarding these variables for all schools and they are only missing for one student, whom we were unable to match to the registers. That is, attrition is a highly unlikely source of imbalance. All differences appear small and we find one significant difference ($p < 0.05$)–there are significantly more mothers in the treatment group, whose disposable incomes are in the highest quintile.

Table A1: Balance Tests for Schools

	Treatment Schools	Control Schools	
Variable	Mean (SD)	Mean (SD)	Difference
Mother Danish	.707 (.21)	.758 (.18)	-.05
Disposable income q1	.352 (.124)	.395 (.154)	-.043
Disposable income q2	.337 (.089)	.33 (.139)	.007
Disposable income q3	.177 (.091)	.175 (.089)	.002
Disposable income q4	.096 (.068)	.072 (.061)	.024
Disposable income q5	.023 (.036)	.009 (.019)	.015**
Single	.182 (.111)	.197 (.114)	-.015
Missing population register	.008 (.018)	.002 (.01)	.005
Basic education	.259 (.157)	.248 (.115)	.011
College education	.168 (.091)	.13 (.09)	.038*
Missing education register	.017 (.029)	.01 (.022)	.006
Employed	.674 (.101)	.646 (.131)	.028
Unemployed	.036 (.042)	.034 (.037)	.003
Out of labor force	.267 (.096)	.296 (.122)	-.029
Missing labor force register	.014 (.024)	.007 (.015)	.007
Parents live together	.690 (.135)	.662 (.126)	.028
Child Danish	.662 (.222)	.716 (.194)	-.054
Boy	.529 (.114)	.477 (.139)	.052*

8.3 Sensitivity Analysis

This section reports results from alternative specifications of our primary analysis of the ITT effects.

Running the primary analysis at the school level rather than the student level yields results that are very similar but slightly less precise. Excluding covariates, the estimates are 2.33 for decoding ($p = 0.014$) and 0.54 for letter knowledge ($p = 0.024$). Dropping the 5% worst-matched pairs reduces the sample to 72 schools and yields an estimate of 2.28 for decoding ($p = 0.022$) and 0.55 for letter knowledge ($p = 0.030$).

We showed in Table 2 that our results are close to the primary analysis when we omit all covariates and just compare the means in the treatment and control group. In Table A2, we show the results from four additional variations of the primary specification. The set-up mimics Table 4: Odd columns use the decoding test as the outcome, even columns the letter knowledge test. Column 1 and 2 include the full sample, 3 and 4 the students at or below the 20th percentile in the literacy pre-test distribution, and 5 and 6 the students above the 20th percentile in the same distribution.

Overall, we find relatively similar results with these alternative specifications. The school-level treatment effect for decoding is always significant and largest when we include strata fixed effects (i.e., pair/triple dummies) and smallest when we use only register-based covariates. The primary analysis estimate is relatively close to the mid-point between these two estimates.

The only estimates that differ markedly from the primary analysis are the Lin-specifications for students at or below the 20th percentile, shown in Panel A. Both the decoding and the letter knowledge effects are much higher than in the primary analysis. However, the Lin-specification interacts all covariates with the treatment-indicator, and the estimate for the treatment variable shown in the table is the estimate for students at the mean of the two within-school literacy and language rank-variables who were not at risk according to the at risk indicators (and did not have missing data on any of these variables). Especially in this smaller sample, this might be a small and unrepresentative group.

Table A2: Alternative Specifications of the Average Treatment Effects

	(1)	(2)	(3)	(4)	(5)	(6)
	All	All	$\leq$ 20th	$\leq$ 20th	$>$ 20th	$>$ 20th
Variables	Decoding	Letters	Decoding	Letters	Decoding	Letters
Panel A: Lin-specification						
Treatment	2.18**	0.33*	5.88**	3.25**	2.04**	0.21
	(0.91)	(0.18)	(2.37)	(1.29)	(0.95)	(0.17)
Constant	8.57***	26.60***	3.04***	25.16***	8.87***	26.71***
	(0.50)	(0.15)	(0.87)	(0.86)	(0.53)	(0.13)
Covariate set	Lin	Lin	Lin	Lin	Lin	Lin
Students	2014	2044	426	426	1533	1563
R^2	0.19	0.17	0.13	0.12	0.15	0.11
Panel B: Including strata fixed effects						
Treatment	2.72***	0.57**	2.32***	1.82***	3.05***	0.33
	(0.86)	(0.25)	(0.84)	(0.63)	(1.06)	(0.22)
Constant	8.06***	27.28***	2.56***	25.70***	9.14***	27.45***
	(0.61)	(0.14)	(1.04)	(0.54)	(0.74)	(0.14)
Covariate set	Strata FE	Strata FE	Strata FE	Strata FE	Strata FE	Strata FE
Students	2014	2044	426	426	1533	1563
R^2	0.29	0.23	0.27	0.30	0.26	0.17
Panel C: Non-missing register data-based covariates						
Treatment	2.06**	0.57**	1.72**	1.36***	2.13**	0.31
	(0.78)	(0.21)	(0.70)	(0.43)	(0.91)	(0.20)
Constant	7.73***	26.40***	-0.51	25.37***	8.50***	26.92***
	(2.50)	(0.53)	(0.80)	(1.81)	(2.60)	(0.36)
Covariate set	Reg	Reg	Reg	Reg	Reg	Reg
Students	2013	2043	426	426	1532	1562
R^2	0.08	0.05	0.07	0.05	0.08	0.04
Panel D: Pretest and non-missing register data-based covariates						
Treatment	2.16***	0.55***	2.19***	1.70***	2.21***	0.26
	(0.70)	(0.19)	(0.69)	(0.48)	(0.80)	(0.17)
Constant	5.94***	26.55***	-0.72	25.78***	6.45***	26.89***
	(2.05)	(0.56)	(1.65)	(1.10)	(2.16)	(0.43)
Covariate set	Pre+Reg	Pre+Reg	Pre+Reg	Pre+Reg	Pre+Reg	Pre+Reg
Students	2013	2043	426	426	1532	1562
R^2	0.22	0.18	0.15	0.13	0.19	0.14

Note. All covariates are omitted from the table for brevity. Specifications in Panel A include the following pre-test covariates: the within-school pre-literacy and language score rankings, indicators for missing pre-tests, and interactions between all these variables and the treatment indicator. The exception is the indicator for missing the language rank in columns 2 and 3, which is omitted because there were no missing values in the 20th-percentile group. Average treatment effects are given by the coefficient on the Treatment variable. The constant represents the average in the control group for students with non-missing pre-tests and at the mean of the rank variables. Standard errors clustered on the 40 pairs/triples are shown in parentheses below the coefficient. *** $p < 0.01$, ** $p < 0.05$, $p < 0.1$.

8.4 Testing Mechanisms

To examine the extent through which the school-level treatment effects are mediated by the literacy skills of the students in the bottom 20 percentiles of the pre-test literacy distribution, we use a method developed by Kwon and Roth (2025). The method tests the sharp null hypothesis of complete mediation–that is, the treatment effect is fully explained by a proposed set of mediators.

The version of the method we use rests on two main assumptions. First, treatment is independent of potential outcomes. This assumption is likely satisfied by the random assignment to treatment groups in our case. Second, the treatment effects on the mediator obey monotonicity. In our case, the treatment does increase the number of students who score zero on the decoding test in any school. This assumption is difficult to test but seems relatively plausible given the pattern of results we observe. As mentioned in the main text, the assumptions are much weaker than those needed for regular mediator analysis (e.g., Imai et al., 2011; Huber, 2020). We use the R package *TestMechs* to implement the method (available here: https://github.com/jonathandroth/TestMechs).

To get informative results, there are two additional conditions. The method requires that there are some "always-takers" of the mediator. Because we have one-sided non-compliance – no one in the control group received the program – some interesting mediators fail this condition (e.g., being tutored, having a teacher or a reading specialist as tutor). The second condition is that the mediator must have finite support. This latter condition is also needed on the outcome to obtain point estimates. We therefore discretize the variables in the analysis.

We use the method to test whether the treatment effects on high-achieving students are fully mediated by the literacy skills of low-achieving students. High-achieving is operationalized as the top 80 percent and low-achieving as the bottom 20 percent of the literacy pre-test (i.e., we find most of the tutored students in the low-achieving group). We confine the tests to our primary outcome variable, the decoding test. The mediator is the number of students who score zero on the decoding test (which is a count with finite support), and we follow Kwon and Roth (2025) and discretize the outcome variable into five bins, and also use two

and ten bins.[31]

A school-level regression shows that the treatment has a significant negative effect on the number of students scoring zero (-0.69, $p = 0.004$). That is, the treatment affected the mediator. We then drop the low-achieving students from the sample and use only the high-achieving students in the test of the sharp null of full mediation. We can reject the null with two and ten bins ($p = 0.044$ and $p < 0.001$). With five bins, the p-value is above but quite close to 0.05 ($p = 0.065$). Thus, it seems possible that improved literacy skills for the students at the bottom of the pre-test distribution are not the sole channel for the school-level treatment effects.

[31] Simulations in Kwon and Roth (2025) show that the method often over-rejects the sharp null with many bins.