Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 26-1562

Helping the Many by Tutoring the Few: The Social Multiplier of Early Literacy Instruction

Paul Bingley
VIVE - The Danish Center for Social Science Research

Jens Dietrichson,
VIVE - The Danish Center for Social Science Research

Mette Friis-Hansen
VIVE - The Danish Center for Social Science Research

Vibeke Myrup Jensen
VIVE - The Danish Center for Social Science Research

Families frequently receive local signals about children's skills, such as class rank in high school, yet the college market in which most students ultimately compete is much broader. Using nationally-representative data, we show that student effort, parental investment, and beliefs about college degree attainment all respond predominantly to local ranks as opposed to global ranks, despite both strongly predicting college completion. We develop and estimate a dynamic model separating optimistic beliefs, information frictions, and peer interactions. Counterfactual-policy experiments show that more-precise information, provided sufficiently early, raises skill accumulation and college completion. Eliminating optimism has the opposite effects on all outcomes.

VERSION: August 2026

Suggested citation: Bingley, Paul, Jens Dietrichson, Mette Friis-Hansen, and Vibeke Myrup Jensen. (2026). Helping the Many by Tutoring the Few: The Social Multiplier of Early Literacy Instruction. (EdWorkingPaper: 26-1562). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/xnz7-3z53

Helping the Many by Tutoring the Few: The Social Multiplier of Early Literacy Instruction*

Paul Bingley Jens Dietrichson Mette Friis-Hansen
Vibeke Myrup Jensen

August 14, 2026

Abstract

Tutoring is almost always evaluated by randomizing students within schools, a design that differences out any spillover onto classmates. We instead randomize 81 Danish schools and 2,583 kindergartners, tutoring a few children in each and testing all. Decoding improved 0.38 standard deviations among the pre-registered target group and 0.27 SD school-wide. Bounds requiring only the test's range and random assignment show that never-tutored peers gained at least 0.19 SD, implying a social multiplier of at least three. Tutors were the schools' own staff, so costs were low. Evaluations of targeted tutoring understate its value two- to threefold.

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 on 2022-06-22 (protocol number 2022/4) and, for the updated testing procedures, on 2024-03-20 (protocol number 2024/8). Disclosure: no author contributed to the design or development of *Læseklar*, which was developed in Sweden by special educator Anna Aldenius Isaksson. Isaksson is, however, an advisor on this project, and the Danish version of the program was developed with her together with teachers and reading specialists in Hvidovre municipality in a project administered by Dietrichson and Friis-Hansen. In addition, the research team is not independent of the program's prior evaluation record: Dietrichson co-authored the Swedish randomized trial of the program with Isaksson (Bøg et al., 2021), and Bingley, Dietrichson and Jensen co-authored the Danish efficacy trial (Seerup et al., 2025). No author has a financial interest in the program. Isaksson supplied the pricing information used in Section 6. 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 and 0219-00004B) is gratefully acknowledged. All authors are at VIVE–The Danish Center for Social Science Research. The order of authors is alphabetical. 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

Mastering reading is arguably the most critical milestone in early human capital formation. Children who struggle with early decoding are not simply delayed in a single subject; they face a compounding disadvantage because reading is the primary medium through which all later learning is delivered. A central premise in the economics of human development is that the return on such educational investments is highest when children are young, as developmental plasticity declines with age (Carneiro and Heckman, 2003; Knudsen et al., 2006; Heckman, 2013). Despite the importance of early skill formation, roughly one in four 15-year-olds across OECD countries still fails to reach minimum proficiency in reading: on average, 26 percent score below Level 2, the PISA baseline (OECD, 2023). This persistent gap has made targeted tutoring one of the most widely advocated remedies in education policy, and a large experimental literature has been assembled to evaluate it. That literature asks what tutoring does for the child who is tutored. For a school or a ministry deciding whether to adopt a program, however, the relevant question is different: what does a school gain by adopting it, counting every child in the building? These are not the same quantities, and we show that the gap between them is large.

We evaluate *Læseklar* ("Reading Ready"), a structured early literacy tutoring program delivered by existing school staff rather than externally recruited tutors. Using a large cluster-randomized field experiment in 81 Danish schools involving 2,583 kindergarten students, we estimate the effects of school-wide adoption under implementation conditions that closely resemble routine school practice.

Our central claim is that the return to this program at the level of the school is substantially larger than the return a conventional evaluation of it would have found, and that this gap can be bounded rather than assumed. Almost all tutoring experiments randomize tutoring among eligible students *within* a school and compare them with untreated students in the same school (e.g., Nickow et al., 2024; Kraft et al., 2026). If tutoring spills over onto classmates, the control group in such a design is contaminated: its members *are* the peers who benefit. The design does not merely fail to count the spillover; it differences it out. Randomizing schools and testing every kindergartener, as we do, instead recovers what a school gains by adopting the program, including the children who never sat with a tutor.

That difference turns out to be large. Using only the range of the test and random assignment, we bound the gain to non-tutored peers below at 0.19 standard deviations, and show that the implied social multiplier, the ratio of the school-level return to the return a within-school design would have implied, is at least three. A conventional trial of this program would have recovered less than a third of its social return. The evidence base on targeted tutoring is built almost entirely on within-school designs, and if our estimate is representative, that evidence understates the value of these programs by a factor of two to three.

The setting makes this credible, not just a mathematical artifact, and it is worth being precise about what the design can and cannot support. Recent work on scaling asks three questions of any program claiming to be scalable: what is the binding resource constraint, what share of eligible providers will adopt it, and does the effect survive the loss of fidelity that scale produces (Al-Ubaydli and List, 2024; List, 2022; Kraft et al., 2024). Our design speaks to each. The binding constraint in the tutoring literature is the supply of tutors, and this program does not draw on it: tutors are staff the school already employs, so adoption by one school does not reduce another's capacity to adopt. We observe the adoption margin directly: between 10 and 13 percent of invited schools took up the program when it was offered free of charge, and the invited pool was drawn at random from eligible schools in the first round and comprised all of them in the second. We observe fidelity loss and can price it: schools delivered an average of 22 of the recommended 32 to 35 sessions, about two-thirds of the intended dosage, selected students using their own criteria, and the program still raised school-level decoding by 0.27 standard deviations. For comparison, Kraft et al. (2026) report pooled effects of 0.16 to 0.22 standard deviations for tutoring programs at scale, against 0.40 in their full meta-analytic sample.

That tutors are the school's own staff is not incidental to the multiplier. It opens a channel absent from tutoring models that hire externally: staff who spend ten weeks teaching a pedagogy may carry it into whole-class instruction. We present evidence against the school-level effect operating solely through the literacy gains of the tutored children, and show that gains appear even among the highest-achieving students in the class, who were not tutored. That is what such a transfer would predict and what the alternative mechanisms would not.

Not all tutoring models are equally exposed to the frictions that erode effects at scale.[1] Programs that depend on externally supplied tutors, intensive coaching, or specific scheduling are vulnerable to all of them, and a design that works under tight research control may say little about what happens when schools decide for themselves whom to tutor, who tutors them, and when (Kraft et al., 2026). *Læseklar* was designed to teach a narrow set of foundational skills in kindergarten: phonological awareness, letter–sound knowledge, and decoding. Its narrow focus and structured format may make implementation more consistent than broader models, and interventions targeting specific skills need not fade more rapidly than broader programs (Watts et al., 2025). Crucially for what follows, it is delivered by the school's own staff.

Læseklar is not entering our study as an entirely untested idea. The program has pre-

[1]Recent evidence highlights alternative models designed specifically for scalability. Cortes et al. (2024) evaluate an early reading program that embeds specialized part-time tutors directly into classrooms for short instructional bursts. In contrast, secondary school math programs have begun incorporating computer-assisted learning alongside human tutors to ease labor constraints (Guryan et al., 2023).

viously been evaluated in two individually randomized trials. A Swedish trial found large effects on decoding and letter knowledge. In contrast, a Danish trial found positive but smaller, statistically insignificant effects in a setting marked by COVID-related disruption and implementation problems (Bøg et al., 2021; Seerup et al., 2025). These earlier studies reduce reliance on evidence from a single setting, but they do not resolve whether the program remains effective when implemented at scale. Like the present study, neither evaluation tested a "best-case" scenario in which the research team supplied substantial additional personnel. Instead, the central question is whether a school-based tutoring model can work when most implementation decisions remain with the school. Taken together with the two earlier randomized evaluations, our findings provide evidence on which tutoring designs may be more robust to scale-up.

To answer this question, we conducted a large cluster-randomized field experiment. The program consisted of 15-minute one-to-one or small-group sessions, three to four times per week, for 10 to 12 weeks. Treatment schools received only limited additional resources from the research team: program materials, instructional videos, and a follow-up phone call shortly after implementation began. Schools retained substantial discretion over student selection, tutor assignment, and scheduling, and they used it. The experiment therefore asks not whether the program works when delivered as designed, but whether it works when delivered as schools actually deliver it. We find that it does.

Our pre-registered primary specification estimates the effect on students below the 20th percentile of the within-school pre-test distribution, the group the program targets. These students read 2.3 more words on the decoding test than comparable students in control schools, corresponding to 0.38 standard deviations ($p = 0.001$), and gain 0.37 standard deviations in letter knowledge. Comparing all kindergarten students in treatment and control schools, we also estimate positive and statistically significant school-level effects: students in treated schools read, on average, 2.4 more words on the decoding test, corresponding to 0.27 standard deviations, and the effect on letter knowledge is 0.16 standard deviations. That the school-level effect is nearly as large as the effect on the targeted group is the first indication of what follows: the effects cannot be explained by gains among tutored students alone.

Our approach also sidesteps a difficulty that has dogged the economics of peer effects. Estimating how one student's achievement responds to their classmates' is notoriously hard. The reflection problem makes the structural parameter difficult to identify (Manski, 1993), and estimates drawn from natural variation in peer composition are prone to bias (Angrist, 2014). Worse, Carrell et al. (2013) show that such estimates can fail badly at predicting what happens when peer groups are actually manipulated by policy. The magnitude of peer effects in education remains genuinely uncertain (Sacerdote, 2011). We estimate no structural peer-effect parameter and need none. We randomize a *treatment* and measure its reduced-form effect on students who did not receive it, which is the object a policymaker faces and the one that determines whether the program is worth buying. On the channel, Duflo et al. (2011) show that peer composition affects outcomes partly by changing what teachers do rather than only through student-to-student interaction, which is suggestive in a setting where the tutors are teachers and reading specialists employed by the school. The closest antecedent to our design is Berlinski et al. (2023), who exploit a school-randomized remedial literacy program in Colombia that used external tutors and find that non-treated children in treated schools gain 0.11 standard deviations; our lower bound on the peer effect exceeds their point estimate.

Our evidence does not resolve every scaling concern. Participation was voluntary, and the program was free, so the schools we study are those willing to adopt a literacy program at zero marginal cost. The 10 to 13 percent adoption rate bounds how much of the school population this describes. Still, it does not tell us how adoption would respond to a positive price, nor what would happen under compulsory implementation. We also measure outcomes at the end of kindergarten and cannot yet speak to persistence, a point we return to in Section 8. What we can say is narrower and, we think, more useful than a general claim about scalability: a tutoring program that draws on no external labor, that schools implement at roughly two-thirds of intended dosage while choosing their own participants, and that costs little beyond materials, raises reading outcomes for the school as a whole.

The remainder of the paper proceeds as follows. Section 2 describes the institutional setting and the program. Section 3 presents the experimental design, data, outcome measures, attrition, and implementation. Section 4 outlines the empirical strategy. Section 5 reports the main results, including the bounds on peer effects and the social multiplier. Section 6 presents the cost-effectiveness analysis, Section 7 discusses mechanisms, and Section 8 concludes.

2 Background

We implemented the tutoring program in kindergarten classes, 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 the letters of the alphabet. 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., Follmer, 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. At the same time, phonics training develops knowledge of letters and the ability to map letters or letter patterns to sounds. Explicit instruction in these areas appears particularly important for children at risk of reading difficulties: a large literature finds that interventions including phonological awareness, phonics and decoding improve literacy skills among at-risk children (e.g., Snowling and Hulme, 2011).

2.2 The *Læseklar* Program

The program has three main objectives: (1) to develop students' awareness that words have both semantic meaning (e.g., *lion* refers to an animal) and a phonological form consisting of individual sounds; (2) to strengthen knowledge of letter–sound correspondences; and (3) to develop students' ability to decode simple words by translating written letters into spoken sounds.

A distinctive feature of *Læseklar* is its use of play-based, multisensory instruction: figurines, finger-spelling and movement engage visual, auditory, tactile and kinesthetic modalities at once. Instruction is organized into three types of sessions that support progressive skill development, which all feature play and multisensory elements. The first type of session develops phonological awareness using clay figurines representing 24 Danish letter sounds, together with finger-spelling and other multisensory activities.[2] Previously learned letter sounds are reviewed before new ones are introduced. The second type of session is introduced once students have acquired approximately 12–15 letter sounds and focuses on decoding, which is a prerequisite for reading comprehension (Hoover and Gough, 1990). Students practice segmenting simple words into phonemes and linking speech sounds to their written forms, with opportunities for independent or

[2]The letters q, w, x, and z, which are not commonly used in Danish, and c, which can be pronounced in different ways depending on the context, are not included.

parent-supported practice. Early mastery of the alphabetic principle may initiate a self-teaching process that strengthens word representations, vocabulary, and comprehension (Blachman et al., 2014).

The third type of session aims to improve decoding speed and fluency through repeated practice decoding short syllables while progress is monitored over time. The second and third session types are alternated to maintain engagement, while the first is revisited as new letter sounds are introduced. Together, the three session types provide a structured yet flexible framework for developing early decoding skills.

2.2.1 Program Dosage and Organization

The program recommends 15-minute sessions per student or group, conducted three to four times per week over a 10 to 12 week period, for an intended total of 32 to 35 sessions, or 8 to 10 hours of tutoring. It is a targeted program, and deliberately small: schools were recommended to tutor two to three students per class, either one-to-one or in small groups (Bingley et al., 2024). In practice, they tutored 205 of the 1,299 kindergarten students in treatment schools, about one in six. Which students, and how many, was left to the schools. Schools had the flexibility to organize 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. Other schools relied on reading specialists or used a combination of classroom teachers and specialists as tutors (see Online Appendix Table A7).

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. In practice, the research team reached 32 treatment schools by phone; the remainder could not be contacted within the intervention window (Online Appendix Figure A1). Tutors also completed a brief diary documenting the content of each session.

3 Data

We use data collected by the research team (pre-tests, literacy outcomes, diaries) alongside administrative register data (descriptive statistics, covariates, social-emotional skills and well-being outcomes). We first describe sample construction and randomization, then summary statistics, student selection for tutoring, attrition and balance tests, followed by the pre- and post-test measures, implementation, and a discussion about blinding and participation effects.

3.1 Sample Construction and Randomization

For administrative reasons, the trial was conducted in two rounds. We recruited in spring 2022 and 2023, with each respective recruitment year corresponding to treatment start 7-12 months later (in the spring semester of 2023 or 2024). In both rounds, we targeted school units where relatively few parents held a tertiary education, as these were likely to have many at-risk students. In round 1, we used a cutoff of at most 50% with tertiary education and invited 120 randomly selected school units out of 394 eligible. In round 2, in response to a relatively low adoption rate in round 1, we used a cutoff of 70% and invited all 674 eligible school units. We describe the details of the recruitment process in Online Appendix A1.1, including a flowchart indicating attrition at each stage.

A total of 83 school units, nested within 81 schools, accepted the invitation, sent us the required pre-test data, and were subsequently randomly assigned to treatment or control.[3] Units sharing management were treated as one school, even if geographically separated; randomization therefore occurred at the school level. Because two schools each comprise two units, the two counts differ. All subsequent statistics and statements regarding randomization concern schools, the unit we randomized.

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). Bai shows that the optimal design pairs units on the baseline outcome. No baseline measure of decoding is available, because kindergarten students are not expected to be able to decode before reading instruction begins, so the pre-tests do not include a decoding test. We therefore matched on the closest available proxy: schools were ranked by the proportion of students at risk of reading difficulties based on pre-tests of literacy and language skills (Section 3.3), and then matched into pairs and one triple according to adjacent ranks. The 81 schools form 39 pairs and one triple, giving the 40 matched strata on which we cluster. We used Stata 17's random number generator to assign one school per pair to treatment and one school in the triple to treatment, resulting in 40 treatment and 41 control schools.

3.2 Summary Statistics, Attrition, and Balance

At randomization, 1,299 kindergarten students were enrolled in treatment schools, and 1,284 students were 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 2,013 students, with an overall attrition rate of 22 percent, which differs by arm (81 percent response in treatment schools, 75 percent in control schools). Of 205

[3]There were 16 units randomized in round 1 and 67 units in round 2, indicating an adoption rate of invited school units of 13% and 10%, respectively.

students who began tutoring, 180 had a decoding test (12% attrition). In five schools, three treatment and two control schools, no student had a decoding post-test, whether because the school left the study after randomization or because its students could not be tested.

Table 1: Observations by Treatment and Control Schools

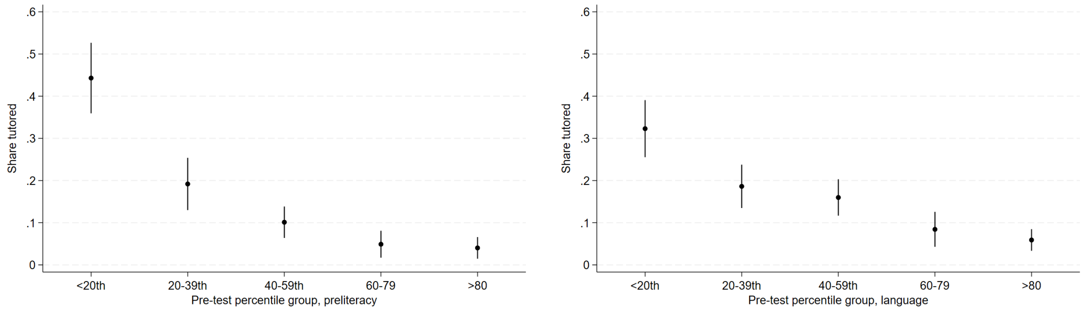
	Tutored Students		**Treatment Schools**			**Control Schools**		
	# Students	%	# Schools	# Students	%	# Schools	# Students	%
Pre-test	196	96	40	1250	96	41	1257	98
Post-test (decoding)	180	88	37	1055	81	39	958	75
Pre- and post-test	173	84	37	1024	79	39	943	72
At risk	145	71		455	35		523	41
Total	205	100	40	1299	100	41	1284	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 (labeled Total). The pre-test row displays the number of students and schools with an observation for at least one of our pre-test based covariates (At risk language, At risk literacy, Language rank, and Literacy rank); 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 a decoding outcome and at least one pre-test based covariate. The At risk-row indicates the students at risk of reading difficulties according to either the language or preliteracy pre-test or both.

Table 2 presents summary statistics and balance tests for students with a primary outcome. The table includes register data on mother and child characteristics, pre-test variables, and the two post-test literacy 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.

Tutored students are clearly more at risk than the average student. Table 2 shows they more often come from ethnic minority families and have mothers who are less likely to be employed, have lower income, and have lower levels of education. Furthermore, tutored students are more frequently identified as being at risk of reading and language difficulties based on the pre-tests. 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, both 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, suggest schools used additional information beyond the pre-tests to select tutored students. 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.

There are three main threats to balance: improper randomization, chance bias, and attrition. Online Appendix Table A1 reports school-level statistics based on register



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 (left panel) and language pre-tests (right panel). The shares are estimated using the sample in treatment schools without missing observations on the respective pre-test.

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

data with complete information at randomization (i.e., before attrition). We find no evidence of randomization failure or large chance bias: differences are small, and only one is statistically significant on the 5-percent level (henceforth, "significant" refers solely to statistical significance).

Table 2, in which attrition at both the student and school-level may cause imbalances, indicates no problematic imbalances between the treatment and control school students. Aside from the outcome variables, most differences between the students in the treatment and control schools are small. One difference is significant at the 5-percent level: the treatment group has more college-educated mothers.[4] 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.[5] The p-value is 0.285.[6] Thus, we cannot reject that the treatment and control groups are balanced, even after attrition.

[4]We use a regression including only an intercept and the treatment indicator with standard errors clustered at the pair/triple level. Three covariates are significant at the 10-percent level: mother is outside the labor force, mother is missing from the registers, and the indicator for being at risk of literacy difficulties at pre-test.

[5]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 the mean (continuous) or 0 (binary), with indicators included in regressions.

[6]The p-value is 0.701 if we instead use father characteristics. We cannot include all variables in one regression because the number of pair/triples imply too few degrees of freedom to calculate an F-statistic.

Table 2: Summary Statistics and Balance Tests

Variable	(1) Tutored Mean (SD)	(2) Treatment Mean (SD)	(3) Control Mean (SD)	(4) Diff (SE)	(5) N
Register data					
Mother Danish	0.66 (0.48)	0.71 (0.46)	0.74 (0.44)	-0.030 (0.047)	2013
Disposable income q1	0.41 (0.49)	0.33 (0.47)	0.37 (0.48)	-0.045 (0.034)	2013
Disposable income q2	0.32 (0.47)	0.32 (0.47)	0.34 (0.47)	-0.016 (0.026)	2013
Disposable income q3	0.18 (0.39)	0.20 (0.40)	0.18 (0.38)	0.026 (0.022)	2013
Disposable income q4	0.05 (0.22)	0.10 (0.30)	0.08 (0.27)	0.021 (0.018)	2013
Disposable income q5	0.01 (0.07)	0.03 (0.18)	0.01 (0.12)	0.019 (0.012)	2013
Single	0.18 (0.39)	0.18 (0.39)	0.19 (0.39)	-0.009 (0.028)	2013
Basic education	0.32 (0.47)	0.23 (0.42)	0.24 (0.43)	-0.010 (0.028)	2013
College education	0.12 (0.32)	0.19 (0.39)	0.14 (0.35)	0.049 (0.020)	2013
Employed	0.63 (0.48)	0.69 (0.46)	0.65 (0.48)	0.043 (0.028)	2013
Unemployed	0.04 (0.21)	0.04 (0.19)	0.03 (0.17)	0.006 (0.008)	2013
Outside labor force	0.28 (0.45)	0.25 (0.43)	0.29 (0.46)	-0.045 (0.027)	2013
Parents live together	0.66 (0.48)	0.70 (0.46)	0.67 (0.47)	0.031 (0.028)	2013
Child Danish	0.61 (0.49)	0.66 (0.48)	0.70 (0.46)	-0.038 (0.048)	2013
Boy	0.52 (0.50)	0.50 (0.50)	0.48 (0.50)	0.026 (0.026)	2013
Age in years	7.03 (0.38)	6.99 (0.34)	6.96 (0.34)	0.026 (0.018)	1985
Missing, any register	0.05 (0.22)	0.03 (0.15)	0.01 (0.10)	0.014 (0.007)	2013
Pre-tests					
At risk language	0.48 (0.50)	0.25 (0.43)	0.28 (0.45)	-0.035 (0.033)	1929
At risk literacy	0.60 (0.49)	0.22 (0.41)	0.27 (0.44)	-0.048 (0.027)	1957
Language rank	12.74 (13.19)	20.69 (15.96)	20.76 (17.78)	-0.074 (2.780)	1965
Literacy rank	9.77 (9.86)	20.52 (15.81)	21.04 (17.78)	-0.518 (2.739)	1920
Post-tests					
Decoding	6.16 (6.81)	9.47 (9.09)	7.12 (8.52)	2.358 (0.786)	2013
Letter knowledge	24.82 (4.02)	26.25 (3.13)	25.65 (3.73)	0.598 (0.220)	2013

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, conditional on having a decoding outcome. Parentheses in columns (1) to (3) contain standard deviations. Column (4) reports the difference in means between treatment and control school students, estimated from a regression with an intercept and the treatment indicator, with standard errors clustered at the pair/triple level in parentheses.

However, while we cannot reject balance, attrition is the most serious threat to our estimates, and because it is differential, it warrants extra attention. Attrition arises at two levels. Of the 244 treatment-school and 326 control-school students without a decoding post-test, 90 are in the five schools where no student has a decoding post-test, 62 in treatment schools and 28 in control schools. The other 480 are students not tested within schools that otherwise participated, 182 in treatment schools and 298 in control schools. The differential response rate is therefore a within-school phenomenon: school-level losses are, if anything, slightly more common in treatment schools, whereas the excess attrition in control schools comes from students not tested within schools that otherwise participated. School-level departures came from management changes and staff turnover, plausibly unrelated to outcomes, whereas selective non-testing within surviving schools would be harder to dismiss.

We first test whether the difference in response rates is statistically significant, regressing an indicator for having a decoding post-test on the treatment indicator, with standard errors clustered at the pair/triple level. The treatment coefficient is 0.066 (standard error 0.056, $p = 0.25$), so the higher response in treatment schools is not statistically distinguishable from the control rate once clustering is accounted for. We then ask whether the students who attrit differ between arms, regressing each pre-randomization characteristic on the treatment indicator within the sample of attritors and conducting a randomization-inference omnibus test of joint orthogonality among attritors, mirroring the test reported above for the analysis sample (Kerwin et al., 2024). None of these characteristics differ between treatment and control attritors at the 5 percent level. Two differ at the 10 percent level: treatment attritors have a lower within-school rank on the preliteracy (-8.8, $p = 0.10$) and the language pre-test (-9.56, $p = 0.06$). The omnibus test does not reject joint orthogonality among attritors ($p = 0.48$), so selection on observable characteristics is unlikely to explain the estimated effects.

Observable balance among attritors is reassuring but not decisive, since selection may operate on unobservables. We therefore report bounds on the treatment effect following Lee (2009), trimming the higher-responding treatment arm so that the response rates match across arms. The lower bound on the decoding effect is positive, both unconditionally and when tightened within pre-test strata, so the estimated effect does not reverse under the worst case consistent with our attrition. We report the full bounds, together with the confidence intervals around them, in Online Appendix A1.3, and we regard these bounds rather than the balance tests as the binding evidence on this question.

3.3 Pre- and Post-Tests

As a pre-test, we primarily used the Danish national kindergarten reading test, developed by Danish and international researchers and inspired by several language proficiency tests, including the MacArthur-Bates Communicative Development Inventory (Bleses et al., 2017). The test assesses vocabulary, language comprehension, rhyming, and letter knowledge, with age- and gender-standardized scores based on a large normative sample.

Most schools administered this test, although any norm-referenced test was accepted. Because scores are not comparable across tests, we use within-school ranks and test-based at-risk indicators rather than raw scores as covariates. Missing ranks are imputed to the mean, and missing at-risk indicators are set to zero.

Our primary outcome is decoding, measured using the 30 real words of *Elbros ordlister*, a subtest of the standardized Danish dyslexia assessment *Ordblinderisikotesten*, since renamed *Risikotest for ordblindhed* (Gellert and Elbro, 2016). The outcome is the number of correctly read words (0–30). The test takes five to eight minutes and has a test-retest reliability of 0.78 (Gellert and Elbro, 2016). It assesses the same skill as the Sight Word Efficiency subtest of the Test of Word Reading Efficiency (Torgesen et al., 1999), differing mainly in that it is scored for accuracy rather than under a speed limit.[7]

As a secondary literacy outcome, we use a standardized letter knowledge test (Borstrøm and Klint Petersen, 2004). The score equals the number of correctly named Danish letters (0–29), with reported reliability of $\alpha = 0.88$ (Petersen, 2005). The project administered post-tests at the end of kindergarten. In treatment schools, testing occurred on average 23 days after the final tutoring session.[8] Blinding testers to treatment status was the standard testing procedure (Torgerson et al., 2011). We describe the testers, the cases where blinding could not be maintained, and why we expect little consequence for our estimates, in Section 3.5.

As secondary outcomes, we use three subscales from the Danish National Wellbeing Survey measuring distress, school satisfaction, and self-efficacy/self-confidence (Damm et al., 2021). Because not all treatment schools had begun tutoring when the survey was administered, and survey timing partly depended on school decisions, these outcomes are analyzed only as sensitivity analyses.

As noted in our protocol, we also collected measures of executive functions for a subset of second-round schools. Because these data cover a different sample, they are analyzed separately and reported in another paper.

3.4 Implementation

School-level compliance with treatment assignment was perfect: no control school implemented any part of the program, and all treatment schools provided tutoring. However, adherence to the program manual varied. To approximate real-world implementation, schools were given considerable flexibility. During the intervention, tutors kept brief diaries documenting each student's progress, providing information on treatment dosage and implementation fidelity (Table 3).

The average session lasted 15 minutes, and the intervention ran for 12 weeks on average, both consistent with the program manual. Schools delivered two sessions per week on average, however, below the recommended three to four. Consequently, students received an average of 22 sessions rather than the recommended 32 to 35, corresponding to 5.7 hours of tutoring rather than the recommended 8 to 10. Overall, 91% of students progressed to session type 2 and 69% to session type 3.

[7]Only 26 students (1.3%) reached the maximum score. Seerup et al. (2025) used the same test in the efficacy trial of Læseklar, while the Swedish trial used a closely related measure (Elwér et al., 2016).
[8]Based on tutor diaries from 133 tutored students.

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	91%		192
% Children pass step 2	100	69%		192
Taught by teacher		44%		151
Taught by reading specialist		48%		151
Taught by both		7%		151

Note. The 'Recommendation' and 'Actual' columns report the recommended and actual treatment dosage, respectively. 'SD' reports the standard deviation of the continuous dosage measures. 'N' denotes the number of treated students for whom information is available for the corresponding variable. Data are obtained from either teacher log reports (diaries) or teacher interviews.

Interviews indicate that 44% of students were tutored by their classroom teacher, 48% by a reading specialist, and 7% by both. They also suggest that tutoring was conducted during school hours at varying times of the day, causing students to miss different activities, including literacy, math, play, and joint sessions with other classes.

3.5 Blinding and Participation Effects

Neither schools nor students can be blinded to treatment status. Schools must know they are implementing the program to implement it, and parents must be informed to consent. As we noted in the pre-analysis plan (Bingley et al., 2024), this leaves three threats that random assignment does not address, and we discuss each in turn.

Tester blinding. Our primary outcome was an individually administered oral reading test. During the assessment, a tester sat one-on-one with each child and recorded the number of words the child read correctly. The assessments were mainly administered by college and university students.[9] Some testers were hired specifically for the project, while others were regular student assistants employed by VIVE on longer-term contracts: they had no role in delivering the program and no stake in the results, and blinding them to treatment status was the standard testing procedure. However, a small number of schools may have disclosed treatment status to the testers.[10] Three features of the setting limit

[9]One retired teacher also conducted assessments. For students hired by the project, we focused on students from teacher colleges for the reading test.

[10]Schools were instructed not to disclose their treatment status, but a few delivered the diaries to a

the scope for this to bias our estimates. First, the testers had no incentive to score treated and untreated children differently. Second, the outcome is scored dichotomously, word by word, rather than rated on a scale, which leaves little discretion. Third, the effects on older cohorts reported in Section 5.4 come from national reading tests, which are computer-administered and recorded in administrative registers, not by project testers; these are positive, though not statistically significant for the average score.

Hawthorne effects. Staff in treatment schools knew they were participating in an evaluation of a program they were delivering, and may have devoted more attention to literacy instruction for that reason alone. We cannot rule this out. It is worth noting, however, that the mechanisms we discuss in Section 7 include precisely this channel: if being observed causes school staff to teach literacy more attentively, that is part of what a school buys when it adopts the program, and it is not obvious that a policymaker should want it netted out. What a Hawthorne effect would undermine is the claim that the *program* produced the gains, not the claim that adoption raised reading scores.

John Henry effects and control-school behavior. Control schools knew they had been assigned to control. They received compensation of DKK 1,000 per class for participating in the post-tests, and were offered the program for free six months after post-testing was completed.[11] Knowing they had been denied a program, control schools might have increased their own literacy efforts, which would bias our estimates toward zero, or adopted a different program, which would mean we are not comparing *Læseklar* against treatment as usual. Our pre-analysis plan anticipated this and specified that we would survey control schools about their instructional methods. While small-group instruction (group size below 7-8 students) was used in some form by around half the control schools, no school mentioned using literacy materials that were connected to a specific tutoring program or method.

Between-school spillovers of the program itself are unlikely: *Læseklar* was not available on the Danish market during the study period, schools that received materials were not permitted to share them, and the program's physical materials were difficult to copy.

4 Empirical Framework

We pre-registered a protocol including primary and secondary outcomes, in the American Economic Association's Social Science Registry (AEARCTR-0011007; Bingley et al., 2024), and note deviations below. Our pre-registered primary specification focused on

tester, which revealed that the school was a treatment school. We do not have information on when the diaries were delivered and therefore do not know the exact extent of the problem (i.e., if diaries were delivered after the testing took place, it would not be problematic concerning the blinding of testers).

[11] We exclude both from the cost calculations in Section 4.5, as they are research expenses rather than costs a school adopting the program would bear.

students at or below the 20th percentile of the within-school pre-test distribution. As we did not specify a full testing hierarchy, our secondary hypothesis tests are subject to multiple hypothesis testing problems, which we discuss in Section 4.3. The instrumental variables (IV) decomposition of effects on tutored students and non-tutored peers (Section A1.7.3) was pre-specified, including the construction of the two instruments by interacting treatment assignment with indicators for being below or above the 20th percentile.[12] The bounds analysis that precedes it (Section 5.3.1) was not; we added this lower-assumption approach as a complement to the IV estimates. The mediation test in Online Appendix A1.10 (Kwon and Roth, 2026) was also not pre-specified, as the study was not originally designed to test mechanisms.

Our analysis of secondary well-being outcomes (Section 5.4) deviates from the protocol in two respects. Because tutoring start dates and survey timing were partly determined by school decisions rather than by the research team and some schools started tutoring after the survey, we treat these outcomes as sensitivity analyses. We also supplement the pre-registered intention-to-treat (ITT) specification with a non-pre-specified treatment-on-the-treated (TOT)/IV estimate that uses the treatment indicator as an instrument for having started tutoring before the survey. Executive function measures, though part of the pre-registered protocol for a subset of second-round schools, are analyzed in a companion paper rather than here, since they cover a different sample. Finally, the analysis of non-targeted older cohorts (Table 6) is not among the pre-registered hypothesis tests; we conducted it to contextualize the program's broader effects. The protocol pre-specifies a full-ingredients cost calculation, valuing personnel, school facilities, and materials (see Section 2.5 of Bingley et al., 2024). We deviate from it and report an incremental cost as our primary figure because only the incremental cost is comparable to the benchmark studies against which we assess cost-effectiveness (Section 6), which cost interventions the same way. We report the full-ingredients cost as a conservative upper bound.

We estimate program effects using a cluster-randomized controlled trial (RCT) in which schools are randomly assigned to treatment or control. Simultaneous assignment mitigates time trends and regression to the mean (Torgerson et al., 2011), while school-level randomization reduces the risk of crossover and spillovers. Spillovers between schools are particularly unlikely because the program was not commonly available in Denmark and participating schools were prohibited from sharing materials (Seerup et al., 2025).

Randomization ensures that treatment is independent of observed and unobserved school and student characteristics in expectation. To improve statistical precision and estimate effects for different student groups, we estimate average treatment effects (ATE)

[12] The protocol specified two further refinements to the IV analysis, the recentering of Borusyak and Hull (2023) and the unbiased estimator of Andrews and Armstrong (2017), which we do not implement; Online Appendix A1.7 explains why neither bears on our conclusions.

using student-level data and covariate adjustment:

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

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 ITT estimate of the effect.

Following de Chaisemartin and Ramirez-Cuellar (2024), we cluster standard errors at the pair/triple level. de Chaisemartin and Ramirez-Cuellar show that in paired designs, clustering at the unit of randomization while including pair fixed effects yields a variance estimator that is downward biased by roughly one half, so that a nominal 5 percent t-test rejects a true null up to 16.5 percent of the time. Clustering at the pair or strata level corrects this, and their simulations show that it attains nominal size whether or not strata fixed effects are included in the regression. Our primary specification omits the pair/triple indicators, and we report the specification including them in Online Appendix Table A2; the two give similar estimates. Their simulations further indicate that the pair-clustered t-test is well approximated by a normal distribution with as few as 20 pairs, and becomes liberal below that. Our 40 pair/triples 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 for those above the 20th percentile. Defining samples using pre-test scores avoids bias from potentially endogenous selection into tutoring within treatment schools.

To contextualize our results, we report effect sizes. Accounting for clustering within schools, we estimate the standard deviation from an intercept-only multilevel model and calculate Hedges' g (Hedges, 1981).[14]

[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).

[14] The effect size is the treatment effect divided by the estimated SD. 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 examine treatment effects across the outcome distribution, we estimate distribution regressions for all students and separately for those below and above the 20th percentile of the pre-test distribution; the resulting estimates are reported in Online Appendix A1.5. Specifically, we estimate a series of linear probability models (LPMs) with binary outcomes indicating whether a student's score exceeds each possible threshold up to the maximum.[15]

We also examine treatment heterogeneity by interacting the treatment indicator with five groups based on the pre-literacy percentiles. We use the pre-literacy test because it is strongly correlated with the overall pre-test score ($\rho = 0.67$) and better predicts post-test outcomes than the language test.

4.2 Effects on Tutored Students and Non-Tutored Peers

The school-level effect is an average over two groups: the students a school chose to tutor, and their classmates. Because schools, not students, were randomized, and because selection into tutoring within treatment schools was not random, how the effect splits between tutored students and non-tutored peers is not directly observed. We recover it in two ways, and we are explicit about which one carries the argument.

Our primary approach requires almost nothing beyond random assignment. Let E denote the school-level effect on decoding, s the share of students who receive tutoring, τ_T the effect on tutored students and τ_P the effect on their non-tutored peers, each relative to comparable students in control schools. By construction,

$$E \;=\; s\,\tau_T \;+\; (1-s)\,\tau_P. \tag{2}$$

The decoding test is bounded between 0 and 30 words, which bounds τ_T from above by the tutored students' mean post-test score. Combined with (2), this bounds τ_P from below without any further assumption. Bounding τ_P from above requires a restriction on how negative τ_T could be, and we report two, following Kang and Keele (2018): that no student is harmed by treatment, and, more conservatively, that tutored students' effect equals the estimate for the lowest pre-test percentile group, who make up most of them.[16] Equation (2) also assumes that the share who would have been tutored is the same in

[15] See Freedman (2008b) for arguments against logit or probit in RCTs. We omit covariates to avoid predictions outside [0,1] and because primary results are similar with and without covariate adjustment. 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.

[16] Kang and Keele (2018)'s TOT approach yields similar bounds, but assumes that potential outcomes depend only on the *number* of treated peers rather than their identity, which is a strong assumption here.

treatment and control schools; because that share is not observed in control schools, we report in Section 5.3.1 how the bounds vary with it.

The same identity delivers the paper's headline quantity without any additional machinery. A within-school randomized design, which is how tutoring is almost always evaluated, has a control group drawn from the same classrooms and therefore contaminated by spillovers; it estimates $\tau_T - \tau_P$ rather than τ_T. The ratio of the true social return to the return such a design would imply is the social multiplier (M; Glaeser et al., 2003), and rearranging (2) gives $s(\tau_T - \tau_P) = E - \tau_P$, so that $M = E/(E - \tau_P)$. The effect on tutored students cancels. The multiplier can therefore be bounded using only the school-level effect and the bounds on the peer effect. We derive and report it in Section 5.3.2.

We also report the IV decomposition that was pre-specified in our protocol (Bingley et al., 2024), which instruments tutoring status and peer exposure using interactions of the treatment indicator with indicators for being above or below the 20th percentile of the pre-test distribution. It yields point estimates rather than bounds, but requires assumptions we cannot defend as exactly true, including a no-cross-effects condition that our setting violates by construction (Bhuller and Sigstad, 2024). The estimating equations, the assumptions, and the estimates are set out in Online Appendix A1.7. We treat the bounds as the evidence and the IV as a consistency check.

Throughout, we focus on decoding, since it is our primary outcome and the letter knowledge test exhibits ceiling effects.

4.3 Testing Multiple Hypotheses

Our empirical approach tests several hypotheses. Multiple testing can inflate the family-wise error rate, the probability of rejecting at least one true null beyond the pre-specified significance level (Romano and Wolf, 2005), and the false discovery rate, the expected share of rejected nulls that are actually true (Benjamini and Hochberg, 1995).

Our response is a design choice made before the data were seen rather than a correction applied after. Following the pre-analysis plan, we designated a single primary outcome, the decoding test, and a single primary specification, students at or below the 20th percentile of the within-school pre-test distribution (Bingley et al., 2024). Every other estimate we report (letter knowledge, the full sample, students above the 20th percentile, the well-being scales, the older cohorts, and the distribution regressions) is secondary or exploratory, and we label it as such. This is the pre-registered strategy for controlling error rates, and it is the one on which our confirmatory claim rests: the effect on decoding among the pre-registered target group is 2.30 words ($p = 0.001$).

The estimates we report as secondary should be read accordingly. For readers who prefer a correction rather than a designation, the school-level effect on decoding survives a Bonferroni adjustment across the three samples we report (dividing the 5 percent level by three; Dunn, 1961).

This designation has an implication we should be explicit about, because it cuts against us. The social multiplier we report in Section 5.3.2 is constructed from the full-sample school-level effect, and the full sample was not our pre-registered confirmatory specification. While the *question* of spillovers was pre-registered, the combination of the full-sample estimate and the bounds, and the multiplier derived from them, were not. We therefore report the multiplier as what it is: an analysis the design permits and the data support, but one specified after the fact. The confirmatory claim of this paper remains the 2.30-word effect on the pre-registered target group. Readers who weight pre-registration heavily should treat the multiplier as exploratory, and we would rather say so than have it inferred.

4.4 Statistical Power

The pre-analysis plan reported minimum detectable effects (MDEs) of 0.25 standard deviations for the primary analysis, students at or below the 20th percentile, and 0.20 standard deviations for students above it, assuming 80 schools, an intra-cluster correlation of 0.10, and a baseline correlation between pre- and post-tests of 0.4 (Bingley et al., 2024). Our estimates are 0.38 and 0.26 standard deviations respectively, so both exceed the effect the study was designed to detect. The margin is comfortable for the primary analysis and narrower for the higher-achieving students, which is the group whose estimate carries the spillover argument. We note this rather than resting on it: the bounds in Section 5.3.1 do not depend on the precision of that particular estimate.

4.5 Cost-Effectiveness

We define cost-effectiveness as the cost per unit increase in effect size (CES) per student, calculated at the school level (Hollands et al., 2013). Because the program's goal is to improve students' decoding skills, our cost-effectiveness analysis focuses on this outcome and its school-level treatment effect. The effect size is expressed as Hedges' g, computed as described in Section 4.

We use the ingredients method to estimate costs (Hollands et al., 2013; Levin and Belfield, 2015). This approach incorporates the opportunity cost of all resources needed to run the intervention. Focusing on school-level treatment effects, we calculate implementation costs at the school level and exclude program development.

The central costing question in our setting is how to value tutor time. Because tutors were existing school personnel rather than external hires, tutoring did not raise school payrolls. It did, however, redirect staff time away from other activities. We therefore report a *marginal-outlay* cost as our primary estimate, which is the incremental cost in the sense of the ingredients literature (Levin and Belfield, 2015), alongside a full-ingredients figure, which values tutor session time at a teacher wage, that we retain as a conservative upper bound and report in Online Appendix A1.9.

The *marginal outlay* to an adopting school is the resources it would have to purchase that it does not already own. These are the program materials and the research team's time for follow-up calls.[17] Tutor time enters at zero, because it is reallocated rather than added. This is the relevant figure for a school or municipality asking what the program would add to its budget.

Valuing reallocated tutor time at zero is a strong assumption, but our design bears on it directly. Because randomization is at the school level and we test every kindergarten student in participating classes, the school-level treatment effect is already *net of* whatever displacement tutoring caused within these classes. If pulling students out of literacy, mathematics, or free play harmed the tutored students, or their non-tutored classmates, that harm is already subtracted from the effect in our numerator. A trial that randomizes individual students within schools cannot make this claim; we can.

What the school-level effect does not internalize is threefold. First, effects on other grades: to assess these, we use national reading tests in grades 2, 3, 4, 6, and 8, obtained for 76 treatment and control schools from Danish administrative registers. The youngest of these students were in grade 1 during our study's implementation. Second, effects on kindergarten classes in treated schools that did not participate in the study, which we can address once national second-grade reading test data become available for our study cohorts. Third, outcomes within the kindergarten cohort that we do not measure, chiefly mathematics.

Literacy tutoring may also spill over to other skills. Previous studies find positive or null spillovers 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., 2020b; Lindström-Sandahl, 2024). We can partly assess effects on social-emotional skills and well-being using a national well-being survey, but not all schools had begun tutoring when it was conducted. Given the prior literature, excluding other skills likely leads us to underestimate, rather than overestimate, cost-effectiveness.

5 Results

This section presents our findings. We first report ITT estimates of program impacts based on school-level random assignment. We then examine heterogeneity across outcome and pre-test distributions and analyze effects on both tutored students and their non-tutored peers.

[17] We reimbursed control schools for participating in post-intervention tests, but exclude this research-only expense from the cost calculation.

5.1 Intention to Treat Estimates of the Program Effect

Table 4 reports ITT estimates from our main analysis. Odd-numbered columns use the decoding test as the outcome; even-numbered columns use the letter knowledge test. Columns 1–2 use the full sample; columns 3–4 restrict to students at or below the 20th percentile of the joint pre-test distribution; and columns 5–6 to 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	2.44	0.55	2.30	1.70	2.42	0.31
	(0.71)	(0.20)	(0.68)	(0.49)	(0.78)	(0.17)
Constant	5.99	26.12	3.77	24.70	6.43	26.35
	(0.82)	(0.29)	(0.83)	(0.72)	(0.85)	(0.28)
Covariate set	Pretest	Pretest	Pretest	Pretest	Pretest	Pretest
Students	2013	2043	425	425	1588	1618
R^2	0.19	0.17	0.12	0.12	0.14	0.12
g (in sample)	0.27	0.16	0.38	0.37	0.26	0.11
Randomization p	0.004	0.007	0.001	0.002	0.004	0.088

Note. All specifications include the following pre-test covariates: within-school pre-literacy and language score rankings, indicators for literacy and language at-risk status, and indicators for missing observations. Exceptions are the indicator for missing the language rank in columns 3–5 and the indicator for missing literacy at-risk status in column 6, which are omitted because there are few missing cases. Covariates are not reported in the table for brevity. Average treatment effects are the coefficients on the Treatment variable. The constant is the mean for students at the average of the rank variables with non-missing pre-tests, and who were not at risk of language or literacy difficulties according to the pre-tests. Standard errors, clustered on the 40 pairs/triples, are in parentheses. Effect sizes (Hedges' g) are the effect estimate divided by the estimated SD for the relevant sample (the in-sample SD for each outcome/sample). We estimate the SD using an intercept-only multilevel model to account for clustering of students within schools.

The estimates show a positive, significant effect on decoding skills, our primary outcome.[18] Students in treated schools read, on average, 2.4 more words than those in control schools ($g = 0.27$). The effect on letter knowledge is also positive and significant ($g = 0.16$). The constant, which is the control group mean for students at the mean of the rank variables and with non-missing pre-tests, indicates many children score at or near the maximum, suggesting substantial ceiling effects on the letter knowledge test.

Decoding effects are positive and significant for both lower- and higher-achieving students. Although point estimates in words are similar (2.3 vs. 2.4), effect sizes based

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

on the in-sample SD are larger for lower-achieving students, who more often received tutoring ($g = 0.38$ vs. 0.26). Effects on letter knowledge are also larger for lower-achieving students ($g = 0.37$ vs. 0.11) and not significant for higher-achieving students. This likely reflects stronger ceiling effects among higher-achieving students, though the constant indicates ceiling effects in both groups.

Given the small share of tutored students and our ITT focus, the effect sizes at the bottom of Table 4 are substantial. Relative to the control mean, they correspond to 34% for the full sample and 82% for low-achieving students on decoding.

The school-level effect is also large by the standard most relevant to scaling. Kraft et al. (2026) report meta-analytic estimates for a policy-relevant target of inference, large-scale tutoring programs evaluated on standardized tests, ranging from 0.16 to 0.22 standard deviations, against 0.40 in their full sample. Our school-level estimate of 0.27 standard deviations exceeds that at-scale benchmark, and it does so despite being an average over all students in treated schools, roughly 83 percent of whom were never tutored, and despite schools delivering only about two-thirds of the recommended dosage. Comparisons across tests and settings should be made cautiously. But the contrast is informative, because the attenuation Kraft et al. document is driven by precisely the features our design was built to avoid: reliance on externally supplied tutors and the organizational strain of delivering a program at scale.

The results are robust to alternative inference procedures and specifications. Results are robust to using randomization test-based p-values: in the last row of Table 4, all randomization p-values are below 0.01, except, consistent with the primary analysis, for letter knowledge among higher-achieving students. As shown in Table 2, school-level treatment effects for decoding and letter knowledge closely match the primary estimates in Table 4 when we exclude covariates and directly compare means. In Online Appendix A1.4, we report several alternative specifications: following Lin (2013), we interact the treatment indicator with covariates,[19] include fixed effects for randomization strata, use only register-based covariates with minimal missingness, and include both pre-test and register-based covariates. Across these models, the sign and significance of the estimates are unchanged, and point estimates remain similar, sometimes slightly larger and sometimes slightly smaller than in the primary analysis.

5.2 Heterogeneity Across the Pre- and Post-Test Distribution

The school-level effect could arise in two very different ways, and distinguishing them matters for what follows. The program might lift the weakest readers off the floor, moving children from zero words to a few, in which case the gain is real but narrow. Or

[19]Because our randomization created treatment and control groups of unequal size, including covariates can introduce bias (Freedman, 2008a). Lin (2013) shows that interacting the treatment variable with all covariates largely removes this bias.

it might shift the whole distribution, including students who were never candidates for tutoring. We find the second, and the pattern differs sharply between the students the program targeted and everyone else.

Distribution regressions, reported in Online Appendix A1.5, characterize where in the distribution these effects fall, and two patterns matter for what follows. Among students at or below the 20th percentile, the group from which most tutees were drawn, the gains are concentrated at low thresholds, the signature of children being lifted off the floor. Among students above the 20th percentile, few of whom were tutored, the gains instead fall across the middle and upper range of the decoding distribution. Because tutored students were rare in the upper part of the distribution (Figure 1), effects there are hard to reconcile with an effect operating only through tutored students, and point instead to gains among non-tutored peers.

Heterogeneity by pre-test quintile (Online Appendix Table A3) tells the same story and sharpens it. Decoding effects are statistically significant below the 60th percentile and, notably, in the top quintile (2.24 words). The top-quintile effect is the one worth dwelling on, and it invites an obvious objection: a small number of tutored students rank high on the pre-tests (Figure 1), so the estimate might simply be picking them up.

It is not. On the pre-registered sample of students above the 20th percentile, the group the protocol designated as peers, we add an indicator for tutored students, so that the treatment coefficient identifies the effect on higher-ranked students who were *not* tutored (Online Appendix Table A4). This sample contains enough tutored students, about 78, to identify the indicator, unlike the top quintile alone, where fewer than one in twenty students received tutoring (Figure 1).

Removing the tutored students leaves the effect intact. The treatment effect on non-tutored students above the 20th percentile is 2.49 words (standard error 0.82), strongly significant and consistent with the bounds we place on the peer effect in Section 5.3.1. At the same time, the coefficient on the tutored indicator is small and statistically insignificant (-0.66). Letter knowledge shows the same pattern: a small treatment effect on non-tutored students (0.36, standard error 0.18) and an insignificant tutored indicator. The gains among higher-achieving students are therefore gains among students who were not tutored. Whatever transmits the program's benefit to them is not their own tutoring, and is unlikely to be a one-student reduction in class size. We return to this point in Section 7.

5.3 Effects on Tutored Students and Non-Tutored Peers

The school-level effects reported above are informative for policy on their own. Still, a natural question follows: are we simply seeing gains among the small number of tutored children, or does tutoring change outcomes for the classroom as a whole? This matters for cost-effectiveness, because a program that only helps the students who receive it is a very different investment than one that raises achievement school-wide. Our answer rests on arithmetic rather than on modeling assumptions. We first show that tutored students alone cannot account for the total effect we estimate, and then bound the split between tutored students and their non-tutored peers using only the range of the test and random assignment. These bounds are our primary evidence on the question, and we show that they identify the paper's headline quantity, the social multiplier, without any further assumptions. The pre-specified IV decomposition delivers point estimates for each group; because it requires assumptions we cannot defend as exactly true, we report it in Online Appendix A1.7 and treat it as a consistency check. Throughout, we focus on our primary outcome, decoding.

5.3.1 Existence of Effects on Non-Tutored Peers and Bounds

The school-level effect we estimate is 2.44 words (Table 4, column 1). Could this be driven entirely by the 180 tutored students, with no effect on their 875 non-tutored classmates? The arithmetic says no. If the school-level effect reflects only gains among tutored students, and each student is weighted equally, simple back-of-envelope math, dividing the school-level effect by the tutored share of the sample, implies those students would need to read 14.30 *more* words than control-group students, on average.[20] That is larger than the average tutored student's entire post-test score (6.16 words), so it is not a plausible outcome. Some of the total effect must therefore come from gains among non-tutored peers.

To quantify how much, we use the fact that scores are bounded between 0 and 30 words, together with one further assumption: that the share of students who would have been tutored is the same in treatment and control schools (0.171, based on the treatment-school split of 180 tutored to 875 non-tutored students). This lets us bound the two effects directly from the data, without further assumptions. Non-tutored peers gain *at least* 1.67 words (0.19 SD): this follows because the maximum possible effect on tutored students, given their mean post-test score of 6.16 words, is 6.16 words itself, and anything less than a 1.67-word effect on peers would then be inconsistent with the observed 2.44-word school-level effect.[21] This lower bound already exceeds the 0.11 SD peer effect reported in Berlinski et al. (2023). Tutored students, symmetrically, gain *at most* 6.16 words, their observed mean post-test score.

Pinning down an *upper* bound for peer effects requires an additional assumption about the lower end of the effect on tutored students, since a sufficiently negative tutored-student

[20] $2.44 = (180/1{,}055) \times X$, solving for $X = 14.30$.

[21] $2.44 = 0.171 \times 6.16 + 0.829 \times X$, solving for $X = 1.67$. We also confirm that a 6.16-word effect is attainable for tutored students, since more students score zero in control schools (335) than the number who would have been tutored under our assumptions (163).

effect would otherwise make almost any peer effect arithmetically possible.[22] We consider two natural restrictions, following Kang and Keele (2018): that no student is harmed by treatment, and, more conservatively, that tutored students' true effect matches the 2.30-word effect estimated for the lowest-pretest-percentile group, who make up most tutored students. These yield peer-effect upper bounds of 2.94 and 2.47 words, respectively.[23]

Taken together, non-tutored peers gain between 1.67 and 2.94 words, and tutored students gain at most 6.16 words. These bounds rest on little more than the test's floor and ceiling and random assignment. They are sufficient to establish the paper's central claim about the split: a substantial share of the school-level effect accrues to students who were never tutored, and the lower end of the peer range already exceeds the spillover reported by Berlinski et al. (2023).

The bounds do, however, rely on one assumption about behavior we cannot observe: that the share of students who would have been tutored is the same in treatment and control schools. This is not innocuous. Control schools have somewhat more at-risk students than treatment schools (41 versus 35 percent at pre-test, Table 1), and schools select tutees on need, so a control school might have tutored a larger share than the 17.1 percent we observe in treatment schools. We therefore report how the bounds vary with this counterfactual share, s.

Two results follow, and neither depends on getting s right. First, the *existence* of spillovers is insensitive to it. Tutored students alone could account for the entire school-level effect only if $2.44/s \leq 6.16$, that is, only if $s \geq 0.396$. No plausible value approaches this. Treatment schools tutored 205 students while having 455 at-risk students (Table 1); a control school behaving identically would have tutored 18.4 percent of its students. The conclusion that non-tutored peers gained therefore survives any counterfactual tutored share below 40 percent, which is to say any share consistent with schools tutoring fewer than all of their at-risk students.

Second, the peer-effect lower bound is stable across the plausible range. Table 5 reports it for shares from 0.15 to 0.25. The lower bound on the gain to non-tutored peers ranges from 1.20 to 1.78 words, or 0.14 to 0.20 standard deviations, and exceeds the 0.11 SD spillover in Berlinski et al. (2023) for any counterfactual share below 0.286.

[22] The unrestricted upper bound is a mathematical curiosity, not a plausible scenario: it would require an effect of -35.09 words for tutored students, well outside the $[0, 30]$ range of the test.

[23] Applying Kang and Keele (2018)'s TOT approach instead of weighted means gives similar bounds: $[0, 6.16]$ for tutored students and $[1.67, 2.86]$ for peers, though their method assumes potential outcomes depend only on the *number* of treated peers rather than peer identity, a strong assumption here.

Table 5: Sensitivity of the Peer-Effect Bounds to the Counterfactual Tutored Share

Counterfactual tutored share s	Peer effect, lower bound	Peer effect, upper bound: No harm	Peer effect, upper bound: Tutored effect = 2.30
0.150	1.78	2.87	2.46
0.171 (reported)	1.67	2.94	2.47
0.184 (tutored/at risk same in control)	1.60	2.99	2.47
0.200	1.51	3.05	2.47
0.250	1.20	3.25	2.49

Note. Bounds in words on the decoding test, from $2.44 = s \cdot \tau_T + (1 - s) \cdot \tau_P$, where τ_T is the effect on tutored students and τ_P the effect on non-tutored peers. The lower bound on τ_P sets τ_T at its maximum feasible value, the tutored students' mean post-test score of 6.16 words. The reported share of 0.171 is the observed split in treatment schools (180 tutored, 875 non-tutored, among students with a decoding post-test). The 0.184 row assumes control schools would have selected the same ratio of tutored students to at-risk students as treatment schools did (205 to 455).

5.3.2 The Social Multiplier

The bounds deliver more than a split. They identify how much of the return to tutoring a conventional evaluation would miss, without requiring an estimate of the effect on tutored students.

Consider what a within-school randomized trial recovers. Almost all tutoring experiments, including the two earlier trials of this program (Bøg et al., 2021; Seerup et al., 2025), randomize tutoring among eligible students *within* a school and compare them with untreated students in the same school. If tutoring spills over, that control group is contaminated: its members are the non-tutored peers. Such a design therefore does not estimate τ_T, the effect on tutored students relative to an uncontaminated counterfactual. It estimates $\tau_T - \tau_P$. The literature does not merely fail to count the gain to peers; it differences it out.

The social return per student in a school that adopts the program is our school-level estimate, $E = 2.44$ words. The return a within-school design would imply, scaling its estimate by the tutored share, is $s(\tau_T - \tau_P)$. The ratio of the two is the social multiplier (Glaeser et al., 2003). Using the decomposition $E = s\,\tau_T + (1 - s)\,\tau_P$,

$$s\,(\tau_T - \tau_P) = E - \tau_P, \qquad \text{so} \qquad M = \frac{E}{s\,(\tau_T - \tau_P)} = \frac{E}{E - \tau_P}. \tag{3}$$

The effect on tutored students cancels. The multiplier depends only on the school-level effect, which we estimate precisely, and on the peer effect, which we have bounded. It requires no instrument, no point estimate of τ_T, and none of the additional assumptions discussed in Section 4.

Substituting the lower bound on the peer effect gives a lower bound on the multiplier. At the observed tutored share, $\tau_P \geq 1.67$ implies $M \geq 2.44/(2.44 - 1.67) = 3.2$. Across the range of counterfactual tutored shares considered in Table 5, the bound runs from 2.0 to 3.7, and is above one throughout. A within-school randomized trial of this program would therefore have recovered at most about a third of its social return.

We regard this as the paper's central quantitative result. It is not a claim about how much a child's reading improves when their classmates read better, which is the structural parameter the peer-effects literature has found so difficult to pin down (Manski, 1993; Angrist, 2014; Carrell et al., 2013). It is a claim about the wedge between the private and social returns to a policy, estimated from experimental variation and bounded rather than assumed. If it is right, then the evidence base used to evaluate targeted tutoring, which is built almost entirely on within-school designs, understates the value of these programs by a factor of two to three.

The protocol also pre-specified an IV decomposition, which instruments tutoring status and peer exposure using the interaction of treatment assignment with a student's position above or below the 20th percentile. We report it in Online Appendix A1.7. The point estimates are consistent with the bounds above: tutored students gain 3.46 words and non-tutored peers 2.27 words. We do not build on them here. The tutored-student estimate is imprecise, and the identifying assumptions are ones we cannot defend as exactly true, for the reasons set out in Section 4. The bounds and the multiplier require none of these assumptions.

5.4 Effects on Other Skills and Non-Targeted Cohorts

Adopting the program required schools to reallocate staff time. If that reallocation harmed students in ways our primary outcome does not capture, the school-level effect we estimate would overstate what the program achieves. The cost-effectiveness analysis in Section 6 would understate its true cost. This section tests for such costs on two margins: other skills within the treated cohort, and the achievement of cohorts the program did not target.

We use the national survey of well-being conducted during the spring semester in kindergarten to assess the effects on social-emotional skills and well-being. Three scales measure distress, self-confidence, and school satisfaction. For distress, lower scores are beneficial, while higher scores indicate beneficial outcomes for the other two. Because the survey was conducted during early and mid-spring of the kindergarten semester, some treatment schools had not started the tutoring program when students answered the survey. Furthermore, both the tutoring start time and the timing of the survey are, to some extent, choice variables for schools, which may introduce endogeneity. We therefore report both ITT estimates mirroring the primary analysis, and TOT estimates from an IV approach in which the treatment indicator is the instrument for a school-level indicator equal to one if at least one student had started the tutoring program before the survey was conducted. 117 students had started tutoring before the survey, and the mean number of days between the start of tutoring and the survey date was 29.9 (range 1-64).

Online Appendix Table A6 displays the results. There are no significant ITT (Panel A) or TOT (Panel B) estimates. The TOT estimates for distress and self-confidence are, however, not small (g = -0.13 and 0.13) and in a beneficial direction. The TOT estimate for school satisfaction is in a harmful direction but smaller (g = -0.05).

Table 6 reports effects on non-targeted cohorts for the 76 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 6 shows a positive but statistically insignificant average treatment effect (0.036 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 2.8 percentage points lower in treatment schools compared to control schools and significant at the 5 percent level. The proportion in the top two categories is 2.5 percentage points higher, but not statistically significant.

There are thus no strong indications of *negative* effects on other skills or spillovers to older students. There is, however, a suggestion of *positive* ones, and it is worth pausing on it. The students in these cohorts were in grades 1 through 7 when the program ran. None was tutored, and none was in a kindergarten classroom where tutoring took place. A reduction in the share of low performers among them cannot operate through the improved reading of the tutored children, nor through any change in kindergarten class size. If it is real, it can only reflect something that changed at the level of the school: how its staff was deployed, or how they taught. We return to this in Section 7. We treat it cautiously here, because the average score effect is not significant and the result is one of three we test. Given the lack of consistent statistical significance across outcomes for the latter, we treat all 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 6 reflect true effects, our estimates understate program cost-effectiveness.

Table 6: Treatment Effects on Older Cohorts' Test Scores

	(1)	(2)	(3)
	Score (standardized)	Low score	High score
Treated schools	0.036	-0.028	0.025
	(0.047)	(0.013)	(0.022)
Constant	-0.104	0.144	0.650
	(0.115)	(0.039)	(0.056)
Observations	11067	11067	11067
R^2	0.005	0.022	0.012

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. All specifications include grade and year indicators, omitted from the table for brevity. We observe test scores for 76 out of the 81 schools.

6 Cost-Effectiveness

The program raised reading outcomes for kindergarten students in treatment schools at the cost of tutoring a few children. This section asks what that cost was per unit of learning, and how it compares with other interventions. We follow the method set out in Section 4.5, reporting an incremental cost as our primary estimate and a full-ingredients cost, which values tutor time at a teacher wage, as a conservative upper bound (Online Appendix A1.9). The evidence in Section 5.4 bears directly on this calculation: because we find no offsetting harm to other skills or to non-targeted cohorts, we set the opportunity cost of within-school reallocation to zero in the baseline, and note that if the positive patterns in Table 6 reflect true effects, our estimates understate cost-effectiveness.

The Swedish market price of the program materials at the time of the intervention was $946 for the first set purchased by a school.[24] Additional sets were available at a reduced price ($240).[25]

We provided treatment schools with one set per participating class. In total, 69 treated classes across the 37 treatment schools with post-test decoding outcomes (1.9

[24] All amounts are reported in US dollars, converted from Danish kroner using the 2022–2023 average exchange rate from https://nationalbanken.statbank.dk/nbf/100249.

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

classes per treated school; 16 schools contributed only one class) received materials, implying material costs of \$42,682. Follow-up phone calls required approximately one hour per treated school; at a wage of \$80 per hour, this corresponds to \$2,960. We cost this call for all 37 schools, because it is part of the resource package a school adopting the program would need, even though the research team reached 32 schools in practice; costing only the calls actually completed would lower the CES to \$159. Including all inputs, total costs were \$45,642. Dividing by the 1,055 students in treatment schools with post-test decoding outcomes yields a per-student cost of \$43.26. With a school-level effect size of 0.27 SD, the implied marginal-outlay cost-effectiveness statistic (CES) is \$160.

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. One methodological point determines which of our CES figures is the appropriate comparator. Both literatures we benchmark against report costs incrementally, charging the resources a program adds over and above business as usual and leaving the time of existing staff uncosted where that time is reallocated rather than added. Rosholm et al. (2021) do this explicitly: several of their day care arms cost under \$50 per child for a 20-week program, which is possible only because educator time is not priced. Hollands et al. (2013) likewise report incremental cost-effectiveness ratios. For their supplementary programs, which add external or specially funded staff, that added personnel is counted and makes up the large majority of program cost; but for K-PALS, which substitutes for regular classroom instruction using existing teachers, they treat the substitution as neither adding to nor subtracting from the cost of teacher time. Because our tutors are existing staff whose time is reallocated, the like-for-like comparator against both literatures is our incremental figure of \$160, not the full-ingredients figure, which would charge us for personnel that the comparison studies exclude. We therefore benchmark using the incremental figure and report the full-ingredients range only as an upper bound on our own cost.

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. TrygFonden's Centre for Child Research implemented all interventions. Of the 17 arms with a positive estimated effect, only one reports a lower CES than our marginal-outlay figure of \$160: LEAP-OPEN, a 20-week universal preschool program using play-based activities to support language and literacy, with an inflation-adjusted CES of approximately \$140 (for more information, see Bleses et al., 2018).[26] The next-

[26]The eighteenth arm, Suitcase, has a negative point estimate and hence a negative CES, which does not indicate greater cost-effectiveness. LEAP-OPEN targets a different population, all children in a

lowest arm, SPELL, has an inflation-adjusted CES of approximately \$203. Our ranking is thus not sensitive to modest changes in our cost estimate: our CES would have to more than double before a third arm fell below it.

Hollands et al. (2013) examine seven U.S. programs, five involving tutoring.[27] We compare against their alphabetics domain, which Hollands et al. (2013) define to include phonemic and phonological awareness, letter identification, print awareness, phonics, and word reading fluency, and which they measure partly through the Sight Word Efficiency and Phonemic Decoding Efficiency subtests of the Test of Word Reading Efficiency, treated as measures of phonics. Our decoding outcome, a real-word reading list scored for accuracy (Section 3.3), corresponds most closely to Sight Word Efficiency and thus to the phonics construct in this domain; the overlap is close but not exact, since our measure emphasizes accuracy over speed and is not primarily one of fluency. They report incremental cost-effectiveness ratios in 2010 dollars, which we adjust to the price level used here.

Only one program reports a lower CES than our incremental figure of \$160: K-PALS, a same-age peer-tutoring program targeting alphabetics in kindergarten (see Stein et al., 2008), at roughly \$50 (\$38 in 2010 dollars).[28] Two features make even this comparison favorable to us rather than the reverse. First, K-PALS reaches its low cost the same way our incremental figure does: it substitutes for regular classroom instruction using existing teachers, so Hollands et al. (2013) leave that teacher time uncosted, just as our incremental figure leaves reallocated tutor time uncosted. Second, its alphabetics effect was measured with an instrument created by the program's developers, and such treatment-inherent measures yield substantially larger effect sizes than independent ones (Slavin and Madden, 2011); K-PALS also ran for 18 to 20 weeks against our roughly 12. Its favorable ratio therefore rests in part on a non-independent outcome and a longer dosage. Furthermore, while same-age peer-tutoring interventions typically benefit at-risk students (e.g., Slavin 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).

In sum, across the 25 interventions in these two studies, only two have a lower CES than *Læseklar*. Neither LEAP-OPEN nor K-PALS is a tutoring intervention targeting students at risk of reading difficulties. Because effect sizes and costs cannot be fully harmonized across studies, we read this not as a precise ranking but as an indication that the program is comparatively cost-effective.[29]

preschool class rather than students at risk of reading difficulties, and is not a tutoring program.

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

[28]The next-lowest program, Stepping Stones, has an adjusted CES of approximately \$800.

[29]Our end-of-intervention estimates may appear more cost-effective than the tutoring intervention re-

7 Mechanisms

The school-level effects we estimate are not confined to the students who received tutoring. Something must transmit the gains to children who never sat with a tutor. This section asks what. We are not well placed to answer definitively: the study was not designed to identify mechanisms, and for some channels we lack the mediating variables entirely. But the candidate mechanisms make different predictions, and several of those predictions can be checked against evidence we already have. We therefore proceed by elimination rather than by estimation, stating what each mechanism implies and what the data say about it.

7.1 What Can Be Identified

Before turning to the mechanisms individually, it helps to be precise about what our data can and cannot recover. Figure 2 sets out a causal model. Solid arrows denote effects we can at least partially estimate; dashed arrows denote unobserved effects.

Reading from the bottom, Z is the treatment assignment, interacted with pre-test rank where relevant. It determines T_i, the uptake of tutoring and the identity of the tutor. T_i in turn determines T_g, the observable features of implementation: who was tutored, how many students per class, the number of sessions. But T_i also determines two things we do not observe. R is the reallocation of school resources, such as the reading specialist's time, that tutoring frees or consumes. P is the pedagogy used in regular classroom instruction, which may change when a teacher spends ten weeks tutoring with the program's methods. T_g, R and P all affect the literacy skills (L_t) and social-emotional skills (SE_t) of the tutored students, and these in turn may affect the outcome of their non-tutored classmates, Y_{nt}.

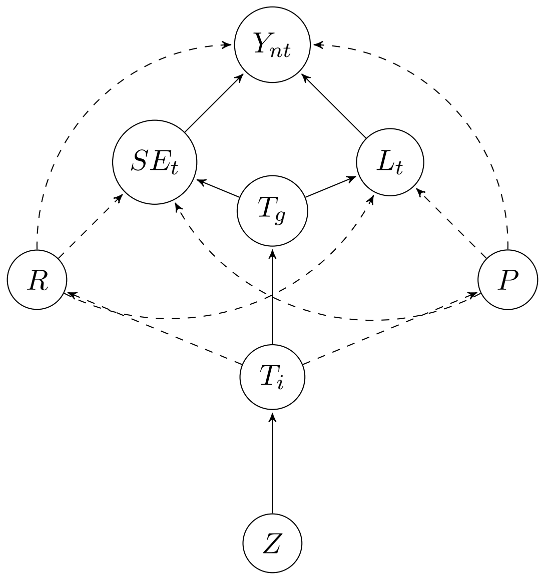
The critical feature of the figure is the pair of dashed arrows running from R and P directly to Y_{nt}, bypassing the tutored students entirely. Three consequences follow, and together they explain the choices we make in this paper.

First, linear-in-means peer effects models are valid here only if those arrows are absent. If resource reallocation or a change in classroom pedagogy affects non-tutored students directly, rather than only through the improved skills of their tutored classmates, then Z has a direct effect on Y_{nt} and the exclusion restriction fails. Treatment is not a valid instrument for tutored students' skills. This is the reflection problem in a particular guise (Manski, 1993), and it is why we do not estimate a structural peer-effect parameter.

ported 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.

Second, the same arrows explain why our IV decomposition is fragile in the way Section 4.2 describes, and why we do not rely on it.

Third, they explain what Kwon and Roth (2026) buys us. A mediation test of this kind permits direct effects of treatment on the outcome; it asks only whether the effect operates *entirely* through a specified mediator. It is therefore the right tool when the dashed arrows may be present, which is precisely our situation. Rejecting full mediation through L_t is equivalent to establishing that at least one dashed arrow is real.



Note. Z: treatment assignment. T_i: uptake of tutoring and tutor identity. T_g: observable implementation. R: reallocation of school resources (unobserved). P: pedagogy used in regular classroom instruction (unobserved). L_t, SE_t: literacy and social-emotional skills of tutored students. Y_{nt}: outcome of non-tutored students. Solid arrows are effects we can partially estimate; dashed arrows are unobserved.

Figure 2: What the Design Can Identify

The figure also tells us what would identify R and P separately, even though neither is observed. Both are consequences of T_i, the identity of the tutor. If the effect on non-tutored students is larger in schools where the classroom teacher tutored than in schools where the reading specialist did, conditional on the effect on tutored students, then R or P has a direct effect on Y_{nt}. We return to this below.

Class-size reduction. Berlinski et al. (2023) note that pulling tutored students out of the classroom mechanically reduces class size for those who remain, which may raise their achievement.[30] This mechanism predicts that peer effects should scale with the number of students withdrawn from a class, and should be absent where no student is withdrawn during instruction. Two features make a large effect unlikely a priori: schools typically withdrew two to three students from a class of roughly twenty, often one at a

[30] 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).

time, and tutoring was for the most part scheduled outside literacy instruction. We also test the prediction directly: we use the students scoring above the 20th percentile on the pre-test who were not tutored, and regress decoding scores on the number of tutored students in their class, together with the treatment indicator and the covariates from our main specification. The gradient is positive but imprecisely estimated: 0.26 words per additional tutored student, with a 95 percent confidence interval from −0.47 to 1.00, and the share of the class tutored gives a similar, insignificant estimate (0.088 words per percentage point, 95% CI = [-0.050, 0.23]). Both are robust to a quadratic term and to controlling for the number of eligible students. The interval is wide enough that we cannot exclude a class-size contribution large enough to account for much of the within-cohort peer effect, and we do not claim to. What we can say is that class size cannot be the whole story: the effect reaches students for whom class size did not change, namely the older cohorts of Section 5.4, in different classrooms and grades.

Reallocation of school resources. Tutoring may free personnel who would otherwise support struggling readers, allowing their time to be redirected. This mechanism has a sharp implication. In schools where the classroom teacher acted as tutor, the reading specialist's time was not consumed by tutoring and remained available for other students. In schools where the reading specialist acted as tutor, that resource was tied up. Considered in isolation, the mechanism therefore predicts *larger* peer effects where teachers tutored than where specialists did. We observe the split: 44 percent of students were tutored by their classroom teacher, 48 percent by a reading specialist, and 7 percent by both (Table 3). The point estimates do not bear out the prediction. Splitting treatment schools by tutor type, the peer effect is 2.42 words where only teachers tutored and 2.95 words where only specialists did, with confidence intervals that overlap almost entirely ($[0.13, 4.71]$ and $[0.48, 5.42]$).[31] We caution against reading this in either direction. Tutor type was a school's choice rather than a randomized assignment, and the two kinds of school differed at pre-test in which students they tutored: students tutored in teacher schools averaged 23.4 on the national language test and 16.9 on the pre-literacy test, against 27.3 and 22.2 in specialist schools (Online Appendix Table A7). The peer effect is measured on the students who were not tutored, so a difference in who was taken out of the classroom is precisely what would confound it. If the reading specialists were more effective tutors, then other mechanisms would also be affected. Thus, we treat the comparison as uninformative rather than as evidence against the channel. Because pedagogical transfer also runs through the classroom teacher, at least where reading specialists

[31] From indicators for treatment schools with only teacher tutors, only specialist tutors, and both types; the small both-tutors group is estimated at 2.91 words with a wide interval. Tutor information is missing for 364 students, and in the sample with information, the ITT effect on students above the 20th percentile is somewhat larger than in the full sample (2.71 versus 2.42 words). As they are estimated on a different sample, these estimates are not directly comparable to the bounds of Section 5.3.1 or the IV estimate of Online Appendix A1.7.

are not participating in classroom instruction, it predicts the same contrast, so this test bears on that channel too.

Improved skills of tutored students. If tutored students read better, their classmates may benefit through peer learning, or teachers may be able to pitch whole-class instruction higher, or attention may be freed for others. This is the mechanism the literature usually has in mind, and it is the only one we can test directly. In Online Appendix A1.10 we apply the method of Kwon and Roth (2026) to test the sharp null that the school-level effect operates entirely through the literacy gains of the lowest-performing students. The evidence runs against full mediation, but is not decisive. We reject the sharp null when the outcome is discretized into two bins ($p = 0.044$) and into ten ($p < 0.001$), but not at the five bins we take as our primary specification ($p = 0.065$). Because Kwon and Roth show that the test tends to over-reject with many bins, we put little weight on the ten-bin result. We therefore read this as evidence that improved skills among the tutored students are unlikely to be the whole story, without treating it as established.

Improved social-emotional skills. If tutoring raises tutored students' confidence or reduces classroom disruption, the learning environment may improve for everyone. The evidence does not support this as a major channel. Our well-being measures show no statistically significant effects, either in ITT or TOT form (Online Appendix Table A6). The TOT point estimates for distress and self-confidence are not negligible ($g = -0.13$ and 0.13) and are signed beneficially. But they are imprecise, and the survey timing is compromised: some treatment schools had not begun tutoring when it was administered. We regard this channel as unsupported rather than refuted.

Pedagogical transfer. A fifth mechanism is available to us and not to most tutoring studies, because our tutors were the schools' own staff rather than external hires. A teacher who spends ten weeks teaching phoneme–grapheme correspondence with figurines and finger-spelling may carry those methods back into whole-class instruction. This is not speculation. After the Swedish trial, several teachers reported using or planning to use the program's methods with all their students (Bøg et al., 2021), and teachers in the Danish efficacy trial expressed strong appreciation for the pedagogy (Seerup et al., 2025). School staff in this trial have mentioned similar practices to us informally, such as using finger-spelling with the whole class.

Pedagogical transfer makes a prediction the other mechanisms do not. Class-size reduction, resource reallocation and peer learning all operate through students who are near the margin of needing help, and should therefore concentrate their benefits low in the achievement distribution. Transfer of pedagogy to whole-class instruction should benefit students throughout the class, including those who were never at risk. Online Appendix Table A3 shows a positive and statistically significant decoding effect in the *top* quintile of the pre-test distribution (2.24 words). Section 5.2 shows that this is not an artifact of the handful of tutored students who rank high on the pre-tests: when they are removed, the effect on high-ranked students strengthens rather than disappears. These are gains among children who were never tutored and who were unlikely to be at risk of reading difficulties. That pattern is difficult to generate from a one-student reduction in class size or from the freed time of a reading specialist redeployed toward struggling readers, and it is what pedagogical transfer predicts.

One further piece of evidence bears on this, and it is the cleanest we have. Section 5.4 reports effects on *older cohorts*: students in grades 2 through 8, tested on national reading tests. These children were in grades 1 through 7 while the program ran. None was tutored. None sat in a classroom where tutoring took place. None had a classmate who was tutored. Whatever reaches them cannot travel through the improved reading of the tutored kindergarteners, cannot be a class-size effect, and cannot be a change in the social-emotional climate of a kindergarten room they were not in.

Yet the share of these students in the two lowest performance categories is 2.8 percentage points lower in treatment schools, significant at the 5 percent level (Table 6). We do not want to lean on this too heavily: the effect on average scores is positive but not significant, and this is one of three outcomes we examine. But the sign and the location of the effect are exactly what resource reallocation and pedagogical transfer predict, and they are the only two of our five mechanisms that could produce it at all. A reading specialist whose time is freed by having the classroom teacher tutor can be redeployed anywhere in the school. A teacher who learns a phonics pedagogy can use it with any class. Neither channel respects the boundary of the kindergarten cohort, and the older-cohort result suggests that in practice neither did.

We therefore read the evidence more cautiously than a clean decomposition would allow. Two things we can establish. Tutored students' own gains cannot account for the school-level effect alone, by the arithmetic of Section 5.3.1, and the mediation test runs weakly against full mediation through their literacy. And the effect is not confined to students who stood to gain from a smaller class or from resources redirected toward strugglers: it appears among never-at-risk students in the top quintile (Online Appendix Table A3) and, more tentatively, among older cohorts the program never reached (Table 6). Because those older students shared neither a tutor, a classroom, nor a class-size change with any tutored child, that pattern, to the extent it is real, is the part of the effect that must operate at the level of the school rather than the classroom.

What we cannot do is say which school-level channel dominates. Resource reallocation and pedagogical transfer both fit the evidence, and both would follow from using a school's own staff as tutors rather than external hires; our two attempts to tell them apart are either confounded or too imprecise to decide. What we are confident in is that a substantial part of the return operates school-wide rather than through the tutored children alone, which is exactly what a within-school design differences out and what the social multiplier of Section 5.3.2 quantifies without reference to any mechanism. Whether the finer choice between teachers and specialists changes the peer return is a question our data are too thin to answer, and we leave it as a hypothesis for adequately powered work.

We close by asking why our peer effects exceed those in Berlinski et al. (2023), who report 0.11 SD. Part of the answer may be larger effects on tutored students, since three of the five mechanisms above transmit tutored students' gains to their classmates. Three further differences are relevant. First, our primary outcome is decoding rather than reading comprehension. Decoding is a narrower skill on which large short-term effects are possibly easier to achieve. While Danish kindergarten guidelines require students to learn the alphabet, they do not require decoding (Ministry of Education, 2019). Many students score zero, and some may need only a small push to grasp the alphabetic principle. Such floor effects may also compress the standard deviation relative to Berlinski et al. (2023) and inflate standardized effect sizes. Our effects remain large in distribution regressions where the significant relative school-level effects range from 14 to 46% of the control group proportion (Online Appendix A1.5). However, that metric has no direct counterpart in their paper. Second, and as discussed above, our tutors were school personnel rather than external hires, which opens the pedagogical-transfer channel. Third, schools in our study chose which and how many students to tutor, and used criteria beyond our pre-tests (Figure 1). This flexibility may have produced assignments closer to optimal than a strict pre-test cutoff would, and teachers in the efficacy trial reported balancing the needs of tutored students against those of the rest of the class (Seerup, 2023).

8 Concluding Remarks

We evaluate an early literacy tutoring program in a cluster-randomized trial of 81 Danish schools and 2,583 kindergarten students. Decoding improved by 0.38 standard deviations among the pre-registered target group and by 0.27 standard deviations across all students in treated schools.

The paper's central claim is about the second of those numbers rather than the first. Because we randomized schools and tested every kindergartener, our estimate includes the children who never sat with a tutor, and they gained. Bounds that require nothing beyond random assignment and the range of the test put the gain for non-tutored peers at a minimum of 1.67 words, a lower bound of 0.19 standard deviations that exceeds the spillover reported by Berlinski et al. (2023). The conclusion that peers gained at all survives any counterfactual tutored share below 40 percent, which is to say any share consistent with schools tutoring fewer than all of their at-risk students.

The consequence is a social multiplier of at least three. A within-school randomized trial, which is how tutoring is almost always evaluated, has a control group drawn from the same classrooms and therefore contaminated by the spillover; it estimates the difference between the effect on tutored students and the effect on their peers rather than the effect on tutored students alone. Such a design would have recovered less than a third of this program's social return. The multiplier is identified from the bounds alone, requires no instrument, and runs from 2.0 to 3.7 across the counterfactual tutored shares we consider. If our estimate is representative, the evidence base on targeted tutoring, built almost entirely on within-school designs, understates the value of these programs by a factor of two to three.

Three features of the setting make this more than an arithmetic curiosity. The program draws on no external labor, because the tutors are staff the school already employs, so its adoption by one school does not reduce another's capacity to adopt. Between 10 and 13 percent of invited schools took it up when offered free of charge. And schools delivered roughly two-thirds of the recommended dosage while choosing their own participants. Yet, the program still raised school-level decoding by 0.27 standard deviations, above the 0.16 to 0.22 that Kraft et al. (2026) report for tutoring programs at scale. That the tutors are the school's own staff is also not incidental to the multiplier: it opens a channel absent from models that hire externally, and the effects we find among the highest-achieving students in the class are what such a transfer of pedagogy would predict and what the alternative mechanisms would not.

The program is correspondingly cheap. Because we measure outcomes for all students, our cost-effectiveness calculations price the social rather than the private return. Both literatures we benchmark against cost interventions incrementally, as we do. These comparisons cannot be fully harmonized across studies. Still, they indicate that the program is comparatively cost-effective, bettered on cost per unit of effect by only two of the 25 interventions in Rosholm et al. (2021) and Hollands et al. (2013).

Our evidence has limits that bear on the policy conclusion. Participation was voluntary, and the program was free, so we observe adoption at a zero price and cannot say how it would respond to a positive one. Schools could not be blinded, and tester blinding could not be guaranteed throughout. And while we can say that the school-level effect is unlikely to operate solely through the literacy gains of the lowest-performing students, we cannot identify which of the remaining channels does the work.

The most consequential limit is that we measure outcomes 23 days after the final tutoring session and can say nothing directly about persistence. This matters, because the literature gives ample reason for caution about the durability of short-run gains in early literacy (e.g., Bailey et al., 2020a; Hart et al., 2024; Dietrichson et al., 2026), and because our cost-effectiveness calculation divides costs by an effect measured at the end of kindergarten. Three considerations bear on how much weight the short horizon can carry. First, the comparison is like for like: the estimates in Rosholm et al. (2021) and Hollands et al. (2013) against which we benchmark are themselves short-run, so our favorable ranking is not an artifact of measuring earlier than others do. Second, interventions targeting a narrow and well-defined skill need not fade more rapidly than broader programs (Watts et al., 2025), and decoding is such a skill. Third, the question is answerable, and we have committed to answering it: our pre-analysis plan specifies a follow-up using the national reading test in second grade (Bingley et al., 2024), which we will report once the data become available for our cohorts. Until then, our cost-effectiveness estimates should be read as short-run figures, and the policy case for the program rests in part on a persistence assumption we have not yet tested.

Two implications follow for how tutoring is evaluated. Designs that randomize students within schools cannot see the social return and, worse, subtract part of it; school-level randomization with outcome measurement for all students should become more common. And because the multiplier appears to operate through the school's own staff, the question of who tutors is not merely one of labor supply. It determines how much of the benefit reaches the children who are never tutored at all.

References

Al-Ubaydli, O. and List, J. A. (2024). Will It Scale? *Issues in Science and Technology*, 41(1):34–36.

Andrews, I. and Armstrong, T. B. (2017). Unbiased Instrumental Variables Estimation Under Known First-Stage Sign. *Quantitative Economics*, 8(2):479–503.

Angrist, J. D. (2014). The Perils of Peer Effects. *Labour Economics*, 30:98–108.

Athey, S. and Imbens, G. W. (2017). The Econometrics of Randomized Experiments. In Banerjee, A. and Duflo, E., editors, *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.

Bailey, D. H., Duncan, G. J., Cunha, F., Foorman, B. R., and Yeager, D. S. (2020a). Persistence and Fade-Out of Educational-Intervention Effects: Mechanisms and Potential Solutions. *Psychological Science in the Public Interest*, 21(2):55–97.

Bailey, D. H., Oh, Y., Farkas, G., Morgan, P., and Hillemeier, M. (2020b). Reciprocal Effects of Reading and Mathematics? Beyond the Cross-Lagged Panel Model. *Developmental Psychology*, 56(5):912.

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.

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*, https://doi.org/10.1257/rct.11007-1.1.

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.

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

Borusyak, K. and Hull, P. (2023). Nonrandom Exposure to Exogenous Shocks. *Econometrica*, 91(6):2155–2185.

Carneiro, P. and Heckman, J. J. (2003). Human Capital Policy. In Heckman, J. J. and Krueger, A. B., editors, *Inequality in America: What Role for Human Capital Policies?*, pages 77–240. MIT Press, Cambridge, MA.

Carrell, S. E., Sacerdote, B. I., and West, J. E. (2013). From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation. *Econometrica*, 81(3):855–882.

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.

Cortes, K., Kortecamp, K., Loeb, S., and Robinson, C. (2024). A Scalable Approach to High-Impact Tutoring for Young Readers: Results of a Randomized Controlled Trial. Working Paper 32039, National Bureau of Economic Research.

Damm, A. P., Mattana, E., Nielsen, H. S., and Rouland, B. (2021). Academic Achievement and Wellbeing of Dual Language Learners: Evidence from a Busing Program. *Journal of Urban Economics*, 126:103358.

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.

Dietrichson, J., Bhatnagar, R., Filges, T., and Vembye, M. H. (2026). On the Mechanisms of Intervention Effect Fade-Out: A Meta-Analytic Review of Interventions Targeting At-Risk Students' Achievement. *Psychological Bulletin*, 152(1):66.

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. Technical Report EdWorkingPaper No. 25-1198., Annenberg Institute for School Reform at Brown University.

Duflo, E., Dupas, P., and Kremer, M. (2011). Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. *American Economic Review*, 101(5):1739–1774.

Dunn, O. J. (1961). Multiple Comparisons Among Means. *Journal of the American Statistical Association*, 56(293):52–64.

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.

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.

Glaeser, E. L., Sacerdote, B. I., and Scheinkman, J. A. (2003). The Social Multiplier. *Journal of the European Economic Association*, 1(2-3):345–353.

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.

Hart, E. R., Bailey, D. H., Luo, S., Sengupta, P., and Watts, T. W. (2024). Fadeout and Persistence of Intervention Impacts on Social–Emotional and Cognitive Skills in Children and Adolescents: A Meta-Analytic Review of Randomized Controlled Trials. *Psychological Bulletin*, 150(10):1207.

Heckman, J. J. (2013). *Giving Kids a Fair Chance*. MIT Press, Cambridge, MA.

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.

Kang, H. and Keele, L. (2018). Spillover Effects in Cluster Randomized Trials with Noncompliance. Technical Report arXiv:1808.06418v2, arXiv.

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. Technical Report IZA DP No.

17217, Institute of Labor Economics (IZA).

Knudsen, E. I., Heckman, J. J., Cameron, J. L., and Shonkoff, J. P. (2006). Economic, Neurobiological, and Behavioral Perspectives on Building America's Future Workforce. *Proceedings of the National Academy of Sciences*, 103(27):10155–10162.

Kraft, M. A., Edwards, D. S., and Cannata, M. (2024). The Scaling Dynamics and Causal Effects of a District-Operated Tutoring Program. EdWorkingPaper 24-1030, Annenberg Institute at Brown University.

Kraft, M. A., Schueler, B. E., and Falken, G. T. (2026). What Impacts Should We Expect from Tutoring at Scale? Exploring Meta-Analytic Generalizability. *Review of Educational Research*, Advance online publication:https://doi.org/10.3102/00346543261446660.

Krueger, A. B. (2003). Economic Considerations and Class Size. *The Economic Journal*, 113(485):F34–F63.

Kwon, S. and Roth, J. (2026). Testing Mechanisms. *Review of Economic Studies*, forthcoming:10.1093/restud/rdag028.

Lee, D. S. (2009). Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. *Review of Economic Studies*, 76(3):1071–1102.

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(1):295–318.

Lindström-Sandahl, H. (2024). *Early Elementary School Interventions in Reading and Mathematics*. Linkopings Universitet (Sweden).

List, J. A. (2022). *The Voltage Effect: How to Make Good Ideas Great and Great Ideas Scale*. Currency.

Manski, C. F. (1993). Identification of Endogenous Social Effects: The Reflection Problem. *Review of Economic Studies*, 60(3):531–542.

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, Børne- og Undervisningsministeriet, Copenhagen.

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.

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.

OECD (2023). PISA 2022 Results (Volume I): The State of Learning and Equity in Education. *OECD Pub.*, pages 1–10.

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.

Sacerdote, B. (2011). Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far? In Hanushek, E. A., Machin, S., and Woessmann, L., editors, *Handbook of the Economics of Education*, volume 3, chapter 4, pages 249–277. Elsevier.

Seerup, J. K. (2023). *Ready to Read: Exploring School-based Tutoring Interventions for Struggling Readers: PhD Thesis*. 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)*, 30(4):365–395.

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.

Slavin, R. E. and Madden, N. A. (2011). Measures Inherent to Treatments in Program Effectiveness Reviews. *Journal of Research on Educational Effectiveness*, 4(4):370–380.

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.

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., and Tymms, P. (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.

Watts, T. W., Hart, E. R., and Bailey, D. H. (2025). How General Is Educational Intervention Fadeout? A Meta-Analysis of Educational RCTs with Follow-Up. Technical report, EdWorkingPaper No. 25-1366. Annenberg Institute for School Reform at Brown University.

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.

Helping the Many by Tutoring the Few: The Social Multiplier of Early Literacy Instruction

Paul Bingley Jens Dietrichson Mette Friis-Hansen Vibeke Myrup Jensen

Section, table and figure numbers without an "A" prefix refer to the main paper.

Online Appendix

A1.1 Recruitment, Randomization, and Testing

The trial was conducted in two rounds with recruitment in spring 2022 and 2023, respectively. In both rounds, we targeted school units where relatively few parents held a tertiary education. The share of parents with tertiary education is strongly correlated with second-grade reading test scores, which concentrated the sample on schools likely to have many at-risk students. In round 1, we used a cutoff of at most 50% of parents with a tertiary degree. This cutoff yielded 394 eligible school units. Among eligible units, we drew a random ranking with the research team blind to school unit identity, and invited units in rank order. We invited school units with ranking numbers 1-80 (April 2022), and added ranking numbers 81-120 by June 2022 as we had a low response rate to our first invitation. In spring 2023, we invited 306 more public school units that had not already agreed to participate. Among these 306 school units, no more than half of the parents had a tertiary education (priority 1 sample). As a response to the low number of participating school units from round 1, we also invited 368 school units where 51-70 percent of the parents had a tertiary education (priority 2 sample). In total, we invited 674 school units in round 2.

We invited school units by e-mailing the manager responsible for kindergarten (typically, the school principal). We conducted follow-up phone calls, stating clearly that participation was conditional on agreeing to end up in the control group. In round 2, we prioritized follow-up phone calls for school units in our priority 1 sample, and a total of 246 school units from our priority 2 sample did not get a follow-up phone call.

In total, we invited 794 school units and 123 accepted. By the time of randomization to treatment and control groups, 40 school units (33 percent) had dropped out, mainly due to changes in school management or teacher leave and turnover, leaving 83 units. As described in Section 3.1, units sharing management were treated as one school in the randomization. Because each of the two schools comprises two units, randomization assigned 40 schools to treatment and 41 to control, corresponding to 42 and 41 school units, respectively.

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

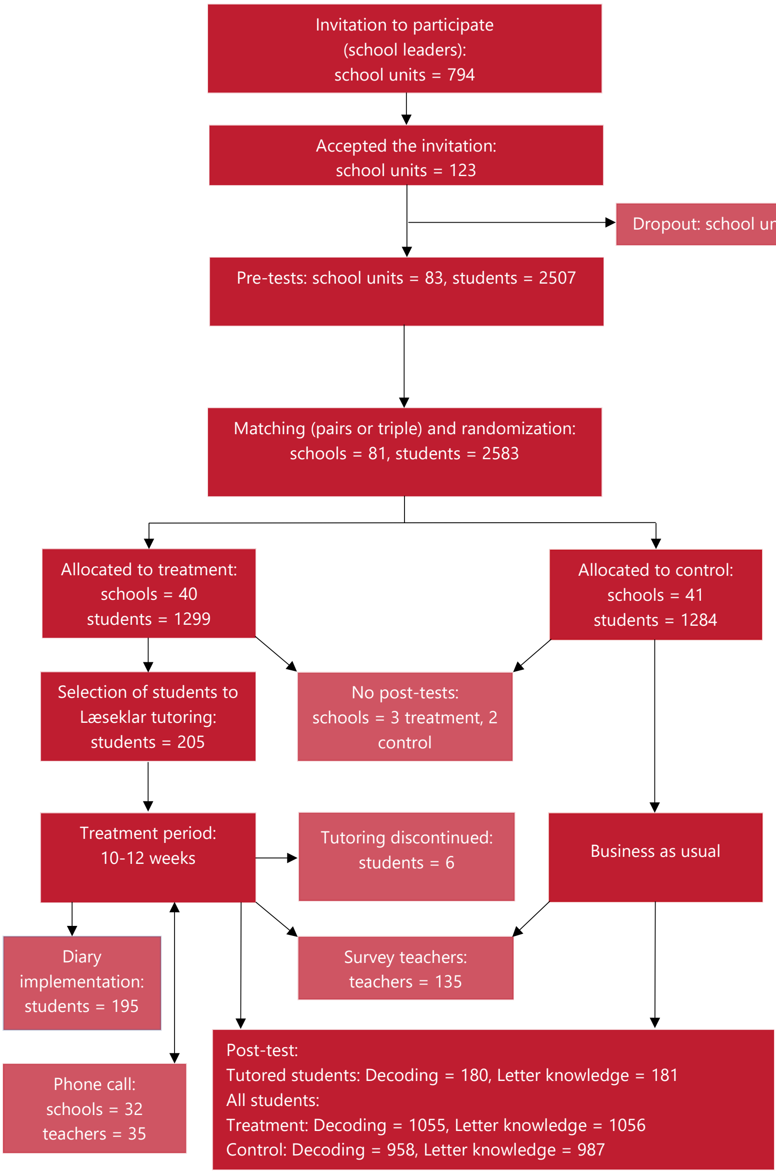


Figure A1: Flowchart of the Recruitment, Randomization, and Testing Process

A1.2 School-Level Balance Test

Table A1 reports school-level means and standard deviations for register-based variables on mothers and participating children. These data are complete for all schools and missing for a single student whom we could not match to the registers, so this table describes the sample as randomized, before any attrition. It is therefore the cleanest available test of whether randomization succeeded.

Differences are small, but three are not zero, and we report all of them. Treatment schools have more mothers in the top disposable-income quintile (a difference of 1.4 percentage points, significant at the 5 percent level) and more college-educated mothers (3.9 percentage points, significant at 10 percent). Both would, if anything, favor treatment schools. Treatment schools also have more mothers with missing information from the population register (0.6 percentage points, significant at 10 percent). Our covariate-adjusted primary specification conditions on the pre-test rankings and at-risk indicators that these characteristics predict; the effect remains in sensitivity analyses that include the characteristics directly (Section A1.4), and the randomization-inference omnibus test reported in Section 3 does not reject balance.

Table A1: Balance Tests for Schools

Variable	Treatment Schools Mean (SD)	Control Schools Mean (SD)	Difference (SE)
Mother Danish	0.71 (0.21)	0.76 (0.18)	-0.045 (0.037)
Disposable income q1	0.36 (0.13)	0.39 (0.15)	-0.029 (0.031)
Disposable income q2	0.33 (0.10)	0.34 (0.13)	-0.009 (0.025)
Disposable income q3	0.18 (0.09)	0.17 (0.09)	0.010 (0.021)
Disposable income q4	0.09 (0.07)	0.07 (0.06)	0.019 (0.016)
Disposable income q5	0.02 (0.04)	0.01 (0.02)	0.014 (0.006)
Single	0.18 (0.11)	0.20 (0.12)	-0.017 (0.026)
Missing population register	0.01 (0.02)	0.00 (0.01)	0.006 (0.003)
Basic education	0.26 (0.16)	0.25 (0.12)	0.006 (0.031)
College education	0.17 (0.09)	0.13 (0.09)	0.039 (0.019)
Missing education register	0.02 (0.03)	0.01 (0.02)	0.006 (0.005)
Employed	0.68 (0.10)	0.64 (0.13)	0.036 (0.026)
Unemployed	0.04 (0.04)	0.03 (0.04)	0.001 (0.008)
Out of labor force	0.26 (0.10)	0.30 (0.12)	-0.035 (0.026)
Missing labor force register	0.01 (0.02)	0.01 (0.02)	0.007 (0.005)
Parents live together	0.69 (0.14)	0.66 (0.12)	0.036 (0.027)
Child Danish	0.67 (0.22)	0.71 (0.19)	-0.047 (0.041)
Boy	0.52 (0.12)	0.48 (0.14)	0.042 (0.027)

Note. School-level means and standard deviations for the 81 schools as randomized, before attrition. The Difference column reports the treatment-control difference in school-level means with standard errors clustered on pair/triple in parentheses.

A1.3 Bounds on the Treatment Effect Under Attrition

Because the response rate is higher in the treatment arm, we bound the treatment effect following Lee (2009) by trimming the treatment group: we discard the highest-scoring and, alternatively, the lowest-scoring 8.1 percent of treated students with a decoding outcome, so that the trimmed treatment response rate equals the control response rate.[1] The trimming is milder than the assumption-free worst case: it removes the 86 marginal treated responders as if they were the highest-scoring treated students, rather than imputing missing scores at the top of the control distribution and the bottom of the treatment distribution.

The unconditional bounds on the decoding effect run from 0.81 to 3.20 words. Tightening the bounds within pre-test strata, defined by the at-risk-of-literacy and at-risk-of-language indicators, narrows them to between 0.57 and 2.72 words.[2] The lower bound is positive in both cases, so the estimated effect does not reverse under the worst case consistent with our attrition. Once sampling uncertainty is added around the bound, the confidence intervals include zero (unconditional, $[-0.22, 1.84]$; tightened, $[-0.56, 1.69]$); this bears on the precision with which the bound is estimated rather than on whether the effect is present, since the point bounds are positive.

The worst case implied by the trimming would require the marginal students retained in treatment schools to be systematically stronger readers, which would tend to show up as imbalances among attritors in the characteristics we observe; it does not (the omnibus test does not reject joint orthogonality among attritors, $p = 0.48$). The pre-test at-risk share is somewhat higher in control schools (Table 1), which, if it survives attrition, would tend to depress control-group post-test scores and inflate the estimated effect. Our covariate-adjusted specifications condition on the at-risk indicators and within-school pre-test ranks, and the randomization-inference omnibus test on the analysis sample does not reject balance ($p = 0.285$).

[1] The trimming fraction is $1 - 0.746/0.812 = 0.081$, or 86 of the 1,055 treated students with a decoding post-test.

[2] The tightened bounds condition on the two binary at-risk indicators rather than the full covariate set, since the trimming is implemented within discrete strata.

A1.4 Sensitivity Analysis

This section presents results from alternative specifications of our primary ITT analysis.

Estimating at the school rather than student level produces very similar but slightly less precise results. Without covariates, the estimates are 2.33 for decoding ($p = 0.015$) and 0.54 for letter knowledge ($p = 0.027$). Dropping the 5% worst-matched pairs reduces the sample to 72 schools and yields estimates of 2.27 for decoding ($p = 0.023$) and 0.54 for letter knowledge ($p = 0.033$).

We showed in Table 2 that our results are close to the primary analysis when we omit all covariates and compare the means in the treatment and control groups (both estimates are significant). 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. Columns 1 and 2 include the full sample, 3 and 4 the students at or below the 20th percentile in the joint 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 midpoint between these two estimates.

Two further observations are worth making. First, the primary estimate of 2.44 words sits above three of the four alternatives (2.19, 2.05 and 2.16) and below the strata fixed-effects estimate (2.72). It is close to the midpoint of the range, but a reader should know that it is not the most conservative specification we could have chosen.

Second, and more importantly, the paper's headline quantity is the social multiplier rather than the ITT estimate itself, and the multiplier is robust across every specification we report. Because $M = E/(E - \tau_P)$ and the lower bound on τ_P is itself a function of E, both move together. Recomputing the bound of Section 5.3.1 under each alternative gives multipliers of 2.7 (Lin), 3.8 (strata fixed effects), 2.4 (register covariates only), 2.6 (pre-test and register covariates), 3.0 (school-level, no covariates) and 2.8 (dropping the worst-matched pairs), against 3.2 in the primary specification. The multiplier exceeds two under every specification, and the conclusion that a within-school design would recover well under half the social return does not depend on the covariate set.

Table A2: Alternative Specifications of the Average Treatment Effects

	(1)	(2)	(3)	(4)	(5)	(6)
	All	All	≤ 20th	≤ 20th	> 20th	> 20th
Variables	Decoding	Letters	Decoding	Letters	Decoding	Letters
Panel A: Lin-specification						
Treatment	2.19	0.33	5.88	3.25	1.97	0.24
	(0.91)	(0.18)	(2.37)	(1.29)	(0.90)	(0.17)
Constant	8.57	26.60	3.04	25.16	8.89	26.66
	(0.50)	(0.15)	(0.87)	(0.86)	(0.50)	(0.14)
Covariate set	Lin	Lin	Lin	Lin	Lin	Lin
Students	2,013	2,043	425	425	1,588	1,618
R^2	0.19	0.17	0.13	0.12	0.15	0.12
Panel B: Including strata fixed effects						
Treatment	2.72	0.57	2.32	1.82	2.83	0.31
	(0.86)	(0.25)	(0.84)	(0.63)	(0.97)	(0.21)
Constant	8.06	27.27	2.56	25.70	9.36	27.45
	(0.61)	(0.14)	(1.04)	(0.54)	(0.71)	(0.14)
Covariate set	Strata FE	Strata FE	Strata FE	Strata FE	Strata FE	Strata FE
Students	2,013	2,043	425	425	1,588	1,618
R^2	0.29	0.23	0.27	0.30	0.26	0.18
Panel C: Non-missing register data-based covariates						
Treatment	2.05	0.56	1.71	1.33	2.01	0.30
	(0.79)	(0.21)	(0.71)	(0.44)	(0.89)	(0.21)
Constant	7.72	26.41	-0.49	25.40	8.79	26.58
	(2.50)	(0.53)	(0.80)	(1.81)	(2.72)	(0.54)
Covariate set	Reg	Reg	Reg	Reg	Reg	Reg
Students	2013	2043	425	425	1588	1618
R^2	0.08	0.05	0.07	0.05	0.08	0.05
Panel D: Pretest and non-missing register data-based covariates						
Treatment	2.16	0.55	2.19	1.70	2.12	0.29
	(0.70)	(0.19)	(0.69)	(0.48)	(0.77)	(0.16)
Constant	5.90	26.53	-0.72	25.79	6.24	26.46
	(2.05)	(0.56)	(1.65)	(1.10)	(2.33)	(0.55)
Covariate set	Pre+Reg	Pre+Reg	Pre+Reg	Pre+Reg	Pre+Reg	Pre+Reg
Students	2013	2043	425	425	1588	1618
R^2	0.22	0.19	0.15	0.14	0.19	0.15

Note. All covariates are omitted for brevity. Panel A specifications include within-school pre-literacy and language score rankings, indicators for missing pre-tests, and their interactions with treatment. The missing language indicator is omitted in columns 3–4 because no values are missing in the 20th-percentile group. The treatment coefficient reports the average treatment effect. The constant represents the control-group mean for students with non-missing pre-tests evaluated at the mean rank values. Standard errors, clustered at the 40 matched pairs/triples, are reported in parentheses.

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.

A1.5 Effects Across the Pre- and Post-Test Outcome Distribution

Figure A2 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; positive coefficients therefore indicate higher probabilities of reaching each threshold in treatment schools.[3]

In the upper-left panel, treatment effects are positive over most of the decoding distribution and statistically significant from 1 to 22 words, with magnitudes of 5–15 percentage points. Relative effects are large: the share able to read at least one word is 14% higher in treatment schools, and 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 begins at 26 words).

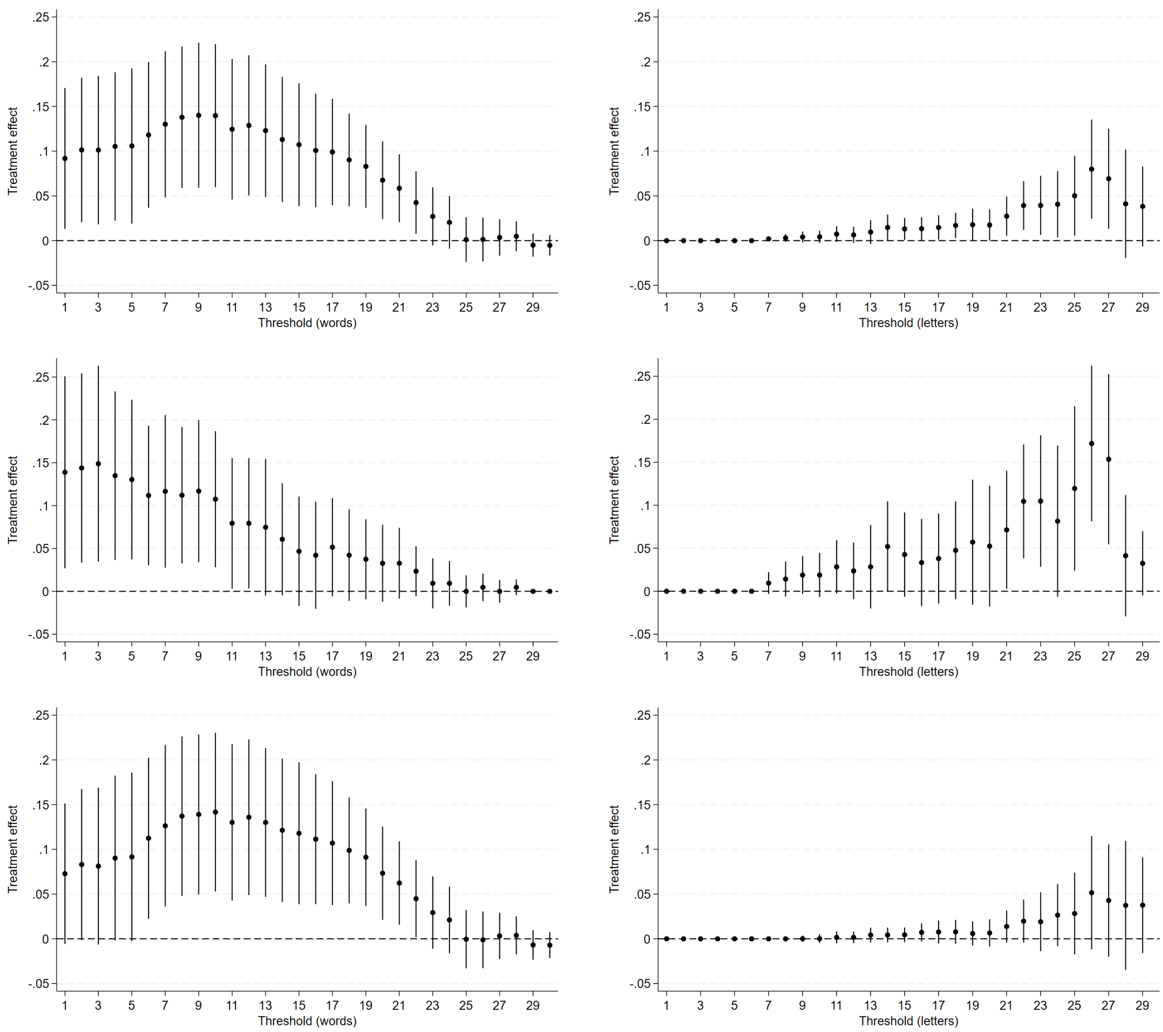
The two groups behave differently, and the difference is informative. Among students at or below the 20th percentile, the group the program targeted and from which most tutees were drawn, effects are about 15 percentage points at low thresholds, decline roughly linearly, and become statistically insignificant at 12 words and above. Relative effects remain large: the share able to read at least one word is 34% higher, and the share reaching 12 words is 94% higher. This is the signature of children being lifted off the floor.

Among students above the 20th percentile, few of whom were tutored, the pattern is the opposite. Effects are about 10 percentage points up to five words and then *increase* in size up to 11 words. The effects are statistically significant from 6 through 22 words. These children were not being rescued from illiteracy; they were reading more words across the middle and upper range of the distribution. A spillover operating purely through the improved skills of the weakest readers, or through the marginal reduction in class size when one child leaves the room, is hard to reconcile with gains concentrated among students who were likely already reading.

The right panels of Figure A2 show results for letter knowledge. Patterns are similar for the full sample and for students at or below the 20th percentile: effects become statistically significant around 21 letters, rise through 26, then decline and become insignificant by 28. Effects are larger for low-achieving students, peaking at 17 percentage points (41%) versus 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 A3 reports treatment effects for five quintiles of the pre-literacy (rank) distribution. In Column (1), decoding effects are statistically significant below the 60th percentile and in the top quintile (2.24 words), and positive but imprecise between the 60th and

[3] These regressions omit covariates, for the reasons given in Section 4.1, and we make no adjustment for testing across thresholds. We use them to characterize where in the distribution effects fall, not to conduct inference on individual thresholds.



Note. The figure shows LPM coefficients on the treatment indicator with 95% confidence intervals (clustered by pair/triple). Outcomes are binary indicators of decoding and letter knowledge post-test performance. The y-axis reports treatment effects; the x-axis indexes thresholds. For threshold k, the outcome equals one if the student decodes (or knows) at least k words (letters), and zero otherwise, for k=1,...,30 (decoding) and k=1,...,29 (letter knowledge). The top panel uses the full sample, the middle panel students at or below the 20th percentile of the pre-test distribution, and the bottom panel those above. Left-hand panels use decoding as the outcome; right-hand panels use letter knowledge.

Figure A2: Average Treatment Effects Across the Outcome Distributions

80th. In Column (2), letter knowledge effects are largest in the bottom quintile and remain positive up to the 60th percentile, but are small and insignificant above, consistent with ceiling effects among higher performers.

Table A3: Heterogeneity Across the Literacy Pre-Test Quantiles

Variables	(1) Decoding	(2) Letters
Treatment$\times 1(i < 20th)$	2.45	1.73
	(0.58)	(0.47)
Treatment$\times 1(20th \geq i < 40th)$	1.96	0.55
	(0.80)	(0.33)
Treatment$\times 1(40th \geq i < 60th)$	4.18	0.97
	(1.05)	(0.33)
Treatment$\times 1(60th \geq i < 80th)$	1.20	0.22
	(1.16)	(0.31)
Treatment$\times 1(i \geq 80th)$	2.24	-0.17
	(1.13)	(0.21)
$1(i < 20th)$	1.47	22.47
	(0.35)	(0.59)
$1(20th \geq i < 40th)$	4.08	25.30
	(0.58)	(0.34)
$1(40th \geq i < 60th$	5.27	25.84
	(0.67)	(0.31)
$1(60th \geq i < 80th)$	9.32	26.63
	(0.87)	(0.24)
$1(i \geq 80th)$	13.64	27.60
	(0.71)	(0.17)
Missing pretest	-4.81	-2.44
	(2.14)	(0.60)
Students	2,013	2,043
R^2	0.59	0.99

Notes: Treatment effects by percentile group x are given by the coefficients on Treatment $\times 1(x)$. The five percentile group indicators are exhaustive, so the model is estimated without an intercept and the reported R^2 is uncentered. It is therefore not comparable to the R^2 in Table 4, and is mechanically close to one for letter knowledge, whose mean is large relative to its standard deviation. Standard errors clustered at the pair/triple level (40 clusters) are in parentheses.

A1.6 Isolating Effects on Non-Tutored High Achievers

Table A4 adds an indicator for tutored students to the treatment regression on the pre-registered sample of students above the 20th percentile, so that the treatment coefficient identifies the effect on students who were not tutored. The tutored indicator is small and statistically insignificant for both outcomes, while the treatment effect remains large and, for decoding, strongly significant. The gains among higher-ranked students therefore do not come from the few tutees among them.

That the coefficient on the indicator for tutored students is negative does not imply that the treatment effect on these students is negative. It implies that tutored students scored (insignificantly) lower than the control group mean, which, because tutored students are selected because they are at risk of reading difficulties, could be consistent with positive treatment effects.

Table A4: Effect on Non-Tutored Students Above the 20th Percentile

	(1) Decoding	(2) Letter knowledge
Treatment	2.49	0.36
	(0.82)	(0.18)
Tutored	−0.66	−0.46
	(0.75)	(0.34)
Covariates	Yes	Yes
Observations	1588	1618

Note. Estimates from the treatment regression on students above the 20th percentile of the within-school pre-test distribution, the pre-registered peer group. Treatment is the school-level treatment indicator; Tutored is an indicator for students in this sample who received tutoring, so the Treatment coefficient identifies the effect on non-tutored students. Covariates are those of the main specification. Standard errors clustered at the pair/triple level are in parentheses.

A1.7 Instrumental Variables Decomposition

This appendix reports the IV decomposition pre-specified in our protocol (Bingley et al., 2024). We present it for completeness and because it was pre-registered, not because we rely on it. The bounds in Section 5.3.1 and the multiplier in Section 5.3.2 require none of the assumptions below, and the estimates here are consistent with them.

A1.7.1 Estimating equations

For point estimates, we use a two-stage least squares (2SLS) IV approach to estimate effects on program-adherent students and their non-tutored peers. School-level adherence was high: only one treatment school discontinued the program, while all others were classified as adherent.

The first-stage equation is:

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

where the outcome variable is either adherence to the program or being a non-tutored peer in a school where at least one student adhered to the 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 equation is:

$$y_{iPost} = \alpha + \rho_1 \hat{T}_i + \rho_2 \hat{T}_i^{peer} + X_i \delta_2 + \epsilon_i \tag{2}$$

where y_{is} denotes the post-test outcome, $\hat{T}_i$ represents the predicted treatment status for students who comply with the program, and $\hat{T}_i^{peer}$ is the analogous measure for their non-tutored peers. Standard errors are clustered at the pair/triple level in both stages, as well as in the reduced-form specification.

A1.7.2 Why we do not rely on these estimates

A key methodological challenge is interference among students within the same school. Although the design limits cross-school interference, assigning tutoring to one student may lower the chance that peers receive tutoring, and tutoring may also affect the outcomes of non-tutored peers. Interference may therefore arise through both instrument-to-treatment and treatment-to-outcome spillovers (e.g., Nibbering and Oosterveen, 2025).

Even without interference, 2SLS with two endogenous regressors requires much stronger identifying assumptions. Beyond the usual conditions of relevance and exclusion, Bhuller and Sigstad (2024) show that two further assumptions are needed. First, a monotonicity-type assumption: for every agent, instrument 1 has a non-negative partial effect on treatment 1, at least on average. Second, no "cross effects": conditional on the instruments, instrument 1 has no partial effect on treatment 2 for any agent.[4]

The no-cross-effects assumption is unlikely to hold in our setting, and we do not claim otherwise. Schools tutored a limited number of students, so assigning a student to tutoring mechanically reduces the probability that a classmate is tutored rather than becoming a non-tutored peer. The instrument for one treatment therefore has a partial effect on the other, which is precisely the violation Bhuller and Sigstad (2024) identify, and the two treatments are close to mutually exclusive within a school, the case in which cross effects matter most. A related warning comes from Kang and Keele (2018), whose bounds we use in the main text: they show that in cluster-randomized trials with noncompliance, instrumental variables do not identify the usual complier average causal effect once treatment spills over to non-compliers, and that no analysis of such data can unbiasedly estimate local network effects. Our setting has exactly this structure.

Finally, Keane and Neal (2023) argue that standard 2SLS t-tests perform poorly and recommend the Anderson–Rubin test in just-identified models and a conditional likelihood-ratio test in overidentified models. Our specification has two instruments and two endogenous regressors, so the order condition holds with equality and the model is exactly identified. We therefore follow their recommendation for the just-identified case and report Anderson–Rubin confidence sets alongside conventional standard errors below.

Despite these problems, 2SLS may still yield less biased estimates than regression or matching, offering a better though not fully correct decomposition. We report the estimates for completeness and because they were pre-specified, and we note where they agree with the bounds. We do not build the paper's argument on them.

The protocol specified two further refinements to the IV analysis that we do not implement, and we explain why here. The first is the recentering of Borusyak and Hull (2023), which corrects for non-random exposure to a shock. In our design, the shock is the school-level treatment assignment, randomized with known probability within matched strata, and the non-random pre-test indicators used to form the instruments enter the specification as controls. The conditional expectation of each instrument is therefore known and absorbed, the recentering leaves the estimator unchanged, and the non-random-exposure problem the method addresses does not arise. The second is the unbiased estimator of Andrews and Armstrong (2017), which removes finite-sample bias when the sign of the first stage is known; it is derived for a single endogenous regressor, and we are aware of no implementation for the two-endogenous-regressor case. In any event, our first stages are strongly identified, with F-statistics of 90 and 2,418, so finite-sample bias of the kind it targets is not a first-order concern. More fundamentally, the decomposition is a

[4]Fusejima (2024) propose an alternative strategy based on rank similarity, which assumes that each student's rank in the outcome distribution is invariant across treatment states, up to non-systematic, identically distributed "slippages." This is also a strong assumption (Chernozhukov et al., 2017).

pre-specified robustness exercise that we already regard as suggestive at best, because within-school interference violates the assumptions it requires, so neither refinement bears on the conclusions, which rest on the bounds.

A1.7.3 Estimates

Table A5 reports IV estimates of the effect on tutored students and non-tutored peers separately, with an OLS benchmark in column (1) for comparison. Tutored students gain 3.46 words and non-tutored peers gain 2.27 words (column 5), both consistent with the bounds derived in Section 5.3.1. The peer effect is statistically significant at conventional levels ($p < 0.05$). The tutored-student estimate is imprecise and not significantly different from zero, with a standard error of 2.30, and the two estimates are not statistically distinguishable from one another. Following Keane and Neal (2023), we also report weak-instrument-robust confidence sets. Our specification has two instruments and two endogenous regressors and is therefore exactly identified, so we use the Anderson–Rubin approach, with projection-based sets for the individual coefficients. The joint Anderson–Rubin test rejects the hypothesis that both coefficients are zero ($\chi^2(2) = 11.9$, $p = 0.003$). The projection-based 95 percent confidence sets are $[-2.88, 10.16]$ for tutored students and $[-0.73, 5.09]$ for non-tutored peers using 50 x 50 gridpoints. Because projection-based sets are conservative, both are wider than the conventional intervals. The peer set includes zero even though the conventional estimate is significant. The imprecision of the tutored-student estimate is expected given that only 180 students in our sample were tutored. These are further reasons to rely on the bounds rather than the point estimates when characterizing the split.

The identification strategy instruments for tutoring status using the interaction of school-level treatment assignment with a student's position (above or below the 20th percentile) in the pre-test distribution. The instruments are strong: first-stage F-statistics of 90 (tutored students) and 2,418 (peers), and the reduced-form estimates in column (2) confirm the pattern seen throughout the paper. The IV approach requires stronger assumptions than a single-instrument design (formally discussed in Online Appendix A1.7.2 following Bhuller and Sigstad (2024)), since two endogenous variables (tutoring and peer exposure) are instrumented simultaneously. We view the IV estimates as a useful complement to, rather than a replacement for, the bounds: the OLS estimates in column (1) fall outside the bounds under some assumptions and imply a smaller effect for tutored students than for their peers, which is difficult to justify given that tutoring is targeted at struggling readers; the IV estimates do not have this problem, and are likely less biased than OLS if, as seems plausible, selection into tutoring reflects unobserved factors beyond the pre-tests.

Converting the IV estimates to effect sizes, tutored students gain $g = 0.39$ SD and non-tutored peers gain $g = 0.25$ SD (full-sample SD). Using the SD among the lowest-pretest-percentile students, a more comparable benchmark to prior tutoring studies, which typically restrict to similarly at-risk samples, the tutored-student effect rises to $g = 0.57$. This is larger than recent meta-analytic averages (Kraft et al., 2026; Nickow et al., 2024) and than the Danish efficacy trial of the same program ($g = 0.15$; Seerup et al., 2025), though smaller than the Swedish trial ($g = 1.07$; Bøg et al., 2021), which achieved closer-to-recommended dosage.[5] The tutored-student effect is therefore well within the range of plausible values given prior evidence on this program.

Table A5: 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				3.46
	(0.65)				(2.30)
Non-tutored peers	2.60				2.27
	(0.77)				(1.09)
Treatment $\geq$ 20th		2.38	0.11	0.88	
		(0.81)	(0.01)	(0.01)	
Treatment $<$ 20th		2.74	0.41	0.59	
		(0.73)	(0.04)	(0.04)	
Covariate set	Pretest	Pretest	Pretest	Pretest	Pretest
Students	2013	2013	2013	2013	2013
R^2	0.19	0.19	0.22	0.73	0.19
F-statistic			90.16	2417.66	

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. Column (1) reports OLS, column (2) the reduced form, columns (3) and (4) the two first stages, and column (5) the 2SLS estimates. Standard errors clustered on the pair/triple are shown in parentheses below the coefficient.

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

A1.8 Effects on Social-Emotional Skills and Well-Being

Table A6: Treatment Effects on Social-Emotional Skills and Well-Being

Panel A: Intention to treat estimates			
	(1)	(2)	(3)
Variables	Distress	Self-confidence	Satisfaction
Treatment	-0.04	0.04	-0.01
	(0.04)	(0.04)	(0.04)
Constant	-2.46	2.27	2.70
	(0.04)	(0.04)	(0.03)
Covariate set	Pretest	Pretest	Pretest
Students	2260	2250	2273
R^2	0.01	0.02	0.01
g	-0.08	0.09	-0.03

Panel B: Treatment on the treated estimates			
	(1)	(2)	(3)
Variables	Distress	Self-confidence	Satisfaction
Tutoring started	-0.06	0.06	-0.02
	(0.07)	(0.06)	(0.06)
Constant	-2.46	2.27	2.70
	(0.04)	(0.04)	(0.03)
Covariate set	Pretest	Pretest	Pretest
Students	2260	2250	2273
R^2	0.02	0.02	0.01
g	-0.13	0.13	-0.05
1st stage F	54.26	55.65	54.35

Note. The table reports treatment effects on outcomes derived from the national survey of well-being conducted during the spring semester in kindergarten. Panel A displays results from a specification mirroring the primary analysis, where random assignment to treatment and control schools defines the treatment indicator. Panel B shows the IV estimate from a 2SLS approach where the treatment indicator is used as an instrument for a school-level indicator equal to one if at least one student had started tutoring before the well-being survey was conducted (*Tutoring started*). Outcome variables are the distress scale in column (1), the self-confidence scale in column (2), and general school satisfaction in column (3). Effect sizes (Hedges' g), calculated as the effect estimate divided by the estimated SD. We estimate the SD using an intercept-only multilevel model to account for clustering of students within schools. Standard errors are clustered on the pair/triple in parentheses.

A1.9 Full-Ingredients Cost

The full-ingredients cost values tutor session time at the opportunity cost of a teacher hour, as though that time were labor added on top of business as usual. We report it as a conservative upper bound rather than as our comparator. As Section 6 explains, the studies we benchmark against price existing staff time incrementally, charging it only where it is added rather than reallocated, so setting our tutor time to its full teacher wage while they do not would overstate our cost relative to theirs.

Valuing tutor session time at a teacher wage yields a cost that lies between \$226 and \$347, at \$50 per teacher hour and bounding the average tutoring group size between one and three. Even at its upper end, \$347, the program remains among the most cost-effective interventions we consider, so our conclusions do not turn on where within the range the true value falls.

We calculate as follows: the 180 tutored students with a decoding post-test received 5.72 hours of tutoring on average (Table 3, estimated over the 157 tutored students with recorded dosage), or 1,030 student-hours; tutor-hours equal this figure divided by the average group size, which we bound at three because 205 tutored students were spread across 69 treated classes and the program forms groups within class. Adding the \$42,682 in materials and \$2,960 in follow-up calls, dividing by the 1,055 students with a post-test decoding outcome, and dividing by the school-level effect of 0.27 SD, the full-ingredients CES is \$341 at one-to-one tutoring, \$251 in groups of two, and \$220 in groups of three. We value a teacher hour at \$50, from average gross monthly pay including pension for Danish primary school teachers and an annual working-time norm of 1,680 hours (source: Kommunernes og Regionernes Løndatakontor); about half of tutors were reading specialists commanding qualification supplements, so this is a conservative lower bound on the tutor wage.

A1.10 Testing Mechanisms

To assess how much school-level treatment effects are mediated by the literacy skills of students in the bottom 20 percentiles of the pre-test literacy distribution, we apply the method of Kwon and Roth (2026). This method tests the sharp null of complete mediation, i.e., that a specified set of mediators fully explains the treatment effect.

Our implementation relies on two key assumptions. First, treatment is independent of potential outcomes, which is plausible given random assignment to treatment groups. Second, treatment effects on the mediator satisfy monotonicity; in our setting, this requires that the treatment weakly reduces the number of students scoring zero on the decoding test in every school, which is the direction in which the treatment operates on average (-0.69, $p = 0.004$; see below). While this is hard to test, it appears plausible given the observed results. As noted in the main text, these assumptions are weaker than those required for standard mediation analysis (e.g., Imai et al., 2011; Huber, 2020). We implement the method using the R package *TestMechs* (https://github.com/jonathandroth/TestMechs).

Two further conditions are required for the test to be informative. First, the method needs some "always-takers" of the mediator. Because we have one-sided non-compliance, in that no control-group students received the program, some otherwise interesting mediators (e.g., being tutored, having a teacher or reading specialist as tutor) do not satisfy this condition. Second, the mediator must have finite support. This condition also applies to the outcome to obtain point estimates, so we discretize the variables.

We use the method to test whether the literacy skills of low-achieving students fully mediate treatment effects on high-achieving students. High-achieving students are the top 80 percent and low-achieving students the bottom 20 percent of the literacy pre-test (most tutored students are low-achieving). We focus on our main outcome, the decoding test. The mediator is the number of students scoring zero on the decoding test (a finite-support count). Following Kwon and Roth (2026), we discretize the outcome into five bins, and also consider two and ten bins.[6]

A school-level regression shows that the treatment significantly reduces the number of students scoring zero (-0.69, $p = 0.004$), indicating an effect on the mediator. We then exclude low-achieving students and test the sharp null of full mediation using only high-achieving students. We can reject the null with two and ten bins ($p = 0.044$ and $p < 0.001$). With five bins, the p-value is slightly above 0.05 ($p = 0.065$). Thus, improved literacy skills among the lowest-performing students do not appear to be the only channel for the school-level treatment effects.

Table A7 documents the pre-test differences between treatment schools that used teachers, reading specialists, or both as tutors, and between the students those schools

[6] Simulations in Kwon and Roth (2026) show that the method often over-rejects the sharp null with many bins.

tutored. These differences are why we treat the tutor-type comparison in Section 7 as confounded rather than as a clean test of the resource-reallocation and pedagogical-transfer channels. Note that Table A7 uses only the subset of schools that answered our follow-up phone-call after one month of implementation and conducted the national kindergarten test at pre-test, not our full sample. Therefore, the shares of schools using reading specialists, teachers, or both as tutors differ from the shares in Table 3. We use this subset because the comparison of means is then based on the same test.

Table A7: Pre-Test Characteristics by Tutor Type

		Reading specialist		Teacher		Both	
	Control	Tutored	Treatment	Tutored	Treatment	Tutored	Treatment
Language	38.4	27.3	39.03	23.42	42.69	14.29	36.17
	(29.08)	(27.76)	(29.82)	(25.76)	(30.5)	(11.11)	(29.14)
	[1097]	[58]	[347]	[62]	[358]	[11]	[94]
Preliteracy	35.29	22.15	42.38	16.93	38.85	11.91	37.8
	(27.16)	(23.33)	(29.05)	(19.33)	(27.56)	(13.03)	(28.64)
	[1097]	[58]	[347]	[62]	[351]	[11]	[95]

Note. Pre-test means for treatment schools grouped by the type of tutor the school used, with tutored students separated and control-school students for reference. The Treatment averages include the tutored students. Standard deviations are in parentheses and the number of students in brackets. Language and Preliteracy statistics are from the schools that conducted the national kindergarten pre-tests, i.e., not from our full sample. The comparison across tutor types is descriptive, since tutor type was a school's choice rather than a randomized assignment.

References

Andrews, I. and Armstrong, T. B. (2017). Unbiased Instrumental Variables Estimation Under Known First-Stage Sign. *Quantitative Economics*, 8(2):479–503.

Bhuller, M. and Sigstad, H. (2024). 2SLS with Multiple Treatments. *Journal of Econometrics*, 242(1):105785.

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*, https://doi.org/10.1257/rct.11007-1.1.

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.

Borusyak, K. and Hull, P. (2023). Nonrandom Exposure to Exogenous Shocks. *Econometrica*, 91(6):2155–2185.

Chernozhukov, V., Hansen, C., and Wüthrich, K. (2017). Instrumental Variable Quantile Regression. *Handbook of Quantile Regression*, pages 119–143.

Fusejima, K. (2024). Identification of Multi-Valued Treatment Effects with Unobserved Heterogeneity. *Journal of Econometrics*, 238(1):105563.

Huber, M. (2020). Mediation Analysis. *Handbook of Labor, Human Resources and Population Economics*, pages 1–38.

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.

Kang, H. and Keele, L. (2018). Spillover Effects in Cluster Randomized Trials with Noncompliance. Technical Report arXiv: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.

Kraft, M. A., Schueler, B. E., and Falken, G. T. (2026). What Impacts Should We Expect from Tutoring at Scale? Exploring Meta-Analytic Generalizability. *Review of Educational Research*, Advance online publication:https://doi.org/10.3102/00346543261446660.

Kwon, S. and Roth, J. (2026). Testing Mechanisms. *Review of Economic Studies*, forthcoming:10.1093/restud/rdag028.

Lee, D. S. (2009). Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. *Review of Economic Studies*, 76(3):1071–1102.

Nibbering, D. and Oosterveen, M. (2025). Policy-Relevant Causal Effect Estimation Using Instrumental Variables with Interference. Technical Report arXiv:2509.12538, arXiv.

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.

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)*, 30(4):365–395.