Reading Gains Through Teacher Pairs: A Co-Teaching Experiment

Anna Holvio[1, 5, 6], Hannu Karhunen[1, 3, 5, 6], Marja-Kristiina Lerkkanen[2, 6], Oskari Nokso-Koivisto[1], Matti Sarvimäki[1,4,5,6], and Minna Torppa[2, 6]

[1]Aalto University
[2]University of Jyväskylä
[3]The Labour Institute for Economic Research LABORE
[4]VATT Institute for Economic Research
[5]Helsinki GSE
[6]EDUCA Flagship

November 16, 2026

Abstract

We conducted a randomized field experiment with the City of Helsinki to evaluate co-teaching in 105 second-grade classrooms across 34 schools receiving needs-based funding. From September 2024 to April 2025, treatment classrooms received an additional qualified teacher for four 45-minute literacy lessons per week, supported by training in co-teaching and literacy instruction. Control classrooms continued regular instruction without additional resources. The point estimate suggests an average improvement in reading skills of 0.06 standard deviations ($p = 0.148$) in the end of the intervention. Estimated gains were larger among Finnish background students (0.13 standard deviations, $p = 0.026$) and students scoring above the baseline median (0.10 standard deviations, $p = 0.095$), while we find no evidence of gains among immigrant background students or initially poorer readers. By the beginning of the following school year, the average estimated effect was close to zero.

Keywords: co-teaching, student–teacher ratio, literacy instruction, reading skills, randomized controlled trial

JEL Codes: H75, I21, I24, I28, J24

This study is part of the EDUCA Flagship funded by the Research Council of Finland (#358924, #358946). We pre-registered the study at the Open Science Framework (OSF) registry (link) on 19 May 2025.

1 Introduction

Children from disadvantaged backgrounds often attend schools that are less well equipped to meet their students' needs than those attended by their more affluent peers (OECD, 2023). To counter these inequalities, countries and school districts sometimes direct additional funding to schools serving disadvantaged communities. In Helsinki, where we conduct our study, schools receive needs-based funding from both the city and the Finnish central government. In 2024, these two sources together provided €13.2 million for comprehensive education in Helsinki, equivalent to approximately 2.3 percent of the city's budgeted comprehensive education costs.[1] Schools typically use this funding to hire additional teachers and teaching assistants, but differ in how they allocate the resulting support across students (Ramos Lobato, 2023). A central question is how to deploy these additional resources to improve learning and which students benefit from them.

This paper examines one way of using needs-based funding: providing an additional qualified co-teacher to work alongside the classroom teacher. Co-teaching may improve students' academic achievement by creating more favorable conditions for effective instruction. A lower student–teacher ratio may enable improved monitoring of students' learning, more individualized feedback, and more differentiated instruction, while also supporting classroom management and student engagement (Cook and Friend, 1995; Vembye et al., 2024). Furthermore, co-teaching provides opportunities to combine the complementary expertise of the teachers. For example, the classroom teachers can contribute knowledge of curricular content and learning objectives, whereas special education teachers and teachers trained to teach Finnish as a second language (L2) can bring expertise in identifying learning difficulties and adapting instruction to students' individual needs. Their combined expertise may facilitate the identification of learning difficulties and the provision of appropriately targeted support (Scruggs et al., 2007; Friend, 2008).

To evaluate the effects of providing an additional co-teacher, we conducted a randomized field experiment with the City of Helsinki in 105 second-grade classrooms across 34 schools receiving needs-based funding. Within each school, classrooms were assigned to receive additional co-teaching resources or continue regular instruction without additional resources. From September 2024 to April 2025, treatment classrooms received a qualified special education or L2 teacher for four 45-minute literacy lessons per week. Participating classroom teachers and co-teachers also received training in co-teaching practices and literacy instruction. We assessed reading and numeracy skills at baseline, during the final weeks of the intervention, and during the fall term of the following school year, allowing us to examine both immediate effects and their persistence. We also measured reading motivation and self-efficacy in reading during the first and second assessments.

We estimate an average improvement in reading skills of 0.06 standard deviations during the intervention ($p = 0.148$). The point estimates suggest larger gains among initially more advanced readers. Estimated gains were 0.10 standard deviations among students scoring above the baseline median ($p = 0.095$), with no evidence of gains among those below it. A similar

[1] Our calculation combines municipal needs-based funding of €8,836,000 (City of Helsinki, 2024b) with national equity grants of €4,387,669 (City of Helsinki, 2024a). Their sum is approximately 2.3 percent of budgeted comprehensive education costs of €571,100,500 (City of Helsinki, 2023).

pattern emerges when students are grouped by immigrant background, which is closely associated with initial reading skills in our sample. Estimated gains were 0.13 standard deviations among students with a Finnish background, defined as having at least one Finnish-born parent ($p = 0.026$), with no evidence of gains among students with an immigrant background.

The average point estimate was smaller at follow-up. At the September assessment in third grade, the average estimated effect was close to zero, but subgroup estimates continued to differ. The estimated effect was 0.09 standard deviations among students with a Finnish background ($p = 0.057$) and -0.08 among students with an immigrant background ($p = 0.153$). Estimates by baseline reading skills followed the same pattern but were smaller in magnitude, at approximately 0.07 standard deviations for initially more advanced readers ($p = 0.240$) and -0.05 for initially poorer readers ($p = 0.244$). Point estimates on reading motivation and self-efficacy in reading were positive but imprecisely estimated, while effects on numeracy skills were close to zero at both assessments.

These findings contribute to three strands of research. First, our study is most closely related to two earlier Nordic field experiments evaluating similar co-teaching interventions. In Norway, providing an additional teacher during literacy instruction in first- and second-grade classrooms had no detectable average effect on reading achievement, but yielded estimated gains of approximately 0.12 standard deviations for boys and 0.23 standard deviations for boys with low baseline skills (Haaland et al., 2024). In Denmark, adding a co-teacher or teaching assistant increased sixth-grade reading scores by 0.09–0.13 standard deviations, with persistent gains for disadvantaged students (Andersen et al., 2020).[2] Although the subgroup definitions and school starting age differ across countries and studies, our findings point to a different distribution of benefits, with larger estimated gains among initially more advanced readers.

Second, our study adds to research on how schools deploy teaching resources through class-size reductions and changes in instructional arrangements (Schanzenbach, 2020; Blanden et al., 2023). The Tennessee Project STAR experiment found that smaller classes improved achievement, particularly among disadvantaged students, and increased subsequent college attendance, while adding a teacher's aide to a regular-sized class had little effect on achievement (Krueger, 1999; Chetty et al., 2011). Quasi-experimental evidence on class-size reductions is mixed. Studies found achievement gains in Israel and lasting improvements in educational attainment and earnings in Sweden (Angrist and Lavy, 1999; Fredriksson et al., 2013), but no detectable achievement effects in more recent Israeli cohorts or long-run benefits in Norway (Angrist et al., 2019; Leuven and Løkken, 2020). By contrast, our experiment examines whether additional qualified teachers can improve reading skills development when deployed through co-teaching within existing classrooms.

Third, our findings inform the broader debate on the effects of school funding. Reviews of quasi-experimental evidence from the United States find that additional spending improves student outcomes on average and the benefits tend to be larger for economically disadvantaged populations (Jackson, 2020; Jackson and Mackevicius, 2024). Particularly relevant to our set-

[2] A broader meta-analysis reports a modest positive average effect of co-teaching and related collaborative instructional models, but many of the underlying studies have relatively weak research designs (Vembye et al., 2024). Earlier work has also documented substantial variation in how co-teachers share instructional responsibilities (Scruggs et al., 2007).

ting are evaluations of funding allocated according to schools' socioeconomic composition. These studies find improvements in achievement or educational progression in England, Finland, and Sweden (Machin et al., 2010; Silliman, 2017; Rosenqvist and Sauermann, 2025), but no clear improvements in the Netherlands and France (Leuven et al., 2007; Bénabou et al., 2009). In Finland, Silliman (2017) found that additional funding increased continuation into upper-secondary education among low-performing native students and students with an immigrant background. By contrast, in England, Machin et al. (2010) found the largest gains among students with stronger prior achievement, particularly in more disadvantaged schools.

Our contribution is to evaluate a specific use of needs-based funding and examine how its benefits are distributed within recipient schools. The experiment provides evidence on adding qualified co-teachers to existing classrooms, an alternative to using additional staff to reduce class sizes. In contrast to the patterns reported in earlier Nordic experiments and what policymakers might reasonably expect on that basis, we find larger estimated gains among initially more advanced readers. This contrast highlights the potential importance of details of the intervention, its implementation, and the support already available to pupils when drawing lessons across settings.

The remainder of the paper is organized as follows. Section 2 describes the intervention and experimental design. Section 3 describes the data and estimation methods. Section 4 presents the results. Section 5 concludes.

2 Experiment

2.1 Institutional background

Finland's compulsory basic education comprises grades 1 through 9. Children start it in the fall of the year when they turn seven. In grades 1 to 6, a classroom teacher is responsible for teaching nearly all subjects to the same class. In our setting, single-teacher classrooms had a median size of 19 students, ranging from 12 to 25. Municipalities assign students to a local school based on their home address, and most students attend their assigned school rather than apply elsewhere. The school's student body therefore tends to reflect the socioeconomic composition of the neighborhood.

Finland's education system provides support for students with diverse learning needs, including learning difficulties and disabilities. Special schools provide education for some students with disabilities who require more intensive support, but whenever possible, support is provided in general education classrooms. Special education teachers work alongside classroom teachers to provide targeted interventions for students who struggle with reading. Education providers (usually municipalities), schools, and individual teachers decide how to implement, monitor, and evaluate the curriculum and have considerable autonomy in choosing teaching materials and methods. Instructional practices therefore vary across schools and classrooms. Nevertheless, between-school variation in reading proficiency is below international averages in both PISA and PIRLS studies (Hiltunen et al., 2023; Leino et al., 2023).

Children are not expected to read before starting school. However, they enter school with widely varying skills. Teachers therefore need to adapt literacy instruction to different start-

ing levels (Lerkkanen, 2019). Finnish has exceptionally regular correspondences between letters and sounds, which help children acquire decoding skills relatively quickly once formal instruction begins (Soodla et al., 2015). The national core curriculum groups grades 1–2 together under shared learning objectives for Finnish language and literature. Students in these grades have approximately 23 lessons per week, about one-third of which are devoted to literacy instruction. Instruction emphasizes letter-sound correspondence, decoding, reading fluency, and basic comprehension (Leino and Lerkkanen, 2026). By the end of grade 2, students are expected to read independently and with reasonable fluency (Opetushallitus, 2016). Students whose home languages include a language other than Finnish may follow the Finnish-as-a-second-language (L2) curriculum rather than the Finnish-as-a-first-language (L1) curriculum. L2 students generally attend regular classrooms, but may receive some or all of their Finnish instruction in a separate group.

2.2 Intervention

We ran the field experiment in collaboration with the City of Helsinki. Together with the city, we designed an intervention to evaluate the effects of co-teaching on second-grade students' reading skills in schools receiving needs-based funding. This funding targets schools serving disadvantaged populations identified on the basis of socioeconomic circumstances and language background.[3] The funding aims to reduce disparities between schools and is commonly used to hire additional teachers and teaching assistants.

Treatment classrooms received an additional teacher for four 45-minute Finnish language and literacy lessons per week from September 2024 to April 2025. Schools had discretion over the allocation of lesson hours across subjects, and second-grade classrooms had six to eight Finnish language and literacy lessons per week. Thus, the additional teacher was present for roughly half to two-thirds of their weekly instruction in the subject. The additional teachers were qualified special education or Finnish-as-a-second-language (L2) teachers, either already employed at the school or newly hired for the intervention.[4] The intervention therefore both increased the teacher resources available during Finnish language and literacy instruction and brought specialist expertise into the classroom. The city and participating schools jointly financed the intervention through their needs-based funding allocations.

Classroom teachers and co-teachers in the treatment group attended three mandatory one-day training sessions in August 2024, October 2024, and March 2025. Education researchers and teachers with expertise in co-teaching led the sessions, which covered co-teaching practices and literacy instruction. Schools had flexibility in how they implemented co-teaching. However, students following the Finnish-as-a-first-language (L1) and Finnish-as-a-second-language (L2) curricula were required to be taught together during these four lessons per week using co-teaching practices in planning and instruction. Students with additional learning needs could also receive individual support outside these co-teaching lessons.

[3]The municipal funding criteria include household income, parental education and unemployment, and students' registered first language. National equity grants use indicators of educational attainment, unemployment, and language background in the school's surrounding population.

[4]For example, one large school allocated the intervention lessons to four existing teachers and one newly hired teacher. The existing teachers received additional compensation for extra teaching hours, while the newly hired teacher provided the lessons as part of their regular teaching hours.

Control classrooms continued with regular instruction without additional resources from the experiment. In a survey of control group classroom teachers, 88 percent reported that L2 students received Finnish-language instruction partly in separate groups and partly together with L1 students. The remaining 12 percent reported that the two groups were always taught together. Among classrooms with separate instruction, L2 students received an average of 2.2 Finnish-language lessons per week in their own group. Joint instruction of L1 and L2 students was therefore already relatively common in the schools. Under the intervention, however, two teachers were required to plan and deliver this instruction together for four lessons per week.

Co-teaching was also rather common in control classrooms. Across the teacher survey waves, 60 percent of control group classroom teachers consistently reported using some form of co-teaching in Finnish language and literacy instruction, 25 percent consistently reported none, and 15 percent gave different answers across waves. Teachers may have understood co-teaching broadly to include another teacher pulling out a small group for separate instruction, consultation with a special education teacher, or joint lesson planning. The survey responses therefore do not establish whether co-teaching in control classrooms involved the same arrangements or intensity as the intervention. Taken together, these features mean that the experiment estimates the effects of additional co-teaching resources and associated training relative to existing instructional practices.

2.3 Randomization and compliance

The intervention targeted second-grade classrooms in the 37 schools receiving municipal needs-based funding. Three schools withdrew before randomization, leaving 34 participating schools. We excluded 19 classrooms that provided special education, used English or Swedish as the language of instruction, or were located in separate annexes serving students whose backgrounds differed substantially from those targeted by the funding. The randomization sample comprised 95 classrooms with 2,083 students. Randomization took place in June 2024, and the student and classroom counts therefore refer to the end of first grade.

We assigned classrooms to treatment and control groups within schools. To reduce the risk of chance imbalances given the relatively small number of classrooms and to increase statistical power, we used the re-randomization procedure of Morgan and Rubin (2012). We generated 1,000,000 candidate assignments and assessed their balance using the Mahalanobis distance based on three classroom-level characteristics reported by schools: the student–teacher ratio, the share of students with special education needs integrated into the classroom, and the number of teachers. After removing duplicate assignments, we retained the 1,001 assignments with the smallest Mahalanobis distances and added their mirror images, which reverse treatment and control status for every classroom. We selected the final assignment uniformly at random from the resulting 2,002 assignments.[5] The final allocation comprised 47 treatment and 48 control classrooms.

The assessment data include 105 classrooms, more than the 95 randomized. This discrepancy arises because nine classrooms were recorded as a single combined classroom in the randomiza-

[5]In schools with an odd number of classrooms, the allocation allowed one more classroom in either treatment or control while maintaining overall balance. We also use this set of assignments to compute exact Fisher randomization test p-values, as described in Section 3.2.

tion data but appear as separate classrooms in the assessment data. Students in these split classrooms retain the treatment assignment of the combined classroom that was randomized. The assessment data comprises 53 treatment and 52 control classrooms.

Compliance with treatment assignment was high. Only four classrooms deviated from their assignments. These deviations occurred in two schools, each of which provided the intervention to a classroom assigned to the control group instead of one assigned to treatment. We could not establish whether these switches reflected administrative errors or deliberate decisions. The affected classrooms contained 71 students, of whom 62 are included in the estimation sample described in the next section.

3 Data and empirical approach

3.1 Data and measurement

We assessed students' reading and numeracy skills in September 2024, early in second grade; in March 2025, near the end of the intervention; and in September 2025, during the fall term of third grade. Students completed the digital assessments on the ViLLE platform over two 45-minute lessons (Laakso et al., 2018). They received audio instructions through headphones and completed a short practice task before each assessment task. The second-grade assessments used tasks from FUNA-Junior, while the third-grade assessment used more age-appropriate tasks from FUNA covering the same skill domains (FUNA Consortium, 2024). We also collected student self-assessments of motivation and self-efficacy in reading and mathematics, together with teacher evaluations of each student. Appendix B describes the assessment battery, and Appendix C reports response rates by outcome and assessment wave.

Our primary outcome is a reading skills index that combines tasks assessing several dimensions of literacy. The second-grade assessments cover word recognition, reading fluency, morphology, reading comprehension, dictation, and vocabulary. Following the pre-analysis plan (Holvio et al., 2025), we exclude responses with unusually long response times and those completed in less than 350 milliseconds. We score each task by the number of correct answers and exclude task scores with an accuracy rate below 65 percent to reduce measurement error from guessing.[6] The index is the average of the standardized task scores. Factor analysis supports treating these tasks as measures of a common underlying skill (Appendix D). We standardize second-grade scores relative to the control group at baseline and third-grade scores relative to the control group at follow-up. The follow-up assessment omits dictation and vocabulary and uses more difficult items in the remaining skill domains.

Statistics Finland linked the assessment results to its register data and administrative records from the City of Helsinki. The register data provide information on students' sex, first language, date and country of birth, and their parents' income, education, and occupation. The city's records identify schools and classrooms, treatment status, Finnish-language curriculum, special education status, and absences. We conducted all analyses using pseudonymized data in Statistics Finland's secure remote access environment.

[6]The threshold comes from the binomial distribution. With two-option items, the probability of *guessing* 65 percent or more of the items correctly is approximately 0.05.

3.2 Estimation

We estimate intent-to-treat (ITT) effects of being assigned to the intervention with the following model:

$$Y_{ics} = \beta_0 + \beta_1 Treatment_c + \boldsymbol{X}_i'\beta_2 + \mu_s + \varepsilon_{ics}, \quad (1)$$

where Y_{ics} is the outcome for student i in classroom c at school s, $Treatment_c \in \{0, 1\}$ indicates the classroom treatment status, and μ_s are school fixed effects. The control vector $\boldsymbol{X_i}$ includes baseline skills and prespecified controls for students' sex, first language, age at assessment, and immigrant background; parental income percentile, years of education, and first language; and a single-parent indicator.

We examine treatment effect heterogeneity by immigrant background, baseline reading skills, and sex. The pre-analysis plan identifies immigrant background as the primary dimension of heterogeneity (Holvio et al., 2025). We define students as having an immigrant background if both parents, or the only known parent, were born abroad. Students with at least one Finnish-born parent are classified as having a Finnish background. We evaluate heterogeneity by baseline reading skills by splitting students into two subgroups based on the median.

We estimate subgroup heterogeneity with the following model:

$$Y_{ics} = \gamma_0 + \gamma_1 Treatment_c + \gamma_2 Z_i + \gamma_3 \left(Treatment_c \times Z_i\right) + \boldsymbol{X}_i'\gamma_4 + \left(Z_i \times \boldsymbol{X}_i\right)'\gamma_5 + \mu_s + \varepsilon_{ics}, \quad (2)$$

where Z_i is the subgroup indicator, γ_1 is the treatment effect for the reference group ($Z_i = 0$), $\gamma_1 + \gamma_3$ is the treatment effect for the comparison group ($Z_i = 1$), and γ_3 measures the difference in treatment effects across the two groups. If Z_i is one of the covariates already included in the vector $\boldsymbol{X_i}$, we exclude it from the covariates. Covariates are fully saturated with the subgroup indicator ($Z_i \times \boldsymbol{X}_i$).

Our primary inference relies on exact Fisher randomization test p-values (Fisher, 1935; Young, 2019). We construct the reference distribution from the 2,002 assignments eligible for selection under our randomization procedure. The two-sided p-value is the proportion of assignments yielding a t-statistic at least as large in absolute value as the observed statistic. These tests evaluate the sharp null that the intervention had no effect on any student. We also report standard errors clustered at the classroom level.

3.3 Sample and balance

Our target population consists of 2,003 students, of whom 1,002 were assigned to treatment and 1,001 to control. Appendix C reports response rates by outcome and assessment wave, and sample attrition. Reading skills data are missing for 6.9 percent of students near the end of the intervention and 14.4 percent at follow-up. Missing assessments primarily concern individual students; two entire classrooms did not complete the end-of-intervention assessments. Near the end of the intervention, response rates were 95 percent in the treatment group and 91 percent in the control group (Table C4). At follow-up, the corresponding rates were 86 and 85 percent.

Our main estimation sample includes students with reading skills measured in both September 2024 and March 2025 and complete register data for the covariates. For the third-grade follow-up, we require reading skills measured in September 2024 and September 2025 and com-

Table 1: Baseline balance and sample descriptives

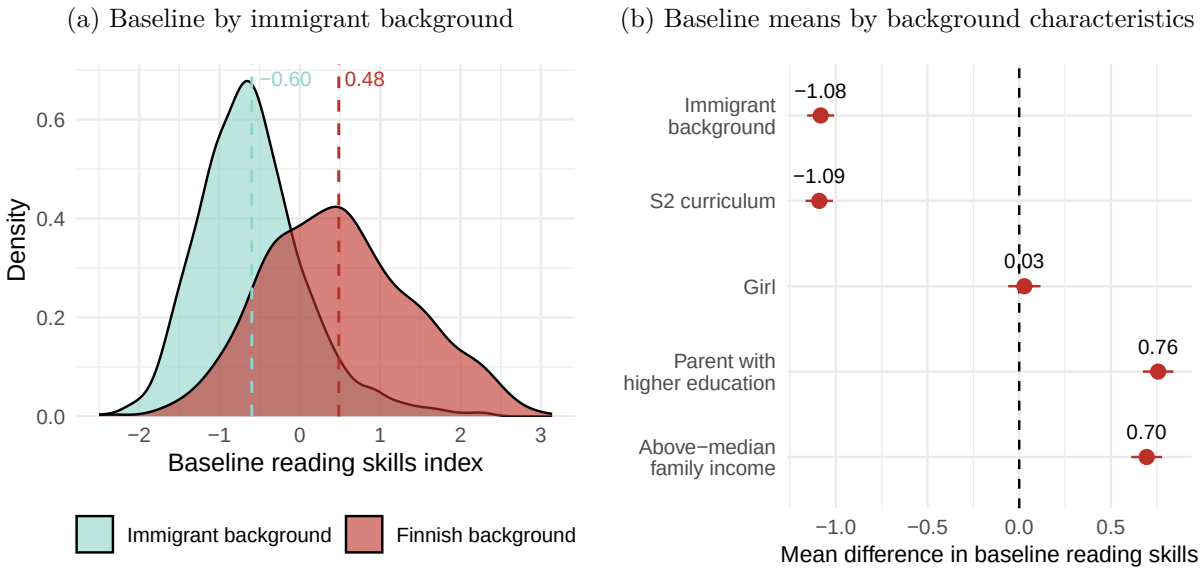
	Estimation sample			Population (2016 cohort)	
	Treatment ($N = 863$)	Control ($N = 828$)	Diff	Helsinki ($N = 6,421$)	Finland ($N = 55,926$)
Girl	0.54	0.50	0.04	0.49	0.49
Finnish-speaking	0.61	0.61	0.01	0.74	0.86
S2 student	0.39	0.39	−0.01	–	–
Immigrant background	0.43	0.44	0.00	0.27	0.14
Born in Finland	0.87	0.89	−0.01	0.90	0.93
Intensified or special support	0.25	0.24	0.01	–	–
Single-parent household	0.31	0.29	0.02	0.25	0.21
Mother Finnish-speaking	0.55	0.54	0.00	0.69	0.84
Father Finnish-speaking	0.51	0.53	−0.02	0.68	0.84
Mother's years of education	12.87	12.96	−0.08	14.12	13.68
Father's years of education	12.93	13.00	−0.07	14.02	13.30
Family income (euros)	60,066	60,095	−29	99,410	80,410
Baseline reading skills index	0.00	0.06	−0.06	–	–
Baseline numeracy skills index	0.01	0.01	0.01	–	–
Baseline motivation index	0.11	0.01	0.09*	–	–
Baseline self-efficacy index	0.05	0.02	0.03	–	–
Asymptotic p-value	–	–	0.18	–	–
Exact p-value	–	–	0.36	–	–

Note: The table reports means for the main estimation sample by randomized treatment assignment. SD denotes standard deviation, and Diff reports treatment–control differences. Binary variables are reported as proportions. S2 denotes the Finnish-as-a-second-language curriculum. Immigrant background is defined as both parents, or the only known parent, having been born abroad. Family income reports the average of past 3 years. Baseline outcome indices are standardized using the control group baseline distribution. Helsinki and Finland report population means for the full 2016 birth cohort, for comparison; a dash indicates no population figure is available for that row. The final two rows report asymptotic and exact randomization inference p-values for a joint test of baseline balance. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

plete covariate data. Figure C1 shows exclusions from the main estimation sample separately by treatment assignment. Of the 2,003 students in the target population, we retain 1,691 (84.4 percent) in the final sample, comprising 863 students assigned to treatment (86.1 percent of those allocated) and 828 to control (82.7 percent of those allocated). The rate is slightly higher in the treatment group, but the difference is not statistically significant (Table C5; $p = 0.331$). Several observable background characteristics are associated with attrition. Immigrant background, Finnish-as-a-second language curriculum, lower parental education, and lower reading skills predict exclusion from the estimation sample (Table C5). However, these associations are similar across treatment and control groups.

Table 1 reports baseline characteristics for the main estimation sample, and shows a comparison to the same cohort of children in Helsinki and Finland as a whole. Approximately 44 percent of students have an immigrant background, while 12 percent were born outside Finland. About 39 percent follow the Finnish-as-a-second-language curriculum (S2), and one-quarter receive intensified or special support for learning. The treatment and control groups have similar observed demographic and socioeconomic characteristics. A joint test of baseline balance yields

Figure 1: Baseline reading skills



Notes: Panel (a) shows the distribution of baseline reading skills index by immigrant background. Dashed lines mark the group means. Students are classified as having an immigrant background if both parents, or the only known parent, were born abroad. The comparison group comprises students with at least one Finnish-born parent. Panel (b) shows unadjusted mean differences and their 95% confidence intervals accounting for classroom-level clustering. Higher education refers to bachelor's degree or higher. Comparison groups are students with at least one Finnish-born parent, students who follow the first-language (S1) curriculum, boys, students with no parents with a higher education degree, and students from families with below-median income. Reading scores are standardized using the control group baseline distribution.

a randomization inference p-value of 0.36. At baseline, the treatment group's reading skills index is 0.06 standard deviations lower and its reading motivation index is 0.09 standard deviations higher than those of the control group. Our preferred specification controls for the corresponding baseline outcome to improve precision and account for these differences.

Relative to Helsinki and Finland as a whole, the sample is noticeably more disadvantaged. The share of students with an immigrant background is over one-and-a-half times the Helsinki average (27 percent) and three times the national average (14 percent). Family income in the sample is around 60 percent of the Helsinki average (€99,410) and 75 percent of the national average (€80,410). These comparisons are consistent with the schools' needs-based funding status.

Figure 1 describes differences in baseline reading skills across student backgrounds. Panel 1a shows the distributions by immigrant background, our pre-registered primary dimension of heterogeneity. Students with an immigrant background scored an average of 1.08 standard deviations lower than those with at least one Finnish-born parent. They account for 72 percent of students below the baseline median and 21 percent of those above it. The distributions nevertheless overlap, with students from both backgrounds represented among initially weaker and stronger readers. Panel 1b places this difference alongside mean differences by Finnish-language curriculum, sex, parental education, and family income. The gap by Finnish curriculum (1.09 standard deviations) is almost identical in magnitude to the immigrant-background gap. Baseline reading skills do not differ by sex in our sample. Both parental education and family income are strongly associated with baseline reading skills: students with at least one parent

that has a higher education degree score 0.76 standard deviations higher than those without, and students from above-median income families score 0.70 standard deviations higher than students from below-median income families.

4 Results

4.1 Effects during the intervention

Figure 2 presents our main results. The top row shows that students assigned to the treatment group scored an estimated 0.06 standard deviations higher in reading than those assigned to the control group, though this estimate is not statistically significant at conventional levels (exact randomization p-value = 0.148). For comparison, average reading scores in the control group increased by 0.38 standard deviations between baseline and the spring assessment. The point estimate therefore corresponds to approximately one-sixth of the improvement observed in the control group over the same period.

The estimates by baseline reading skills in Figure 2 suggest larger gains among initially more advanced readers. The estimated effect was 0.10 standard deviations for students scoring above the baseline median, compared with 0.01 for those below it ($p = 0.211$ for the difference). As Figure 1a shows, a majority, but not all, of the students below the median had an immigrant background. A similar pattern emerges when we group students by immigrant background, our pre-registered primary dimension of heterogeneity. The estimated effect was 0.13 standard deviations among students with a Finnish background and –0.02 among students with an immigrant background ($p = 0.053$ for the difference). The remaining rows of Figure 2 report estimates by sex. Estimated gains were 0.07 standard deviations for girls and 0.05 for boys ($p = 0.68$ for the difference).

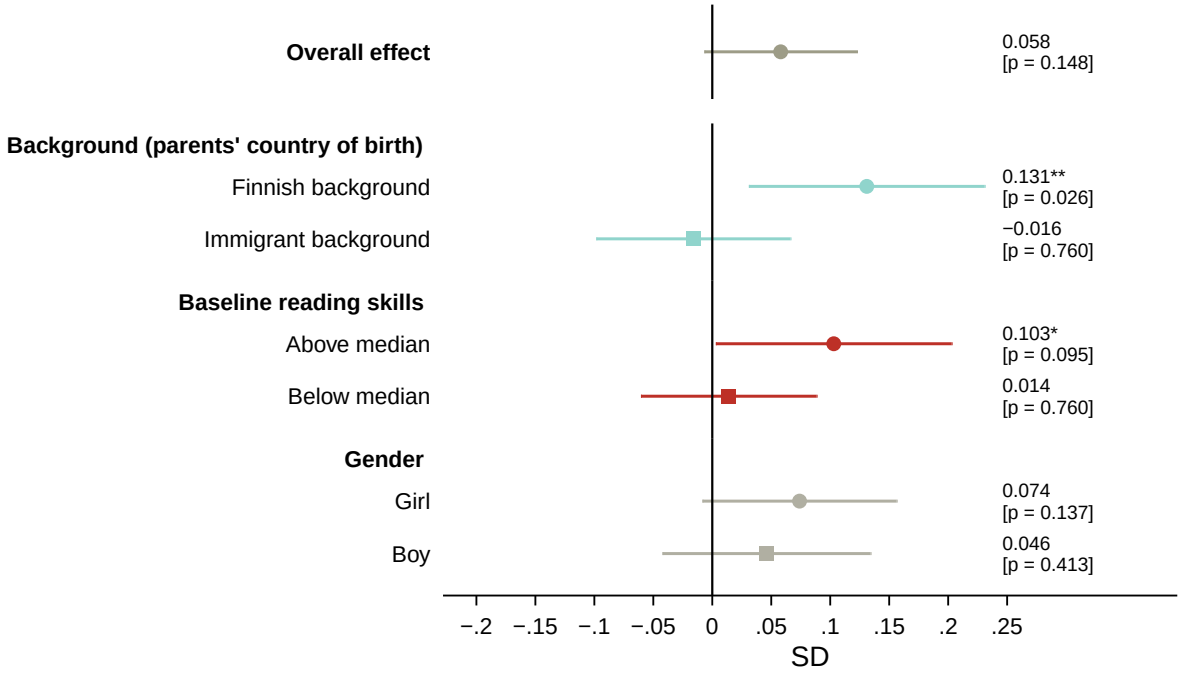
We also examine effects on reading motivation, self-efficacy in reading, and numeracy skills (Appendix Table A1). Estimated average effects on reading motivation and self-efficacy were positive but imprecise, at 0.08 standard deviations ($p = 0.22$) and 0.05 standard deviations ($p = 0.33$), respectively. The effect on motivation was positive for students with a Finnish background and girls, with gains of 0.12 standard deviations ($p = 0.088$) and 0.15 standard deviations ($p = 0.086$), respectively. We observe no statistically significant effects for the other subgroups. The estimated average effect on numeracy skills, which the intervention did not target, was zero ($p = 0.96$).

4.2 Effects after the intervention

Figure 3 reports effects on reading skills in September 2025, during the fall term of third grade and approximately five months after the intervention ended. The estimated average effect was close to zero, at 0.01 standard deviations ($p = 0.708$), compared with 0.06 standard deviations during the intervention.

Subgroup estimates continued to differ after the intervention. The estimated effect was 0.09 standard deviations among students with a Finnish background and –0.08 for students with an immigrant background ($p = 0.014$ for the difference). Among students with above-median

Figure 2: Effects of co-teaching on reading skills in spring of second grade



Note: The figure reports intent-to-treat estimates of the effects of co-teaching on reading skills in spring of second grade. The top row reports the overall effect (Equation 1). The remaining rows report subgroup estimates (Equation 2). Reading scores are standardized relative to the control group's baseline distribution. Confidence intervals are calculated as the point estimate $\pm 1.96\times$ the standard error clustered at the classroom level. Reported p-values are based on exact randomization inference and need not correspond to these confidence intervals. Immigrant background was pre-specified as the primary dimension of heterogeneity. Appendix Tables A1 and A2 report the corresponding estimates, standard errors and p-values.

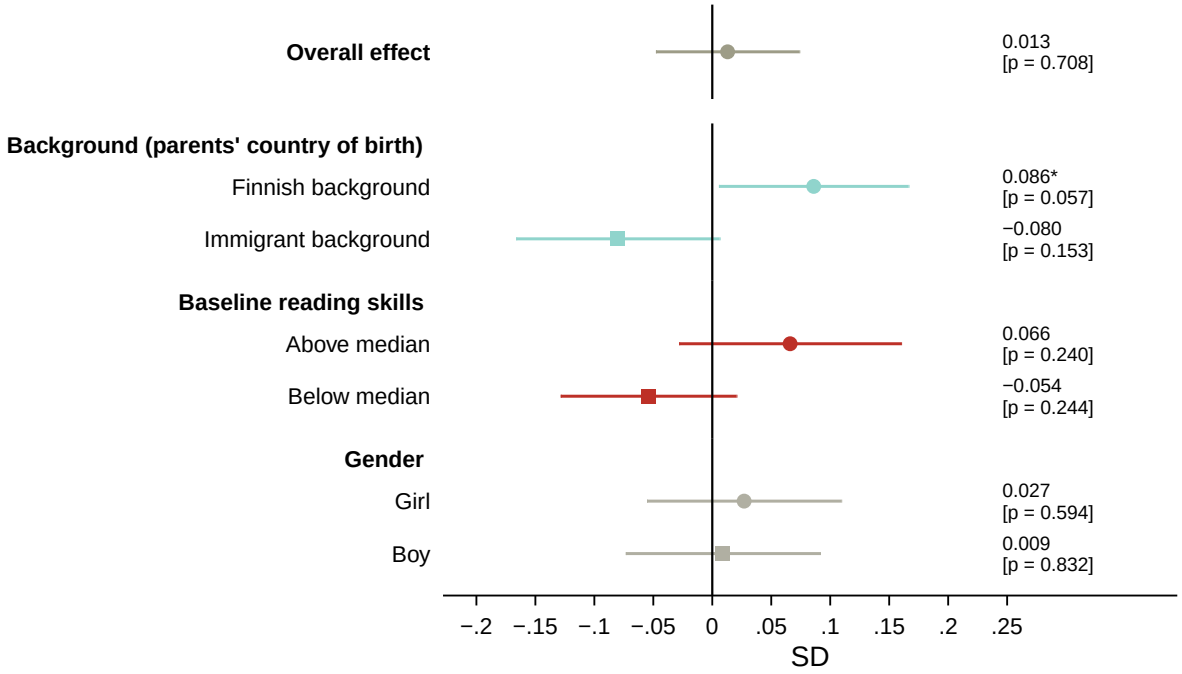
baseline reading skills, the estimated effect was 0.07 standard deviations, compared with –0.05 among students with below-median baseline reading skills ($p = 0.114$ for the difference).

The decline in estimated effects after the intervention may reflect more than the dissipation of initial learning gains. Point estimates shifted downwards across several student groups, while remaining larger among initially stronger readers. For students with an immigrant background, whose estimated effect was already close to zero in spring, this shift resulted in a negative follow-up estimate. Conversations with teachers and administrators suggest that control classrooms may have received additional support after the intervention to compensate for not receiving an additional teacher during the experiment. Such support could have contributed to this pattern by improving subsequent outcomes in the control group. Unfortunately, we lack systematic data on support after the intervention and cannot assess this hypothesis directly.

5 Conclusion

We estimate a positive but imprecise average effect of providing an additional qualified teacher alongside training in co-teaching on reading skills in schools receiving needs-based funding. The estimated gain during the intervention was 0.06 standard deviations, corresponding to approximately one-sixth of the improvement observed in the control group over the same period. The estimates are larger among initially more advanced readers, while we found no evidence of

Figure 3: Effects of co-teaching on reading skills in fall of third grade



Note: The figure reports intent-to-treat estimates of the effects of co-teaching on reading skills in the fall term of third grade. The specification and inference procedures follow Figure 2; see its note for details. Reading scores are standardized relative to the control group's follow-up distribution. Appendix Tables A1 and A2 report the corresponding estimates, standard errors and p-values.

gains among initially poorer readers. The average estimated effect was close to zero at follow-up, but subgroup estimates continued to differ.

The perhaps surprising concentration of gains among initially more advanced readers could arise through several mechanisms. Previous research links lower academic skills to greater teacher-reported individual support and more frequent teacher attention (Chaudhuri et al., 2022). Our working hypothesis is that, when pupils with greater learning needs already receive substantial support, an additional teacher creates more opportunities to challenge and support more advanced readers. In future work, we will explore this hypothesis further using teacher survey data and findings from qualitative interviews about the experiment.

The support already available to initially poorer readers may also help explain the contrast with an earlier Norwegian experiment, which also targeted literacy instruction in the early grades (Haaland et al., 2024). Finland emphasizes early support for pupils with learning difficulties, whereas the prevalence of formal special education in Norway increases substantially in later grades. Although we cannot yet establish what explains the contrasting results, they caution against extrapolating from apparently similar interventions without considering their implementation and the support they supplement. Accumulating evidence across settings is therefore important for understanding both whether additional teacher resources improve learning and who benefits.

More speculatively, benefits for more advanced readers could also have implications for school segregation. If families of initially more advanced students perceive that schools serving disadvantaged communities provide good learning opportunities for their children, they may be more

willing to choose these schools. Co-teaching could therefore help schools attract and retain a broader mix of students, potentially benefiting initially poorer readers indirectly even when their immediate learning gains are limited. Exploring these possibilities would be a promising direction for further research and collaboration between practitioners and researchers. New experiments could also examine whether more targeted instructional support or changes in the intensity and duration of co-teaching would improve outcomes for initially poorer or struggling readers.

References

Andersen, Simon Calmar, Louise Beuchert, Helena Skyt Nielsen, and Mette Kjærgaard Thomsen, "The effect of teacher's aides in the classroom: Evidence from a randomized trial," *Journal of the European Economic Association*, 2020, *18* (1), 469–505.

Angrist, Joshua D. and Victor Lavy, "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *The Quarterly Journal of Economics*, 1999, *114* (2), 533–575.

_ , _ , Jetson Leder-Luis, and Adi Shany, "Maimonides' Rule Redux," *American Economic Review: Insights*, 2019, *1* (3), 309–324.

Bénabou, Roland, Francis Kramarz, and Corinne Prost, "The French Zones d'éducation Prioritaire: Much Ado about Nothing?," *Economics of Education Review*, 2009, *28* (3), 345–356.

Blanden, Jo, Matthias Doepke, and Jan Stuhler, "Educational Inequality," in Eric A. Hanushek, Ludger Woessmann, and Stephen J. Machin, eds., *Handbook of the Economics of Education*, Vol. 6, Elsevier, 2023, pp. 405–497.

Chaudhuri, S., H. Muhonen, E. Pakarinen, and M.-K. Lerkkanen, "Teachers' Visual Focus of Attention in Relation to Students' Basic Academic Skills and Teachers' Individual Support for Students: An Eye-Tracking Study," *Learning and Individual Differences*, 2022, *98*, 102179.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan, "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR," *The Quarterly Journal of Economics*, 2011, *126* (4), 1593–1660.

City of Helsinki, "Helsingin kaupungin talousarvio 2024 ja taloussuunnitelma 2024–2026," Helsingin kaupungin keskushallinnon julkaisuja 2023:40, City of Helsinki, Helsinki 2023. Approved by the City Council on 22 November 2023. Comprehensive education costs on p. 124.

_ , "Perusopetusjohtajan päätös koulutuksellisen tasa-arvon edistämiseen myönnetyn erityisavustuksen jakamisesta 2024," January 2024. Director of Basic Education decision, 16 January, section 12, HEL 2023-001787.

_ , "Tarveperusteisen rahoituksen jakaminen varhaiskasvatukseen ja perusopetukseen 1.8.2024 alkaen," January 2024. Education Division Executive Director decision, 30 January, section 8, HEL 2024-001216.

Cook, Lynne and Marilyn Friend, "Co-Teaching: Guidelines for Creating Effective Practices," *Focus on Exceptional Children*, Dec. 1995, *28* (3).

Fisher, Ronald Aylmer, *The Design of Experiments*, Edinburgh: Oliver and Boyd, 1935.

Fredriksson, Peter, Björn Öckert, and Hessel Oosterbeek, "Long-term effects of class size," *The Quarterly Journal of Economics*, 2013, *128* (1), 249–285.

Friend, Marilyn, "Co-Teaching: A Simple Solution That Isn't Simple After All," *Journal of Curriculum and Instruction*, 01 2008, pp. 9–19.

FUNA Consortium, *FUNA – Functional Numeracy Assessment Test Battery & Reading and Writing Test Battery [Finnish version]*, Turku Research Institute for Learning Analytics, University of Turku, 2024.

Haaland, Venke Furre, Mari Rege, and Oddny Judith Solheim, "Do students learn more with an additional teacher in the classroom? Evidence from a field experiment," *The Economic Journal*, 2024, *134* (657), 418–435.

Hiltunen, J., A. Ahonen, N. Hienonen, H. Kauppinen, J. Kotila, P. Lehtola, K. Leino, M. Lintuvuori, K. Nissinen, E. Puhakka, M. Sirén, M.-P. Vainikainen, and J. Vettenranta, "PISA 2022 ensituloksia [First Results of PISA 2022]," Opetus- ja kulttuuriministeriön julkaisuja 2023:49, Ministry of Education and Culture, Helsinki 2023.

Holvio, Anna, Matti Sarvimäki, Marja-Kristiina Lerkkanen, Minna Torppa, Hannu Karhunen, and Oskari Nokso-Koivisto, *Pre-Analysis Plan: Reading Gains Through Teacher Pairs: A Co-Teaching Experiment*, OSF, May 2025.

Jackson, C. Kirabo, "Does School Spending Matter? The New Literature on an Old Question," in Laura Tach, Rachel Dunifon, and Douglas L. Miller, eds., *Confronting Inequality: How Policies and Practices Shape Children's Opportunities*, American Psychological Association, 2020, pp. 165–186.

_ **and Claire L. Mackevicius**, "What Impacts Can We Expect from School Spending Policy? Evidence from Evaluations in the United States," *American Economic Journal: Applied Economics*, 2024, *16* (1), 412–446.

Krueger, Alan B, "Experimental estimates of education production functions," *The Quarterly Journal of Economics*, 1999, *114* (2), 497–532.

Laakso, Mikko-Jussi, Erkki Kaila, and Teemu Rajala, "ViLLE –collaborative education tool: Designing and utilizing an exercise-based learning environment," *Education and Information Technologies*, 2018, *23* (4), 1655–1676.

Leino, K. and M.-K. Lerkkanen, "Literacy Teaching and Learning in Finland," in N. Flynn, H. Joseph, P. Orellana García, D. Powell, and W. Slater, eds., *The Bloomsbury Handbook of Literacy*, Bloomsbury, 2026, pp. 1–47. In press.

_ , M. Sirén, K. Nissinen, and E. Puhakka, "Puoli tuntia lukemista: kansainvälinen lasten lukutaitotutkimus (PIRLS 2021) [Reading for Half an Hour: Progress in International Reading Literacy Study 2021]," Koulutuksen tutkimuslaitos: Tutkimuksia 37, Finnish Institute for Educational Research, University of Jyväskylä, Jyväskylä 2023.

Lerkkanen, M.-K., "Early Language and Literacy Development in the Finnish Context," in D. Whitebread, V. Grau, K. Kumpulainen, M. M. McClelland, N. E. Perry, and D. Pino-Pasternak, eds., *The SAGE Handbook of Developmental Psychology and Early Childhood Education*, SAGE, 2019, pp. 403–417.

Leuven, Edwin and Sturla A Løkken, "Long-term impacts of class size in compulsory school," *Journal of Human Resources*, 2020, *55* (1), 309–348.

_ , Mikael Lindahl, Hessel Oosterbeek, and Dinand Webbink, "The Effect of Extra Funding for Disadvantaged Pupils on Achievement," *The Review of Economics and Statistics*, 2007, *89* (4), 721–736.

Lobato, Isabel Ramos, "Koulusegregaatio ja tarveperusteinen rahoitus: miten Helsingin kouluissa käytetään lisämäärärahoja ja mitä haasteita kouluissa tunnistetaan?," *Kvartti*, May 2023. Published 22 May.

Machin, Stephen, Sandra McNally, and Costas Meghir, "Resources and Standards in Urban Schools," *Journal of Human Capital*, 2010, *4* (4), 365–393.

Morgan, Kari Lock and Donald B. Rubin, "Rerandomization to improve covariate balance in experiments," *The Annals of Statistics*, 2012, *40* (2), 1263–1282.

OECD, *Equity and Inclusion in Education: Finding Strength through Diversity*, Paris: OECD Publishing, 2023.

Opetushallitus, *National core curriculum for basic education 2014* Publications 2016:5, Helsinki: Finnish National Board of Education, 2016.

Rosenqvist, Olof and Jan Sauermann, "The Effects of Increasing Compensatory Resource Allocation on Student Achievement: Evidence from the Equity Grant in the Swedish Compulsory School," Working Paper 2025:15, Institute for Evaluation of Labour Market and Education Policy (IFAU) 2025.

Schanzenbach, Diane Whitmore, "The Economics of Class Size," in Steve Bradley and Colin Green, eds., *The Economics of Education: A Comprehensive Overview*, 2 ed., Academic Press, 2020, pp. 321–331.

Scruggs, Thomas E., Margo A. Mastropieri, and Kimberly A. McDuffie, "Co-Teaching in Inclusive Classrooms: A Metasynthesis of Qualitative Research," *Exceptional Children*, 2007, *73* (4), 392–416.

Silliman, Mikko, "Targeted Funding, Immigrant Background, and Educational Outcomes: Evidence from Helsinki's "Positive Discrimination" Policy," VATT Working Papers 91, VATT Institute for Economic Research, Helsinki 2017.

Soodla, P., M.-K. Lerkkanen, P. Niemi, E. Kikas, G. Silinskas, and J.-E. Nurmi, "Does Early Reading Instruction Promote the Rate of Acquisition? A Comparison of Two Transparent Orthographies," *Learning and Instruction*, 2015, *38*, 14–23.

Vembye, Mikkel H., Felix Weiss, and Bethany H. Bhat, "The Effects of Co-Teaching and Related Collaborative Models of Instruction on Student Achievement: A Systematic Review and Meta-Analysis," *Review of Educational Research*, 2024, *94* (3), 376–422.

Young, Alwyn, "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results," *The Quarterly Journal of Economics*, 2019, *134* (2), 557–598.

Zhang, Xiao, Jari-Erik Nurmi, Noona Kiuru, Marja-Kristiina Lerkkanen, and Kaisa Aunola, "A teacher-report measure of children's task-avoidant behavior: A validation study of the Behavioral Strategy Rating Scale," *Learning and Individual Differences*, 2011, *21* (6), 690–698.

A Supplementary tables and figures

Table A1: Intent-to-treat effects on reading skills, motivation, self-efficacy, and numeracy

Outcome	End of intervention (Grade 2 Spring)	Follow-up (Grade 3 Fall)
Reading skills index	0.058	0.013
	(0.033)	(0.031)
	[$p = 0.148$]	[$p = 0.708$]
Reading motivation	0.077	—
	(0.048)	
	[$p = 0.222$]	
Self-efficacy in reading	0.052	—
	(0.046)	
	[$p = 0.326$]	
Numeracy skills index	0.002	−0.005
	(0.037)	(0.043)
	[$p = 0.958$]	[$p = 0.918$]
School fixed effects	✓	✓
Student controls	✓	✓
Baseline index	✓	✓
Observations	1,691	1,572
Number of clusters	103	102

Note: Intent-to-treat estimates of β_1 from Equation (1). Grade 2 outcomes are standardized relative to control group baseline, and grade 3 outcomes relative to control group follow-up. Motivation and self-efficacy were measured in the fall and spring of second grade, but not at the third-grade follow-up. Controls: student sex, first language, age at assessment, and immigrant status; parental income percentile and years of education; single-parent indicator. Observations refer to the reading skills estimation sample. Standard errors clustered at the classroom level in parentheses; exact randomization inference p-values in brackets (Young, 2019). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A2: Heterogeneity in reading skills effects by subgroup

Subgroup	End of intervention effect	Follow-up effect
Background (parents' country of birth)		
Finnish background	0.131**	0.086*
	(0.051)	(0.041)
	[$p = 0.026$]	[$p = 0.057$]
Immigrant background	−0.016	−0.080
	(0.042)	(0.044)
	[$p = 0.760$]	[$p = 0.153$]
Difference	$p = 0.053$	$p = 0.014$
Baseline reading skills		
Above median	0.103*	0.066
	(0.051)	(0.048)
	[$p = 0.095$]	[$p = 0.240$]
Below median	0.014	−0.054
	(0.038)	(0.038)
	[$p = 0.760$]	[$p = 0.244$]
Difference	$p = 0.211$	$p = 0.114$
Gender		
Girls	0.074	0.027
	(0.042)	(0.042)
	[$p = 0.137$]	[$p = 0.594$]
Boys	0.046	0.009
	(0.045)	(0.042)
	[$p = 0.413$]	[$p = 0.832$]
Difference	$p = 0.673$	$p = 0.779$

Note: ITT estimates from interaction regressions based on Equation (2). Grade 2 reading skills index is standardized relative to control group baseline, and grade 3 reading skills index relative to control group follow-up. Standard errors clustered at the classroom level; exact randomization inference p-values in brackets. Immigrant status was pre-registered (PAP section 5.4) as the primary heterogeneity dimension. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A3: Heterogeneity in reading motivation and self-efficacy effects by subgroup

Subgroup	Motivation effect	Self-efficacy effect
Background (parents' country of birth)		
Finnish background	0.119*	0.090
	(0.053)	(0.055)
	[$p = 0.088$]	[$p = 0.164$]
Immigrant background	−0.003	−0.031
	(0.075)	(0.084)
	[$p = 0.980$]	[$p = 0.763$]
Difference	$p = 0.272$	$p = 0.344$
Baseline reading skills		
Above median	0.053	0.104
	(0.065)	(0.055)
	[$p = 0.523$]	[$p = 0.112$]
Below median	0.101	0.009
	(0.079)	(0.087)
	[$p = 0.336$]	[$p = 0.929$]
Difference	$p = 0.739$	$p = 0.433$
Gender		
Girls	0.146*	0.039
	(0.066)	(0.059)
	[$p = 0.086$]	[$p = 0.573$]
Boys	−0.005	0.083
	(0.071)	(0.069)
	[$p = 0.955$]	[$p = 0.307$]
Difference	$p = 0.163$	$p = 0.647$

Note: ITT estimates from interaction regressions based on Equation (2). Motivation and self-efficacy were measured in the fall and spring of second grade, but not at the third-grade follow-up. Indices are standardized relative to control group baseline. Standard errors clustered at the classroom level; exact randomization inference p-values in brackets. Immigrant status was pre-registered (PAP section 5.4) as the primary heterogeneity dimension. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

B Description of the student assessment

The literacy and numeracy sections consist of tasks from the FUNA-Junior assessment tool, targeted to children aged 6–8; the Grade 3 assessment uses the FUNA tool (ages 9–15), which measures the same skills with more age-appropriate items (FUNA Consortium, 2024).

Literacy section

1. **Phonological word recognition.** The student sees an image and four words and selects the word matching the image; distractors (one real word, two pseudowords) are phonologically

similar. Up to 140 items in two minutes.

2. **Reading fluency.** The student reads a simple sentence and judges whether it is true or false (e.g. "Pigs can fly"). Up to 100 items in 1.5 minutes.

3. **Dictation.** The student hears and writes down a pseudoword that follows Finnish phonotactics. Eight items, 3.5 minutes; ends after two incorrect responses. *(Grade 2 only.)*

4. ~~**Vocabulary.** The student reads a word and judges whether it is a real Finnish word. 41 items. Versions are harmonized via item response theory using three common anchor items.~~ *(Dropped from data. Task had multiple versions, and students completed one of them. Around 25 percent of students lacked common anchor items used to harmonize the score across different versions.)*

5. **Morphology 1 (inflection).** The student completes a sentence containing a pseudoword by choosing the correctly inflected form from three options. Nine items, four minutes.

6. **Morphology 2 (derivation).** As above, but the options are derivations of the pseudoword. Nine items, four minutes.

7. **Morphology 3 (compound word).** The student splits a compound word into its component stems (two to four stems). Nine items, three minutes.

8. **Reading comprehension 1 (statements).** The student reads a 1–4 sentence text and selects the best-fitting statement from three options. 30 items, three minutes.

9. ~~**Reading comprehension 2 (text).** The student reads a one-paragraph text (which then disappears) and answers multiple-choice questions. Six items, five minutes; ends after four incorrect responses.~~ *(Dropped from data. Task was being piloted, and deemed too difficult for the age.)*

10. **Peabody Picture Vocabulary Test (PPVT).** The student hears a word, sees four pictures, and selects the matching picture. 30 items, six minutes. *(Grade 2 only.)*

Numeracy section

Correspondence of integers and quantities; comparison of integers; number series; addition; subtraction; series reasoning; matrix reasoning; quantitative reasoning. Item counts and time limits are as in the pre-analysis plan.

Student self-assessment

Motivation and self-efficacy in reading and mathematics, on 5-point (facial expressions) and 4-point ("I am sure I cannot", "I maybe can", "I believe I can", "I am sure I can") scales respectively.

Motivation

1. How fun is reading for you at school?
2. How fun is reading for you at home?
3. How fun is doing math tasks for you at school?
4. How fun is doing math tasks for you at home?
5. How good are you at reading?
6. How good are you at math?

Self-efficacy in reading and mathematics

1. How sure are you that you can read the subtitles of a movie?
2. How sure are you that you can read a comic book?
3. How sure are you that you can read a book?
4. How sure are you that you can correctly add up the prices of your candy purchases?
5. How sure are you that you can say what time it will be in two and a half hours?
6. How sure are you that you can correctly solve the most difficult math problems in your math book?

Teacher evaluation

For each student: task-avoidant behavior (Behavioral Strategy Rating Scale; Zhang et al., 2011), individual support given in different areas of literacy, and family support for schooling.

C Response rates and attrition

Table C4: Response rates

Outcome	Round	Treatment ($N = 1,002$)	Control ($N = 1,001$)
Reading skills	Grade 2 Fall	0.96	0.97
	Grade 2 Spring	0.95	0.91
	Grade 3 Fall	0.86	0.85
Numeracy	Grade 2 Fall	0.91	0.94
	Grade 2 Spring	0.91	0.88
	Grade 3 Fall	0.86	0.86
Motivation	Grade 2 Fall	0.77	0.81
	Grade 2 Spring	0.84	0.82
Self-efficacy	Grade 2 Fall	0.77	0.81
	Grade 2 Spring	0.83	0.82

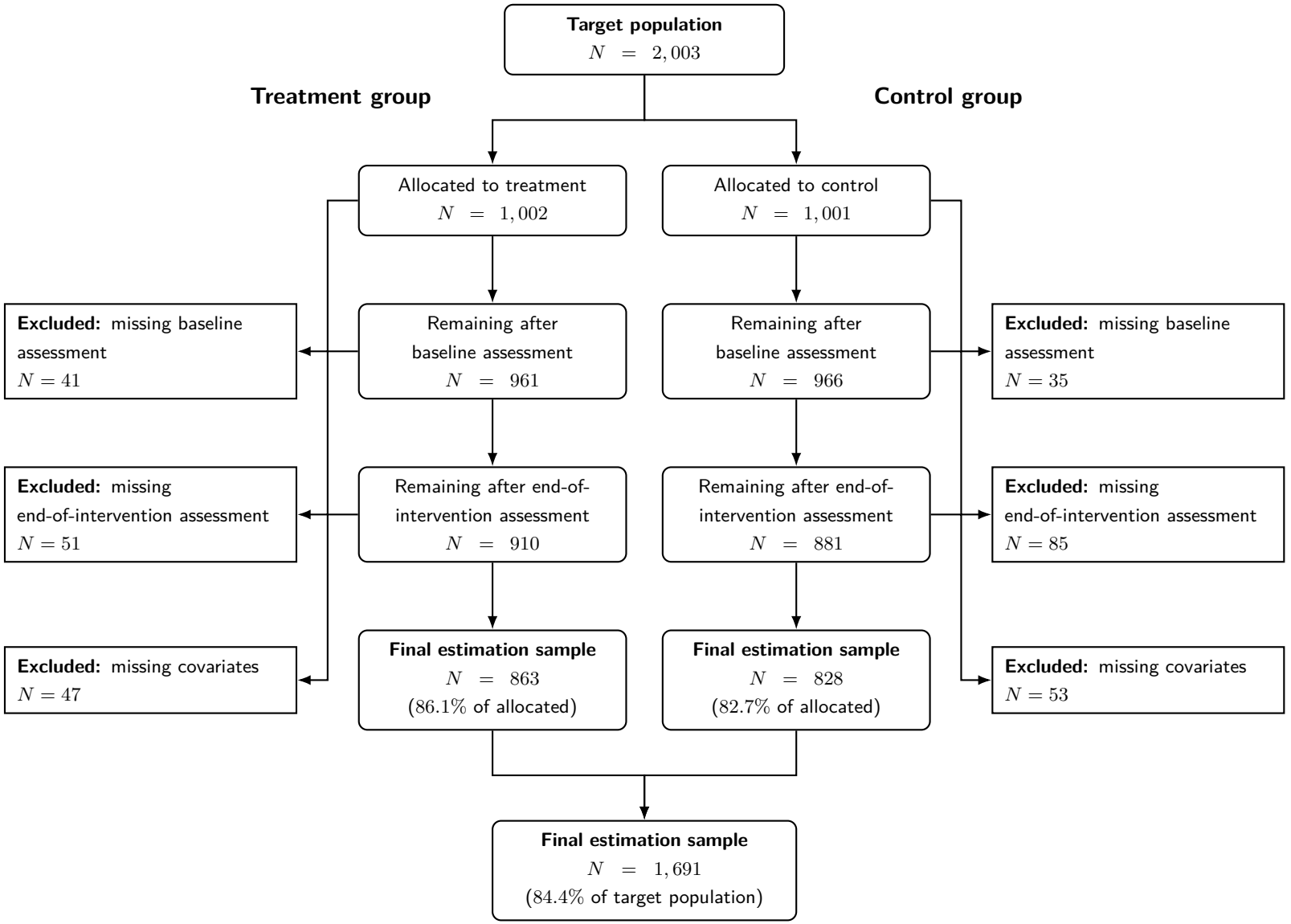
Note: Motivation and self-efficacy were only measured at two points, in the fall and spring of grade 2.

Table C5: Determinants of estimation sample attrition

	Attrition		Attrition × Treat		
Variable	Estimate	SE	Estimate	SE	N
Treatment assignment	−0.019	(0.020)	–	–	2,003
Girl	0.026	(0.016)	−0.019	(0.031)	1,958
Finnish-speaker	−0.090***	(0.018)	0.018	(0.033)	1,927
S2 curriculum	0.122***	(0.018)	−0.021	(0.034)	1,957
Immigrant	0.115***	(0.017)	−0.020	(0.032)	1,956
Born in Finland	−0.288***	(0.032)	0.098	(0.062)	1,958
Learning support	−0.003	(0.022)	0.013	(0.044)	1,958
Single-parent	0.028*	(0.017)	−0.020	(0.032)	1,859
Mother Finnish-speaker	−0.017	(0.014)	−0.002	(0.025)	1,859
Father Finnish-speaker	−0.032**	(0.014)	−0.004	(0.024)	1,859
Mother's years of education	−0.004**	(0.002)	0.005	(0.004)	1,859
Father's years of education	−0.003**	(0.001)	0.001	(0.002)	1,859
Family income percentile	−0.067***	(0.025)	0.049	(0.045)	1,859
Baseline reading skills	−0.047***	(0.008)	−0.001	(0.016)	1,927
End-of-intervention reading skills	−0.052***	(0.008)	0.028*	(0.016)	1,864

Note: Each row reports two separate regressions of an indicator for exclusion from the end-of-intervention (Grade 2 Spring) estimation sample on the listed characteristic. The estimation sample requires the baseline and end-of-intervention reading skills indices and all covariates to be observed. "Attrition" regresses the exclusion indicator on the characteristic alone, and "Attrition × Treat" regresses it on the characteristic, treatment assignment, and their interaction. Table reports the interaction coefficient, which tests whether the characteristic predicts attrition differently by treatment status. Reading skills indices are standardized relative to control group baseline. Both specifications include school fixed effects. Standard errors clustered at the classroom level. Significance based on conventional clustered standard errors (asymptotic p-values). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Figure C1: Sample attrition



D Psychometric properties of the reading skills index

Table D6: Number-of-factors criteria

		Retain?			
k	Eigenvalue	Kaiser (>1)	Scree plot	Parallel analysis	MAP
Panel A: Grade 2 Fall					
1	4.08	Yes	Yes	Yes	**0.035**
2	0.90	No	No	Yes	0.073
3	0.74	No	No	Yes	0.113
4	0.60	No	No	No	0.170
Panel B: Grade 2 Spring					
1	4.42	Yes	Yes	Yes	**0.035**
2	0.80	No	No	Yes	0.075
3	0.65	No	No	Yes	0.115
4	0.62	No	No	No	0.195
Panel C: Grade 3 Fall					
1	3.40	Yes	Yes	Yes	**0.051**
2	0.79	No	No	Yes	0.131
3	0.67	No	No	No	0.292
4	0.46	No	No	No	0.418

Table D7: Cronbach's α

Round	N (items)	Cronbach's α	Average inter-item correlation
Grade 2 Fall	8	0.79	0.43
Grade 2 Spring	8	0.82	0.48
Grade 3 Fall	6	0.73	0.47

Table D8: Exploratory factor analysis loadings

Variable	Reading skills index	Communality
Panel A: Grade 2 Fall		
Phonological word recognition	0.82	0.67
Reading fluency	0.82	0.68
Dictation	0.63	0.4
Morphology 1 (inflection)	0.71	0.5
Morphology 2 (derivation)	0.69	0.48
Morphology 3 (compound word)	0.57	0.33
Reading comprehension (statements)	0.38	0.14
Vocabulary (PPVT)	0.61	0.38
% Var	0.45	
Panel B: Grade 2 Spring		
Phonological word recognition	0.79	0.62
Reading fluency	0.83	0.69
Dictation	0.64	0.42
Morphology 1 (inflection)	0.76	0.58
Morphology 2 (derivation)	0.76	0.58
Morphology 3 (compound word)	0.61	0.38
Reading comprehension (statements)	0.52	0.27
Vocabulary (PPVT)	0.66	0.44
% Var	0.49	
Panel C: Grade 3 Fall		
Phonological word recognition	0.8	0.64
Reading fluency	0.87	0.75
Morphology 1 (inflection)	0.7	0.49
Morphology 2 (derivation)	0.72	0.52
Morphology 3 (compound word)	0.52	0.27
Reading comprehension (statements)	0.51	0.26
% Var	0.49	

Table D9: Confirmatory factor analysis fit measures

Measure	Grade 2 Fall	Grade 2 Spring	Grade 3 Fall
TLI	0.852	0.847	0.910
CFI	0.895	0.891	0.946
SRMR	0.067	0.059	0.050
RMSEA	0.121	0.131	0.115
RMSEA (lower)	0.104	0.117	0.098
RMSEA (upper)	0.138	0.145	0.133
χ^2	161.545	305.872	132.430
df	20	20	9
p	0.000	0.000	0.000

Table D10: Intraclass correlation coefficients

Level	Grade 2 Fall	Grade 2 Spring	Grade 3 Fall
Classroom	0.023	0.016	0.027
School	0.061	0.082	0.051