WORKING PAPER 2024:6

The Boost for reading

Helena Holmlund
Josefin Häggblom
Erica Lindahl

IFAU
INSTITUTE FOR
EVALUATION OF
LABOUR MARKET AND
EDUCATION POLICY

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

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

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

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

ISSN 1651-1166

The Boost for reading

Effects on classroom practices and student outcomes[a]

by

Helena Holmlund,[b] Josefin Häggblom[c] and Erica Lindahl[d]

February 15, 2024

Abstract

We evaluate the "Boost for Reading", an in-service training program for teachers aimed at improving the teaching of literacy and boosting students' reading and writing proficiency. The program provides research summaries about teaching strategies as a basis for group-based discussion, lesson preparations and evaluations under the supervision of a coach. The program was rolled out across Swedish compulsory schools in school years 2015/16–2017/18. We analyze the effects of the intervention using a staggered difference-in-differences strategy excluding treated schools as controls. We find that in lower secondary school, the program shifted the teaching towards a stronger focus on "reading strategies" and raised student test scores in the Swedish language, social study subjects, and science studies by on average 2–5 percent of a standard deviation, respectively. However, we find no effects on teaching practices at stage 1, and accordingly, no effects on the youngest students' test scores.

Keywords: teacher training, professional development, literacy
JEL-codes: I20, I28

[a] We thank Björn Öckert, Sandra McNelly, Chris Karbownik, Caroline Hall as well as seminar participants at the Institute for Evaluation of Labour Market and Education Policy (IFAU) in Uppsala and at the IFAU and UCLS workshop on Education and Human Capital 2023 for valuable comments. We also thank Swedish National Agency for Education for collaborating in the collection of teacher survey data.

[b] IFAU and Uppsala Center for Labor Studies (UCLS).

[c] IFAU; e-mail: josefin.haggblom@ifau.uu.se

[d] IFAU and Uppsala Center for Labor Studies (UCLS); e-mail: erica.lindahl@ifau.uu.se

Table of contents

1 Introduction

The teacher profession is key to successful education. Many studies convincingly show that teacher quality matters and can improve both students' test scores and non-cognitive outcomes (Rockoff 2004; Chetty et al., 2014; Liu and Loeb, 2019; Guryan et al., 2021). Education policies targeting the teaching profession therefore hold the promise of improving students' learning outcomes, but is hindered by the fact that apart from experience, little is known about the determinants of teacher effectiveness (Burroughs et al., 2019; Wiswall, 2013; Hanushek, 2011; Leigh, 2010; Rockoff, 2004). Moreover, policies that affect teacher education and recruitment today will only have a marginal effect in the short term, since the stock of existing teachers will be unaffected by such policies.

An alternative policy that has shown to be beneficial is teacher professional development, in the form of in-service training, coaching, collective lesson preparations, peer-to-peer observation and the like (see e.g. Angrist and Lavy, 2001; Grönqvist et al., 2021; and Briole and Maurin, 2022). When summarizing high quality studies of professional development programs, Yoon et al. (2007) find positive effects on student achievement. However, there are also examples of studies that do not find positive effects of teacher training: see for example Jacob and Lefgren (2004) and (Murphy et al., 2021).

In this paper, we contribute to the literature on teacher professional development by evaluating the "Boost for Reading" program[1], an in-service training program targeting teachers working in Swedish compulsory schools in school years 2015/16–2017/18. The purpose of the program was to improve teachers' proficiency in teaching reading and writing, with the ultimate goal of boosting students' literacy. The model followed the "Boost for Mathematics", a similar program implemented in Swedish schools a few years before the Boost for Reading and evaluated by Grönqvist et al. (2021). In both the Boost for Mathematics and the Boost for Reading, teachers participated in the program during one academic year and the format was inspired by "Lesson study" – collegial lesson preparations and evaluations under the supervision of an experienced coach.[2] However, while the Boost for Mathematics focused on mathematics and was directed towards teachers in this subject only, the Boost for Reading aimed to boost students' literacy in reading and writing, and teachers in all subjects were invited to participate in the program. Thus, from the Boost for Reading we can learn about the importance of literacy for the performance in general, not only for the performance in one subject. From a policy

[1] "Läslyftet" also referred to as "The Literacy Boost" by, for example, The Swedish Agency of Education (Skolverket).

[2] See e.g. Chen and Zhang (2019) for a description of Lesson study.

perspective it is also informative to compare the results from the two programs to provide insights into how to boost student performance in an efficient way.[3] Since both the implementation and to some extent the content of the programs are similar, we follow the evaluation strategy used by Grönqvist et al., (2021). The paper exploits the staggered roll-out of the program over the school years 2015/16–2017/18 in a difference-in-differences (DID) framework. Since there are many untreated schools, we can construct a control group that consists of never-treated schools as suggested by Callaway and Sant'Anna (2021) and Sun and Abraham (2021). The identification assumption is supported by results from balancing tests and tests for parallel pre-treatment trends between the control- and the treatment groups.

During the roll-out of the program, we conducted a yearly teacher survey that was sent to about 5,000 teachers in a representative sample of schools. We follow the schools as they go into treatment or remain in the control group and estimate the DID of treatment on (self-reported) participation in in-service training, collaboration and lesson preparations with colleagues, and teaching practices in the classroom. The results from the teacher survey show that the program was implemented in line with its intentions: Teachers received a boost in in-service training in didactics and methods for teaching literacy. This was achieved through collegial collaboration and through the support of a coach during the implementation year. We observe that teachers in higher grades – who on average had lower skills in teaching literacy to start with – self-reported an improvement in their knowledge and skills after training. Among teachers in grades 7–9, we also find some evidence (in two out of eight outcomes studied) of changed teaching practices that align with a stronger focus on teaching literacy. We did not find any indication of changed teaching practices among teachers in lower grades (grades 1–6).

With respect to student outcomes, we can observe national standardized test scores in Swedish and mathematics for students in all educational stages (grades 3, 6 and 9), while we can observe test scores in social study subjects and science studies only in grade 9. We find that the overall effect of the Boost for Reading on test scores in the Swedish language is positive but small: 1.3 percent of a standard deviation when evaluated over test scores in grade 3, 6 and 9. In general, the estimated effects tend to be slightly larger among students in higher grades. The effect increases to 2 percent of a standard deviation in Swedish when observed in grade 9 only. In mathematics, the corresponding results are statistically insignificant, but in line with those in Swedish. In social study subjects and in science studies, the estimated effects are larger: 4.6 and 4.7 percent of a standard deviation, respectively.

[3] In this context it is interesting to note that the costs of the programs differ: In the Boost for Mathematics, the government grants covered the costs for coaches (20% of full time) and part of the costs for the participating teachers. In the Boost for Reading, the government grants only covered the costs for coaches.

Professional development programs can contain a variety of elements. Several papers show positive effects of peer observation, which is a common theme in Lesson study (Murphy et al., 2021; Burgess et al., 2021; Taylor and Tyler 2012)[4]. However, the Boost for Reading program did not emphasize peer observation but instead consisted of joint lesson planning and evaluation in the presence of a coach, using material such as research summaries and teaching strategies provided by the program. In a meta-analysis of causal studies of effects of coaching, Kraft et al. (2018) find large positive effects on student outcomes. We also know from the literature that it is possible to influence incumbent teachers by providing stricter teaching guidelines in terms of curriculum content and pedagogical practices. For example, Machin and McNally (2008) show that guidelines for teaching literacy had positive effects on students' reading proficiency, and Jackson and Makarin (2018) find that teachers' use of high quality instruction material has positive effects on student test scores. Our results are therefore in line with these studies, which show that coaching and access to didactic support material are successful components of teachers' professional development. We also conclude that the Lesson study model can provide positive effects without peer-to-peer observation.

The paper is organized as follows: section 2 describes the program and the institutional setting. The data and treatment assignment are presented in section 3. Section 4 presents the empirical strategy, section 5 presents the findings from the teacher survey, and section Effects on student test scores presents the effects on student outcomes. Finally, section 7 concludes the paper.

2 The Boost for Reading

The background to the "The Boost for Reading" program was a negative trend in literacy among Swedish students observed in international comparative studies. Both PIRLS 2011 (testing 10-year-olds) and PISA 2012 (testing 15-year-olds) showed declining proficiency in reading and writing among Swedish students. The decline was particularly noticeable in their understanding of informational text. Moreover, teacher surveys indicated that teachers in Sweden taught "reading strategies" – that is, targeted strategies to decode and comprehend the meaning of a text – less frequently compared to their international counterparts (Skolverket, 2012; Mullis et al., 2017). In response to this concerning development, the government launched the Boost for Reading program with the purpose to enhance students' literacy by improving teachers' skills through a large-scale in-service training program for teachers. The National Agency for Education (NAE) was given the responsibility of running the program (Government bill

[4] Murphy, Weinhardt, and Wyness (2021) find positive effects for large schools (with more than one class per grade) and negative effects for small schools (with less than one class per grade).

U2013/7215/S), and it was launched in the academic year 2015/16.[5] The NAE offers the intervention to teachers in all subjects, with the motivation that all teachers have a responsibility for developing students' literacy. Initially, the program focused on compulsory schoolteachers teaching grades 1–9. After a year, pre-school, high school, school librarians and special education schools were also invited to take part in the program.[6]

The NAE based the model on research on "professional learning and development" by Helen Timperley (Timperley et al., 2007) and a review of "collaborative continuing professional development" by Cordingley et al. (2003).[7] The program largely resembles the "lesson study" approach (see Chen and Zhang, 2019) although it does not emphasize peer-to-peer observation. The NAE provides information on its website on how to implement the Boost for Reading program among teachers at the school level. In the following we summarize this information as it was presented in 2015/16.[8] The program was based on frequent and structured group sessions under the supervision of an external coach. Each semester, the group worked with a content "module" over a span of 16 weeks. During a school year each teacher-group worked with two modules.

Figure A1 and Figure A2 in Appendix illustrate how the sessions were organized: each module consisted of eight 2-week parts, where each part included four elements. Element A involved individual preparation, where teachers studied a popular writing research summary. Element B was a group session where teachers and coach jointly discussed the research material and planned a lesson. Element C was the actual teaching, and finally, element D was a group session involving discussion and evaluation of the teaching and lesson plan, supervised by the coach. Coaches were selected based on their qualifications and experience, requiring them to be certified Swedish teachers with a minimum of four years of experience. In parallel with coaching, they also participated in training led by the NAE. Typically, the coach would lead a group consisting of 6–10 teachers, who could be teaching the same or different subjects. The couch could be one of the colleagues at the same school or from another one.

The "modules" were developed in collaboration with educationalists at Swedish universities and peer reviewed by researchers from a different university. Modules targeted different subject areas and grade levels. Examples are "Discussion of texts (grade 1–9)"; "Reading strategies for

[5] In 2014/15 the NAE started a pilot in 32 randomly selected schools. There were a limited number of modules available for these schools to choose from, implying that the program was not as comprehensive as when it was rolled out on a larger scale. One year later, in 2015/16, the Boost for Reading was rolled out on a larger scale and continued in the following years. The focus of this paper is on the school years 2015/16 – 2017/18, and the pilot schools are not used in the analysis.

[6] Nowadays the program is only available to teachers in preschool and in grades 1–3.

[7] Cordingley et al. (2003) do not select studies for the review based on high methodological evaluation standards. In fact, none of the studies were based on randomized controlled trials, and most studies reported correlations between professional development and outcomes.

[8] https://www.skolverket.se/skolutveckling/kurser-och-utbildningar/laslyftet-i-skolan.

informational text (grade 4–9)" and "Promote student learning in Science (grade 4–9)". The modules are publicly available on the NAE website and they were available to all teachers, regardless of officially participating in the Boost for Reading program.[9] Over the course of the program, more modules became available, and by June 2018, there were 29 modules to choose from. Although we do not have information on which modules teachers have used, download statistics from the NAE show that modules with content more generally related to teaching literacy have been downloaded much more frequently than specific modules such as "Literacy in mathematics" or "Literacy for students with Swedish as a second language".[10]

As Sweden has a decentralized education system where the school authority is at the local municipality level or is an independent school provider, the state offered state grants to encourage municipalities and schools to participate. Funding was distributed to regions in proportion to student numbers. Within regions, funding was distributed to school providers in proportion to the amount each school provider had applied for.[11] The grants covered the salary costs for coaches, corresponding to about 10–20 percent (depending on number of teachers the coach is mentoring) of a full-time equivalent teacher wage. The grants did not provide additional funding for participating teachers' time. By the end of the school year 2018/19, 25 percent of all teachers in compulsory school had participated in the program, and the total cost in the spring of 2020 was SEK 640 million (about EUR 58 million at the time) (Skolverket, 2020). Teachers were expected to spend about 60–80 hours on the Boost for Reading during one academic year. This should be done within the contracted "professional development time" of 104 hours (or 11.4 full days) per year, according to their collective agreement.[12] However, surveys show that teachers in Sweden spend less time on training than their contracted hours – about 5 days per year in primary school (grades 1–6) and 4.2 days in lower secondary school (grades 7–9) (Kirsten, 2020). The Boost for Reading is therefore likely to have implied an increase in the total amount of training – but it is also likely that it crowded out other types of training. We will investigate this issue in section 5 where we analyze the teacher questionnaire.

Finally, a short description of the Swedish school system is necessary. We focus on compulsory education, spanning grades 1–9. Schools are typically organized around three stages, which end with standardized tests in grade 3, 6 and 9. At the lower stages (grades 1–3), students typically have a class teacher who teaches most subjects. In grades 4–6, there is more

[9] The official website today offers a large amount of (didactic support) material: https://larportalen.skolverket.se/#/moduler/5-las-skriv/alla/alla.

[10] Download statistics provided by the NAE.

[11] The last year the program existed with state funding in all grades of compulsory school was in 2019/20. In 2018/19, only a very small number of compulsory schoolteachers participated, and we exclude these schools from our analysis. See Section Definition of treatment for more details on the definition of treatment.

[12] The amount of professional development time is stipulated by the main agreement ("Huvudöverenskommelse 21") between the teachers' unions and the Swedish Association of Local Authorities and Regions.

variation across schools and the same teacher may not necessarily cover all core subjects (mathematics, Swedish and English). In contrast, in grades 7–9 (lower secondary school), teachers are "subject teachers" who are specialists in a field or in a combination of fields. The Boost for Reading was also typically organized within a stage at the school.

Grade retention is very uncommon, and students are moved up to the next grade regardless of their performance in individual subjects. After completing compulsory school, students who have failed subjects may retake them within the high school system.

3 Data and descriptive statistics

The paper builds on data from several sources; we combine information on participating teachers with school- and student-level data from Swedish administrative registers as well as survey data on the teachers. Our analysis data builds on the pupil register, which is merged with the results from national standardized tests taken in grades 3, 6 and 9. We also retrieve information on parental background (earnings, education level and immigration background) using family links in the multi-generation register. To arrive at our estimation data, we follow the strategy in Grönqvist et al. (2021) in their evaluation of the Boost for Mathematics. We sample students in the beginning of their stage (grade 1, 4 and 5) and assign treatment based on their expected school at the end of the cycle (grade 3, 6 and 9).[13] The sampling years included in our study are 2010–2016, and the study outcomes are from years 2013–2019.[14]

3.1 Definition of treatment

The NAE required that schools receiving a state grant reported back on the identity of participating teachers. We acquired these data on teachers from the NAE and merged them with the teacher register for the years 2015/16–2017/18. By combining the treated teachers with the full universe of teachers and schools, we can identify 816, 588 and 406 schools with teachers participating in the program at stage 1, 2 and 3, that is in grades 1–3, 4–6 and 7–9, respectively. We use these data to compute descriptive statistics that illustrate the nature of the intervention, which also serve as a basis for how we define treated stages within schools. It's important to note that in our data, we cannot link teachers to the specific classes and students that they teach.

[13] Students in stage 3 are sampled in grade 5 and assigned to the expected grade 9 school given their fifth-grade school assignment. We sample already in grade 5 since school choice is common in stage 3 and we want to avoid endogenous school choices in response to the intervention. In section 4 this is discussed further, and we also show in Appendix Table A2 that there is a high correlation between assigned and actual treatment status, suggesting that endogenous school choice is a minor problem.

[14] Unfortunately, we are not able to follow students after 2019 since national tests were suspended in 2020 and 2021 due to the COVID 19-pandemic.

We can only merge teachers and students at the school level, and treatment is therefore defined at the school-stage level. [15]

It was at the discretion of the headmaster to decide which teachers should participate, and how to target the intervention (Swedish or other subject teachers). Table 1 shows that among participating teachers, the vast majority at stage 1 and 2 teach the Swedish language. This reflects the generalist "class teacher" system at lower levels where the teacher teaches most subjects. However, at the higher stage (grades 7–9), only 38 percent of the participants teach the Swedish language.

Table 1. Share of participating teachers teaching the Swedish language, in any combination with other subjects

Stage 1(grade 1–3)	Stage 2(grade 4–6)	Stage 3(grade 7–9)
89%	71%	38%

Note: Own calculations based on matching participating teachers with the teacher register.

Figure 1 presents the distribution of the share of participating teachers at treated schools, by stage. The shares are calculated by excluding teachers in non-theoretical subjects. At lower stages, a majority of the teachers in participating schools take part in the program. At stage 3, there is a wider dispersion in terms of participation shares.

In the lower stages, there is little ambiguity in terms of what "treatment" represents: most teachers teach a broad set of subjects and most of the teachers in participating schools take part in the program, while at stage 3 there is more heterogeneity with respect to subject-teachers who participate. In this context, it is not obvious how treatment should be defined. We have reasoned as follows: First, the treatment must be at the school level since we cannot link individual teachers to specific students or classes. Second, we want to have *one* definition since we want to compare the estimated effects of the same treatment across all stages and across subjects. Third, an important feature of the program was to invite all types of teachers to improve their teaching in literacy, and we therefore want the treatment definition to include many different types of teachers. Forth, the treatment definition must be theoretically linked to an outcome that we can observe in the end of all stages. In this context we define a treated school as a school where at least one of the participating teachers is teaching the Swedish language.[16] Schools with participating teachers in other subjects but *not* in Swedish are dropped. With this definition, we end up with 814, 572 and 339 treated schools for stage 1, 2 and 3, respectively.

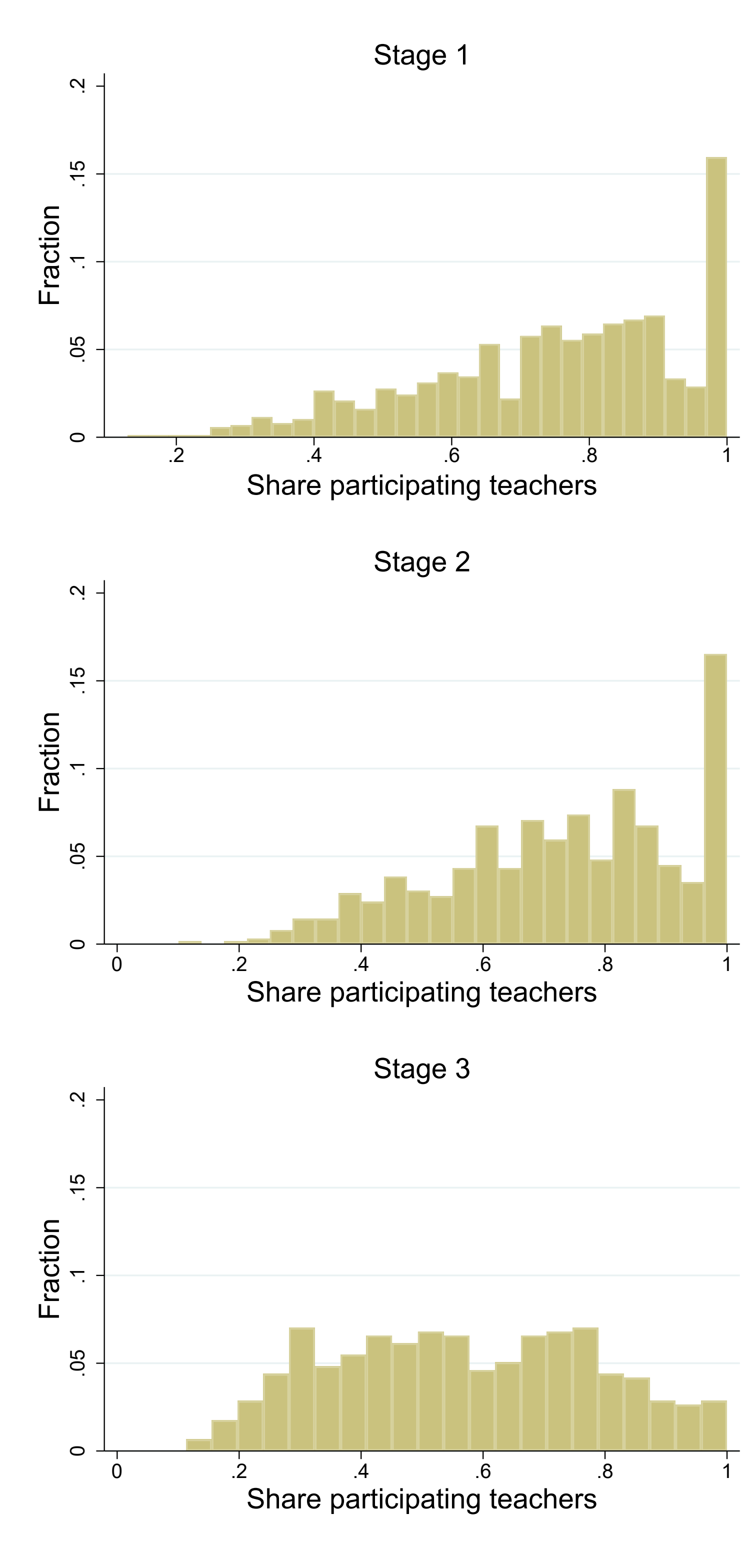
[15] Also note, we cannot rule out endogenous selection of program participation within school. Thus, a treatment definition at class level is not necessarily preferable. Since treatment is defined at the school level, we estimate intention to treat (ITT) estimates.

[16] We have access to outcome variables at all stages in the two core subjects Swedish and mathematics. Since most participating teachers at stage 3 teach Swedish and social study subjects (see Table 2), we have chosen Swedish as our main outcome of interest.

At stage 1 and 2 most teachers in Swedish also teach other subjects, but this is not the case at stage 3 (see Table 1). To learn more about what subjects participating teachers taught at stage 3, we calculate the treatment intensities in different theoretical subjects. Table 2 presents the share of participating teachers at stage 3 by subject, where teachers are weighted by their teaching time in each respective subject, in schools defined as treated. We conclude that the share of teachers who participated in the program (the treatment intensity at the school level) is around 50 percent among all theoretical teachers at the school level, and almost 80 percent if weighted by the number of students in grade 9. However, when we look at specific subjects, the numbers vary. In Swedish and social study subjects the participation rate at the school level is higher, around 60 percent, while the corresponding rate in mathematics and science studies it is lower, around 30-40 percent, respectively.[17]

[17] This is an important difference between this evaluation and the one of the Boost for Mathematics. In the evaluation of the Boost for Mathematics, treated schools are defined as having more than 50 percent (unweighted) of the teachers in mathematics participating in the program.

Figure 1. Distribution of schools' share of participating teachers



Note: Own calculations based on matching participating teachers with the teacher register. Shares are calculated after excluding teachers in non-theoretical subjects. Stage 1 refers to grades 1–3, stage 2 refers to grades 4–6, and stage 3 refers to grades 7–9.

Table 2. Share of participating teachers at stage 3 in schools defined as treated

Share of teachers in:	At school level	Weighted by number of students in grade 9 at school
All subjects	50.48	78.39
Swedish	63.00	60.70
Math	34.62	31.73
Social study subjects	63.31	61.56
Science studies	37.90	35.03

Note: Own calculations based on matching participating teachers with the teacher register and the treatment definition at the school level. Shares are calculated using full-time equivalent teacher positions and the percentage teaching time per subject and teaching position among the theoretical subjects. This data is matched to the individual student population implying that the shares are weighted by the number of students in each school and treatment year.

3.2 Outcome variables

In Sweden, students take standardized national tests in Swedish and mathematics at the end of grades 3, 6 and 9. In grade 9, students are randomly selected to take one subject test in one of the sub-subjects within social study subjects and science studies.[18] The tests in grades 6 and 9 are comprehensive and should reflect the student's overall knowledge and skills in the subject. Test results in grade 3, on the other hand, are less informative as they are primarily intended to assess whether the student has passed the pass mark according to the curriculum at this level. This difference in the outcome variables at stage 1 in comparison to at stage 2 and 3 may affect our possibilities to capture the impact of the Boost for Reading at the lowest stage.

The national tests are graded at the local school by teachers using grading templates. A potential caveat is that the grades are "unfair" due to different grading standards at different schools, and that teachers may change grading standards after participating in the Boost for Reading. We cannot directly address these concerns, but in the analysis, we include school-fixed effects taking care of time invariant differences across schools. We also note that teachers are requested not to correct their own students' tests, and that the written tests should be anonymized before the teacher who corrects them receives them.[19] All test scores are standardized by grade and year to mean zero and standard deviation one.[20]

[18] Social study subjects include history, religion, geography and civics and science studies include biology, physics, and chemistry.

[19] Instructions from the National Swedish Agency for Education about how to correct national tests: https://www.skolverket.se/undervisning/grundskolan/nationella-prov-i-grundskolan/genomfora-och-bedoma-prov-i-grundskolan.

[20] Each exam consists of several sub-tests, where the number of sub-tests differs across grades (3, 6 and 9) and over years. In some cases, sub-test scores are reported, but in some cases only one test grade is reported (pass/fail, or on a 4 or 6 graded scale). In these latter cases, we attribute the grade a merit value. We standardize each test results by year in the population of test-takers.

3.3 Descriptive statistics

Table 3 presents descriptive statistics of participating and non-participating schools. Overall, there are no striking differences between never treated and treated schools, nor between the different treatment waves. Treated schools have slightly higher shares of certified teachers, and more experienced teachers. Treated schools are more likely to be situated in urban municipalities, but less likely to be run by independent school providers. In the formal analysis differences in levels between treated and untreated schools are controlled for by school fixed effects. Potential pre-trend differences are discussed in section 6.1.

Table 3. Descriptive statistics in 2014 (pre-treatment)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Never treated		Wave 1 (2015/16)		Wave 2 (2016/17)		Wave 3 (2017/18)	
	Mean	*Std.Dev.*	*Mean*	*Std.Dev.*	*Mean*	*Std.Dev.*	*Mean*	*Std.Dev.*
Share certified teachers	0.843	0.133	0.865	0.108	0.878	0.101	0.881	0.096
Average teacher experience (years)	12.770	3.358	13.030	2.837	13.160	2.661	12.950	3.002
Urban municipality	0.355	0.478	0.492	0.500	0.388	0.487	0.334	0.472
Immigrant student	0.066	0.248	0.058	0.233	0.059	0.236	0.058	0.233
School cohort size	55.430	41.27	74.63	42.68	63.700	37.85	65.940	38.800
Independent school	0.122	0.327	0.085	0.279	0.039	0.194	0.033	0.179
Predicted test scores	−0.003	0.361	0.009	0.359	−0.000	0.344	0.014	0.351
Test scores	−0.006	1.006	0.022	0.981	0.006	0.998	0.019	0.977
Observations	191,927		27,818		26,638		24,821	

Note: The table displays descriptive statistics of school and student characteristics by treatment status. All characteristics are measured in the year 2014, pre-treatment. Predicted test scores constitute a composite measure of student background characteristics, that is, a prediction from a regression of test scores in Swedish on student's gender, immigration status, age at immigration, mother's and father's birth year, mother's and father's income and indicators for parents' highest level of education. Test scores refer to test scores in Swedish, and are standardized to mean zero, standard deviation 1.

3.4 Teacher survey

In collaboration with the Swedish National Agency for Education we have conducted a survey on teacher practices and professional development. A representative sample of 5,000 teachers in 500 schools were followed yearly for four years, starting in the academic year 2014/15. The questionnaire was sent to teachers teaching the Swedish language and/or social study subjects. The survey focused on teaching literacy, but also included questions about in-service training, self-assessments of own teaching skills and collaborations with colleagues. Across the waves, the response rate was about 41–53 percent. Responding teachers from participating and non-participating schools did not show any systematic differences in their reported background characteristics, see Table A1. However, the response rate seems to be somewhat higher among teachers qualified to teach Swedish in participating schools the year the program was introduced at the school. This is not the case the first and second year after implementation, suggesting that the implementation of the program boosted the response rate among qualified teachers in relation to the non-qualified.

4 Empirical strategy

We exploit the staggered implementation of the Boost for Reading and the fact that many schools never participated in the program. We use a differences-in-differences design and follow the recent literature that removes earlier treated cohorts from the control group.[21] We use only never-treated schools as controls in a stacked regression and as such we estimate the effects of the Boost for Reading by comparing changes in outcomes in treated schools, to the corresponding changes in outcomes in schools that never participated.

Treatment is defined over stage (1, 2 and 3) and the corresponding outcomes in grades $g = \{3, 6, 9\}$, school s, and calendar year t. In detail, for each implementation wave {2015/16, 2016/17, 2017/18}, we keep only schools treated in that year, and schools that never participated.[22] We then stack the data for each wave and estimate all possible treatment effects by grade g and event time τ. We estimate the following event-time regression model:

$$y_{istg} = \sum_{g} \sum_{\tau \neq -1} \theta_{\tau g} D_s 1[\tau, g] + \gamma_{sg} + \mu_{gt} + \varepsilon_{igst}$$

where the outcome y of student i, in grade g, school s, calendar year t is regressed on a set of dummy variables (D) indicating the treatment status for a school and year by grade and event-time τ, a school-fixed effect interacted with grade (γ_{gs}) and a year-fixed effect interacted with grade (μ_{gt}). We estimate a set of parameters $\theta_{g\tau}$ and compute the overall treatment effect θ as a weighted sum of the estimated effects over all grades and event time 0, 1, 2 and 3. Weights are defined as the share of treated students that are used to identify that specific parameter, over all treated students: $\vartheta_{g\tau} = \frac{sum(n_{ig\tau})}{sum(N_i)}; N = 1 \text{ if } D_{sg\tau} = 1;\ n_{ig\tau} = 1 \text{ if } D_{sg\tau} 1[g, \tau] = 1.$ Standard errors are clustered at the school level and weighted using the same procedure.

In our data we include outcome data from the school year 2012/2013 and follow outcomes until the school year 2018/19.[23] We present the overall effect θ, but also alternative configurations by event time (θ_τ) and grade (θ_g). In this setting, event time will capture time from the implementation year to the test year. The first year (τ=0) measures the effect on

[21] When earlier treated cohorts are included as controls, effects may be biased in case there are heterogeneous treatment effects across cohorts and over event time (Callaway and Sant'Anna, 2021; Sun and Abraham, 2021).

[22] Note also that the fixed effects are also interacted by wave.

[23] That means that for the first implementation wave (2015/16), we can evaluate effects up until the fourth year after the program, that is, for students who had not yet started the school-stage and therefore were not directly affected during the implementation year but may have benefited if teacher practices were changed on a permanent basis.

students that were in grade 3, 6 or 9 in the year when the program was running.[24] The second year (τ=1) measures the effect on students that were in grade 2, 5 and 8 in the implementation year, and whose test scores are measured in grade 3, 6 and 9. If the program has permanent effects on teachers' classroom practices, this group potentially benefits from two years of treatment duration. If instead teachers do not continue the methods and content provided by the Boost for Reading in the second year, any positive effects from the implementation year might fade out by the time we observe test scores. Lastly, effects may differ by students' age, e.g. if there are sensitive periods for learning certain skills, or if there are dynamic complementarities in skill production (Cunha and Heckman, 2007). In sum, differences in effects by event time can emerge for several reasons: treatment duration, fade-out, and child's age at treatment.

The difference-in-differences analysis is based on the identifying assumption that treated schools would have followed the same outcome trend as control schools, in the absence of treatment. However, this assumption may be violated if schools that exhibit a positive trend (due to e.g., recruitment of high-quality teachers) are more likely to select into the program, or if parents endogenously choose to opt in or out of participating schools. To handle the latter issue, we sample students in the beginning of each stage (grade 1, 4 and 5) and assign treatment based on their expected school in grade 3, 6 and 9. For three out of four treated cohorts, students had already selected schools when the Boost for Reading was implemented, and we avoid the problem of endogenous school selection correlated with treatment status. Our results are therefore reduced form estimates of assigned treatment status. However, "first stage" estimates of actual treatment status (based on the school attended in the test year) on assigned treatment status (based on expected school) show that actual treatment corresponds to expected treatment status for 82 percent of students. The reduced form effects are therefore close to the treatment effect on the treated (see Table A2 for first-stage estimates).[25]

To further address the parallel trends assumption and potential compositional differences, we provide two standard tests. First, we investigate whether treated and untreated schools are on different compositional trends by tracking how students' *expected* test scores evolve in the post-treatment period. Second, we test for parallel trends in the pre-treatment period in an event-study.

[24] The national tests are mainly conducted during the spring semester; thus, we may observe any potential impact of the program already during the implementation year.

[25] Note that the treatment effect of the treated is at the school level. All teachers in a treated school have not participated in the program, and all students in a treated school are hence not necessarily facing a teacher who has participated in the program.

5 Effects on teachers

We start by presenting results from the teacher survey that show how the program was implemented, how it affected teachers' self-assessed competence, and how it changed teaching practices.

We use the stacked regression model presented in section 4 to estimate the effects. Teachers are assigned to the school where they work (i.e., the school for which they answered the survey), so results should be interpreted as average effects among teachers in treated schools in comparison to untreated schools. The panel with teacher survey answers spans four years (observed during the spring in 2015–2018), and we estimate effects for the implementation year and the two subsequent years. The year before implementation defines the baseline. Of particular interest is whether any potential effect shows up immediately after the implementation of the program and how long it lasts.

5.1 Implementation

The results from the survey data confirm and complement the results from a process evaluation based on interviews and surveys among coaches, teachers and headmasters by Carlbaum et al. (2019). On average, participating teachers received about 20 more hours of training in teaching literacy than teachers in control schools during the implementation year. This is shown in below. They also reported a few more hours of training in didactics and subject knowledge in Swedish, and this increase in training continued after the implementation year, suggesting a more long-lasting effect. Importantly, there was no corresponding effect in subject knowledge in social study subjects, thus, the training was in didactics or subject knowledge in Swedish. Moreover, other training was not crowded out; thus, the program was a boost in in-service training.

We also asked if the training in teaching literacy was via "collegial collaboration" and "support through a coach", which are two essential parts of the Boost for Reading program. Participating teachers reported significantly higher levels of training both via collegial collaboration and via support through a coach the implementation year, but not during (the) other years. This is in line with Carlbaum et al.'s (2019) conclusion that the group-based lesson planning and pedagogical development started to fade out when the official program period came to an end. We have also estimated the corresponding effects by grade level, and the magnitudes of the estimated effects are about the same across all stages. From these results, we conclude that the program was implemented in accordance with its intentions.

To further shed light on how the program affected teachers' work, we asked questions about how they collaborate with other teachers, such as whether they plan lessons together with colleagues, or discuss teaching issues etc. Out of nine questions asked, we find one statistically significant effect: participating teachers planned lessons together with colleagues to a higher degree, which is in line with the intentions of the Boost for Reading program (see Table A3 for the average results across all grades for each specific question).[26] It's worth noting that we had a specific question about the prevalence of peer observation, which has been an important part in earlier successful teacher training programs (Yoon et al., 2007) but not an essential part of the Boost for Reading. The estimated effects of the Boost for Reading on this outcome suggest, as expected, that it was not an important part of the Boost for Reading program.

[26] Analyzing the estimated effects by grade level, it turns out that this statistically significant effect on planning lessons together with colleagues is driven by grade 4–6 teachers.

Table 4. Effects of the Boost for Reading on teachers' different training activities

This academic year, how many hours have you participated in in-service training including				
Teaching literacy (hours per school year)				
	Pooled	Implementation year	One year after	Two years after
Effect	7.8059***	19.7455***	3.0491	-2.9961
SE	[0.862]	[1.104]	[1.482]	[1.966]
Observations	19,806	19,806	19,806	19,806
In addition to training in teaching literacy:				
Didactics/methodology or subject knowledge in Swedish (hours per school year)				
Effect	2.288***	3.6267***	2.5569***	1.6911
SE	[0.584]	[0.810]	[0.930]	[1.088]
Observations	17,612	17,612	17,612	17,612
Control group average	6.634			
Didactics/methodology or subject knowledge in social study science (hours per school year)				
Effect	0.3105	0.0035	0.3598	1.5077
SE	[0.477]	[0.593]	[0.713]	[0.980]
Observations	17,592	17,592	17,592	17,592
Control group average	4.504			
Other (hours per school year)				
Effect	0.3448	-2.6749	3.3811	3.2414
SE	[0.930]	[1.106]	[1.613]	[1.669]
Observations	17,371	17,371	17,371	17,371
Control group average	13.458			
Has the training in teaching literacy included:				
Collegial collaboration (yes/no)				
Effect	0.0787***	0.1273***	0.0679	0.0505
SE	[0.0189]	[0.0214]	[0.0318]	[0.0375]
Observations	13,728	13,728	13,728	13,728
Control group average	*0.882*			
Support through supervision (yes/no)				
Effect	0.0969***	0.2172***	0.0126	-0.0294
SE	[0.0306]	[0.0390]	[0.0444]	[0.0680]
Observations	11,807	11,807	11,807	11,807
Control group average	0.378			

Note: The table shows the estimated effects of the Boost for Reading on teachers' self-reported training activities. All models include school-fixed effects interacted with grade and wave and a year fixed effect interacted with grade and wave. Baseline is the average response reported in the control group. Cluster-adjusted standard errors at the school level are in brackets and */**/*** refers to statistical significance at the 10/5/1 percent level.

5.2 Self-assessed competence

To evaluate the effect on teachers' knowledge and skills in teaching literacy, we asked if they believed they had sufficient knowledge and skills to teach literacy. Interestingly, we find a clear increase of about 20 percentage points in the probability of reporting "good" or "very good" (instead of "neutral", "bad" or "very bad") among teachers at stage 2 and 3 (see Table 5 for these results) during the implementation year. Among teachers at stage 1 we do not observe a corresponding increase. Note also that the baseline observed in the control group is higher among teachers in the lower grades, suggesting that they generally assess their knowledge and skills in teaching literacy higher. This conclusion is also supported by survey answers on formal education in teaching literacy. From these, we learn that there are clear differences across stages. As many as 96 percent of all responding teachers at stage 1 had studied didactics and methods for teaching literacy at university level, while the corresponding shares among teacher at stage 2 and 3 are 0.88 and 0.71 (Holmlund et al., 2021). In this context, it is also interesting to note that Carlbaum et al. (2019) conclude that Swedish language teachers with pedagogical training from higher education should already be familiar with the Boost for Reading content. Taken together, the Boost for Reading program especially boosted self-assessed skills among teachers in higher grades, who on average had lower skills in teaching literacy to start with.

Table 5. Self-assessed competence in teaching literacy by grade

To which degree do you think that you have sufficient knowledge and skills to teach literacy?
Dependent variable is the probability to report "good" or "very good" in comparison to "very low", "low" and "neutral".

	Pooled	Implementation year	One year after	Two years after
Grade 1-3				
Effect	−0.0266	-0.0096	−0.0166	−0.0989
SE	[0.0564]	[0.0534]	[0.0695]	[0.103]
Observations	6,362	6,362	6,362	6,362
Control group average	0.8525			
Grade 4-6				
Effect	0.1727	0.2186***	0.1591	0.0656
SE	[0.0704]	[0.0648]	[0.0855]	[0.126]
Observations	5,469	5,469	5,469	5,469
Control group average	0.7532			
Grade 7-9				
Effect	0.1632	0.2358***	0.1849	−0.0265
SE	[0.0711]	[0.0711]	[0.0998]	[0.102]
Observations	7,985	7,985	7,985	7,985
Control group average	0.6153			

Note: The table shows the estimated effects of the Boost for Reading on teachers' self-assessed competence to teach literacy. All models include school-fixed effects interacted with grade and wave and a year fixed effect interacted with grade and wave. The baseline is the average response in the control group. Cluster-adjusted standard errors at the school level are in brackets and */**/*** refers to statistical significance at the 10/5/1 percent level.

5.3 Teaching practices

Finally, we asked the teachers about their teaching practices. It is difficult to derive clear hypotheses from the modules about how the program is expected to affect teaching practices. We chose to ask eight questions that also appear in PIRLS and that have been claimed to be important tools in teaching literacy (Skolforskningsinstitutet, 2019).27 We estimate dynamic effects for all grades separately. It turns out that the estimated effects among teachers at stage 1 and 2 were generally small and not statistically significant. This is in line with the results presented in Table 5, which show that teachers in lower grades assessed a smaller or no change in self-reported competence in teaching literacy after the Boost for Reading program compared to teachers in higher grades. Among teachers at stage 3, however, we do find evidence of changed teaching practices: Table 6 presents positive effects in two out of eight outcomes on teaching practices for stage 3 teachers. After the program, they more often ask the students "to identify the main ideas of what they have read", and to "compare what they have read with other things they have read". The estimated effects on the other outcomes are not statistically significant. We also know from Carlbaum et al. (2019) that teachers participating in the program did not always perform the teaching activity in their ordinary teaching as proposed by element C (see Appendix Figure A2), suggesting that we should not expect full implementation.

Table 6. Estimated effect on teaching practices among teachers at stage 3

How often do you ask the students to do the following in your teaching about and with text?
Dependent variable is the probability that the teacher answered "often" or "very often" in comparison to "never", "rarely" and "sometimes".

	Pooled	Implementation year	One year after	Two years after
Locate information within the text				
Effect	−0.0438	−0.0552	−0.0282	−0.0432
SE	[0.0257]	[0.0281]	[0.0336]	[0.0426]
Observations	7,970	7,970	7,970	7,970
Control group average	0.843	0.843	0.843	0.843
Identify the main ideas of what they have read				
Effect	0.0936***	0.1037***	0.111***	0.0454
SE	[0.0335]	[0.0372]	[0.0399]	[0.0523]
Observations	7,980	7,980	7,980	7,980

[27] More specifically, we use question 22 in the teacher questionnaire, PIRLS 2016 grade 4: https://nces.ed.gov/surveys/pirls/pdf/P16_TQ_final.pdf.

Control group average	0.764	0.764	0.764	0.764
Explain or support their understanding of what they have read				
Effect	0.0499	0.0275	0.0983	0.0239
SE	[0.0351]	[0.0345]	[0.0465]	[0.0538]
Observations	7,948	7,948	7,948	7,948
Control group average	0.752	0.752	0.752	0.752

Note: The table continues on the next page.

Table 6. Estimated effect on teaching practices among teachers at stage 3, cont.

Compare what they have read with experiences they have had				
Effect	0.0565	0.0473	0.0734	0.0505
SE	[0.0364]	[0.0366]	[0.0468]	[0.0570]
Observations	7,965	7,965	7,965	7,965
Control group average	0.654	0.654	0.654	0.654
Compare what they have read with other things they have read				
Effect	0.1186***	0.0964	0.1423***	0.1303
SE	[0.0405]	[0.0411]	[0.0495]	[0.0665]
Observations	7,942	7,942	7,942	7,942
Control group average	0.521	0.521	0.521	0.521
Make predictions about what will happen next in the text they have read				
Effect	0.0775	0.0549	0.1076	0.0803
SE	[0.0323]	[0.0347]	[0.0437]	[0.0550]
Observations	7,946	7,946	7,946	7,946
Control group average	0.416	0.416	0.416	0.416
Make generalizations and draw inferences based on what they have read				
Effect	0.0483	0.0327	0.0562	0.0694
SE	[0.0375]	[0.0381]	[0.0463]	[0.0629]
Observations	7,946	7,946	7,946	7,946
Control group average	0.706	0.706	0.706	0.706
Describe the style or structure of the text they have read				
Effect	0.0552	0.0539	0.07	0.0356
SE	[0.0395]	[0.0380]	[0.0506]	[0.0665]
Observations	7,938	7,938	7,938	7,938
Control group average	0.458	0.458	0.458	0.458

Note: All models include school-fixed effects interacted with grade and wave and a year fixed effect interacted with grade and wave. The questions asked follow question 22 in the teacher questionnaire: PIRLS 2016 grade 4 survey. The outcome is the probability that the teacher answered “often” or “very often” to the specific question. The scale is “never”, “rarely”, “sometimes”, “often” and “very often”. Baseline is the average response reported in the control group. Cluster-adjusted standard errors at the school level are in brackets and */**/*** refers to statistical significance at the 10/5/1 percent level.

We have concluded that the Boost for Reading improves self-assessed skills especially among teachers in higher grades who more often lack formal education in teaching literacy. To investigate if earlier skills are related to the impact on changed teaching practices, we have estimated the program effects by pooling all the outcomes with respect to teaching practices reported in Table 6 but dividing the sample depending on whether the teacher had formal education in teaching literacy or not. Panel A and B, respectively, in Appendix Table A4 presents the results. It turns out that the estimated positive effects are driven by teachers who lack formal education in teaching literacy, supporting the hypothesis that the Boost for Reading mainly improved skills in teaching literacy among teachers with lower previous knowledge, and that is to a higher degree teachers at stage 3.

5.4 Summary

Taken together, we can conclude that the Boost for Reading program was appropriately implemented, increased training in teachers' collegial collaboration and support through supervision, and that participating teachers at least in higher grades rated their self-assessed competence in teaching literacy higher after taking part in the Boost for Reading program. We also observe that teachers who lack formal education in didactics changed their teaching practices, but we do not observe the same pattern among teachers with formal education in didactics, which could be explained by their higher previous knowledge and skills in teaching literacy. Thus, the higher degree of lack of formal education in didactics among teachers at higher stages may explain the larger impact of the Boost for Reading on self-assessed competence in teaching literacy at higher stages.

6 Effects on student test scores

We now turn to the performance of students after the school has implemented the Boost for Reading program among the teachers. We begin by presenting treatment effects on test scores in the Swedish language and mathematics, that is, the weighted sum of the estimated effects over all waves and event time 0, 1, 2 and 3, pooled over all grades and for each grade separately. This is followed by a sensitivity analysis. We then present the results on test scores in social study subjects and science studies followed by heterogenous effects of the overall impact. However, we start by presenting results from two balance tests.

6.1 Balance tests

Table 7 presents results on i) the probability of taking the test and ii) predicted test scores (within the test-taking population). These two outcomes address compositional differences between treated and control schools.[28] We use our baseline specification, but the outcomes are an indicator for whether a student takes the test and predicted test scores (an index of students' family background characteristics), respectively. We present the pooled effect over all event-time periods for grades 3, 6 and 9 combined (column 1) and separately for the different

[28] The probability to take the test can also be seen as a positive outcome, given that the weakest students are least likely to take the test. However, if there are effects on test-taking, they are likely to also affect the student composition which complicates the effect evaluation.

educational stages (columns 2–4). The table convincingly shows that the Boost for Reading did not affect test-taking, nor are there any compositional changes across treated and control schools, as judged from the close-to-zero coefficients on predicted test scores.

Table 7. Effects on test-taking and predicted Swedish test scores

	(1)	(2)	(3)	(4)
	All grades pooled	Grade 3	Grade 6	Grade 9
Outcome: Non-missing test result				
Pooled effect	0.0024	0.0009	0.0038	0.003
	[0.0035]	[0.002]	[0.0034]	[0.0102]
Observations	5,145,898	1,801,782	1,813,662	1,530,454
R-squared	0.126	0.126	0.091	0.122
Wave*grade*school f.e.	Yes	Yes	Yes	Yes
Wave*grade*year f.e.	Yes	Yes	Yes	Yes
Baseline (in control group):	0.928	0.958	0.937	0.882
Outcome: Predicted Swedish test scores				
Pooled effect	0.0003	-0.0005	0.0009	0.0009
	[0.0014]	[0.0015]	[0.0029]	[0.0032]
Observations	4,774,821	1,725,536	1,699,374	1,349,911
R-squared	0.124	0.202	0.119	0.098
Wave*grade*school f.e.	Yes	Yes	Yes	Yes
Wave*grade*year f.e.	Yes	Yes	Yes	Yes

Note: Each estimate represents a weighted average of wave and event-time effects. Robust standard errors in brackets are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1

6.2 Swedish language and mathematics

Table 8 presents the baseline results of the effects of the Boost for Reading on student test scores in the Swedish language and in mathematics. On average, across all grades, the Boost for Reading has a statistically significant but very small positive effect on test results in Swedish, amounting to 1.3 percent of a standard deviation. This estimated pooled effect is driven by improved results in grades 6 and 9, where the point estimates are closer to 2 percent of a standard deviation. In mathematics, the estimates are smaller, not statistically significant but positive, and the pattern is in line with the results in Swedish, namely larger estimates in higher grades. Studies of teacher interventions in the U.K. indicate that programs which had positive effects on reading also lead to improvement in mathematics (Machin and McNally, 2008; Machin et al., 2018).[29] However, since we do not get statistically significant estimates of the Boost for Reading on mathematics, we cannot confirm those results in this study.

[29] For example, the reading demand (in terms of text difficulty) of mathematics tests can be nearly 70 percent of that of a reading assessment (Machin and McNally, 2008).

Detecting small effects of 1–2 percent of a standard deviation is demanding in terms of precision. In Figure 2, we present the estimated effects by event time for Swedish (panel a) and mathematics (panel b). In neither Swedish nor mathematics, the time-varying estimates are statistically different from zero. The lagged treatment estimates are, in Swedish all close to zero and insignificant, which indicates that treated and control schools are on parallel trends in the pre-period and lends support to the identifying assumption. In mathematics, the corresponding estimates are also all insignificant, which is reassuring, but they are not stable around zero implying that it is harder to get precise results for this outcome.

The main takeaway from the pooled estimates is that the overall effect of the Boost for Reading program on Swedish test scores is small. Apart from the fact that such a small effect is difficult to detect even in large samples with high statistical power, it is a small effect also in comparison to many other educational interventions.[30]

Table 8. Baseline effects on test scores, all grades

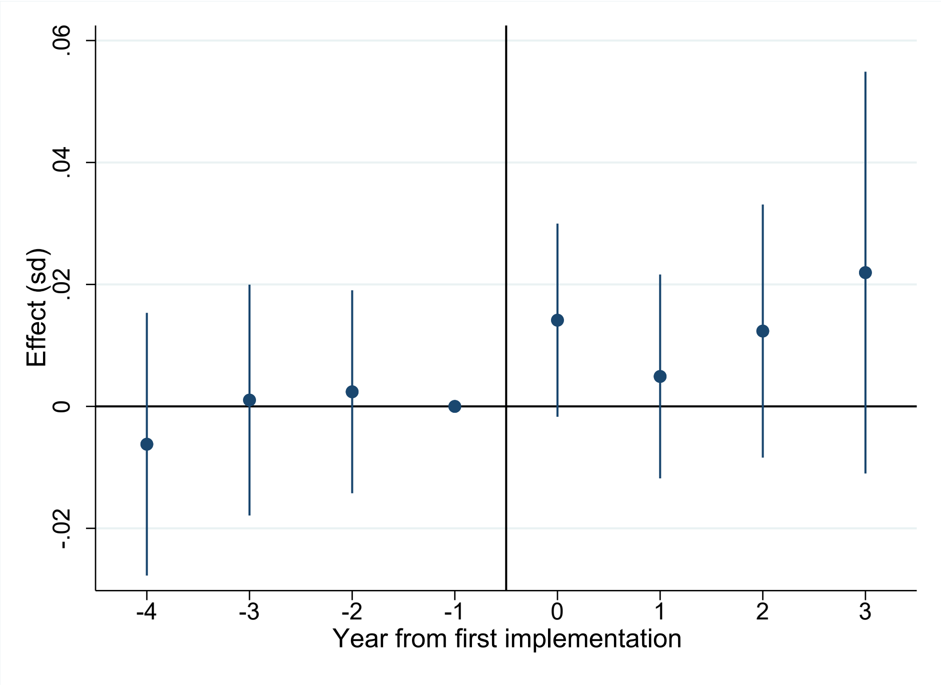
	(1)	(2)	(3)	(4)
	All grades pooled	Grade 3	Grade 6	Grade 9
Test score outcome: Swedish				
Pooled effect	0.013**	0.0053	0.017	0.0204*
	[0.0061]	[0.0091]	[0.0117]	[0.011]
Observations	4,774,825	1,725,536	1,699,374	1,349,915
R-squared	0.068	0.056	0.081	0.067
Wave*grade*school f.e.	Yes	Yes	Yes	Yes
Wave*grade*year f.e.	Yes	Yes	Yes	Yes
Test score outcome: Mathematics				
Pooled effect	0.0088	0.0034	0.0125	0.0139
	[0.0061]	[0.0095]	[0.011]	[0.0102]
Observations	4,677,236	1,735,260	1,702,778	1,239,198
R-squared	0.091	0.067	0.117	0.089
Wave*grade*school f.e.	Yes	Yes	Yes	Yes
Wave*grade*year f.e.	Yes	Yes	Yes	Yes

Note: Each estimate represents a weighted average of wave and event-time effects. Robust standard errors in brackets are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1

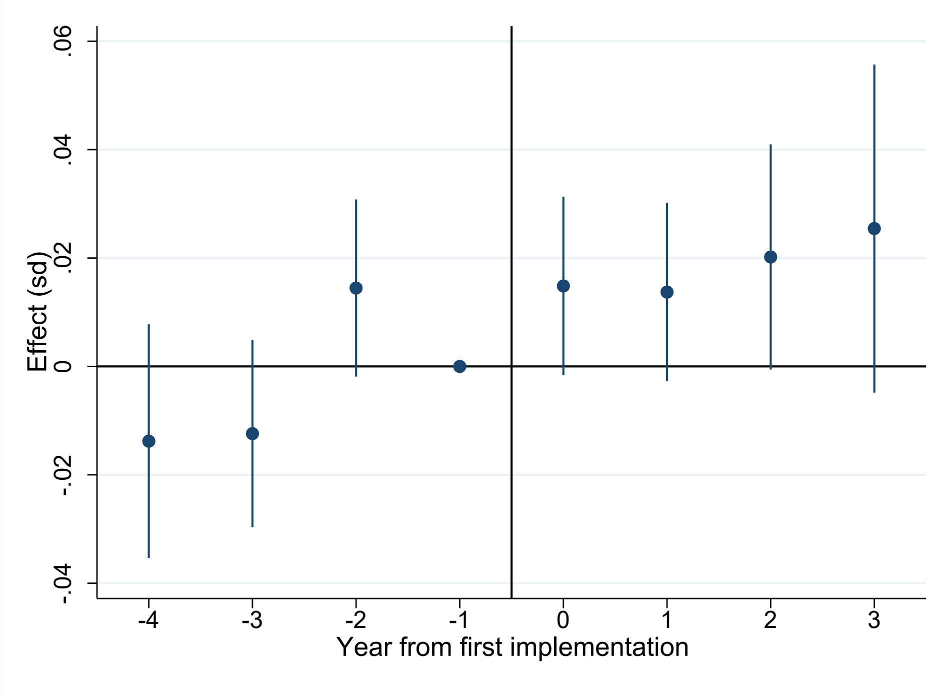
[30] Benchmarks for interpreting effect sizes of educational interventions can be found in Kraft (2020). The benchmarks have been based on estimates from 747 randomized control trials (RCTs) evaluating educational interventions on standardized test scores. According to benchmarks, effects smaller than 0.05 of a standard deviation can be considered small, effects between 0.05–0.20 represent medium effects; and effects larger than 0.2 can be considered large.

Figure 2. Event-time effects on test scores, all grades pooled

Panel a) Swedish



Panel b) Mathematics



Note: Each event-time estimate represents a weighted average of wave-specific effects. The reference year is t-1. The 95-percent confidence intervals are based on cluster-robust standard errors clustered at the school level.

6.3 Sensitivity analysis

We test the robustness of our results in Swedish using a variety of alternative specifications. Table 9, column 1, presents the baseline results. In column 2 we add student background controls, and the estimates remain very similar. In column 3, we include controls for other school interventions that were ongoing at the time (the "Boost for Mathematics" and the "Career teacher reform"), and again, the estimates are robust to the inclusion of these controls.[31]

In Appendix Table A5, we test the sensitivity of the standard errors to the level of clustering. We find that the standard errors are very similar regardless of whether we cluster at the school level (baseline), the municipality level, or at the school-stage level.[32]

Table 9. Robustness test of baseline result, including controls

	(1)	(2)	(3)
Test score outcome: Swedish			
Pooled effect	0.013**	0.0117**	0.0129**
	[0.0061]	[0.0058]	[0.0061]
Observations	4,774,825	4,774,825	4,774,825
R-squared	0.068	0.175	0.068
Student background controls		Yes	
School intervention controls			Yes
Wave*grade*school f.e.	Yes	Yes	Yes
Wave*grade*year f.e.	Yes	Yes	Yes

Note: Each estimate represents a weighted average of wave and event-time effects. Student background controls in column 2 are: gender, first- or second-generation immigrant, mother's and father's income and dummies for mother's and father's level of education. The school intervention controls in column 3 are the share of treated mathematics teachers in the "Boost for Mathematics" and dummies for whether the school took part in the Career teacher reform. Robust standard errors in brackets are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1

6.4 Social study subjects and science studies

Our data allow us to study the effect of the Boost for Reading on test scores also in social study subjects and in science studies, but only in grade 9. Table 10 presents the results: we find positive effects on student performance in both social study subjects and science studies in grade 9 of about 4–5 percent of a standard deviation.[33] These estimated effects are twice as large as the corresponding estimated effect in Swedish. Figure 3 presents the estimated effects on pooled test scores in social study subjects and science studies by event-time. In the post-period,

[31] See Grönqvist m.fl. (2021) and Grönqvist m.fl. (2022) for more details of these reforms.

[32] An additional concern is that when analyzing stage 2 and 3, the effects could be confounded by earlier treatment status in the previous stage. As an example, a student could have been treated in grade 6 in 2015/16 and will emerge as a grade 9 test-taker in 2018/19 (our last outcome year). When dropping 2018/19 from the analysis sample, we find that the baseline effect for all is 0.009 (se 0.0065). In addition, the larger and more precisely estimated estimates presented in section 6 are robust to excluding the last outcome year.

[33] In Table A6 we present the corresponding results on predicted test scores in mathematics, social study subjects, and science studies, which show that treated and control schools are balanced also with respect to expected performance in these subjects.

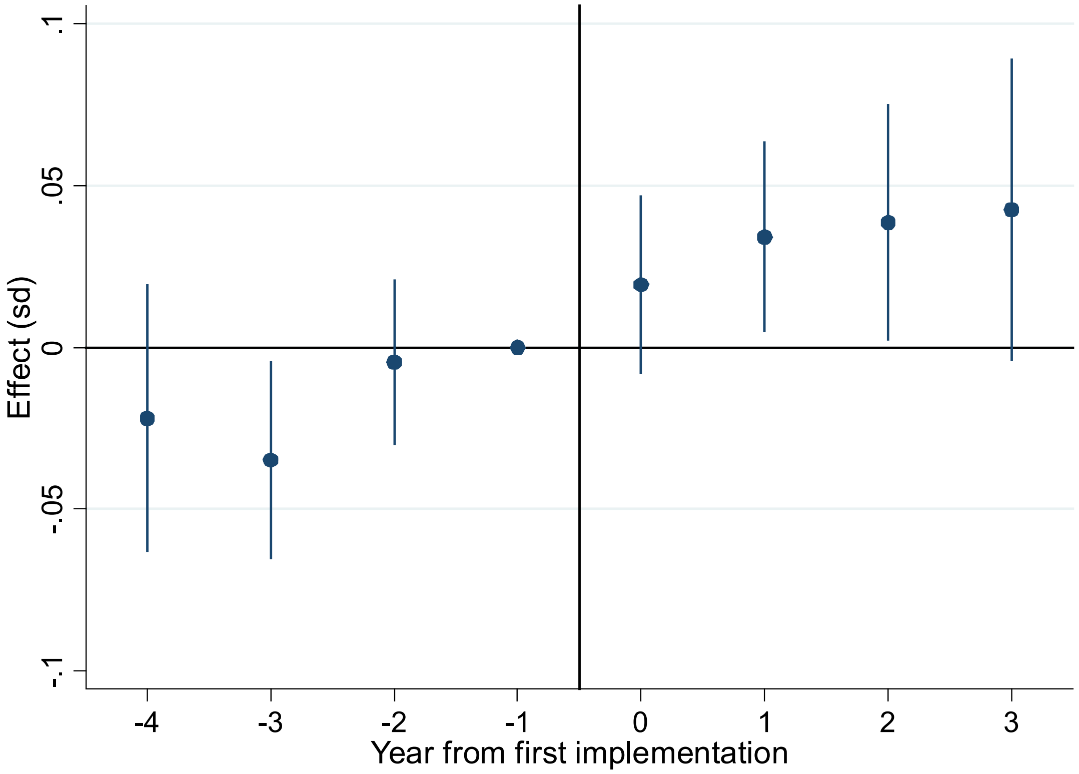
the estimated effects are in general higher and increases somewhat over time. During the pre-treatment period the estimates are lower, and they are jumping around but importantly: they do not suggest any clear difference in pre-trends.

Table 10. Effects on test scores in social study subjects and science studies

	(1) Grade 9	(2) Grade 9
Test score outcome:	*Social study subjects*	*Science studies*
Pooled effect	0.0466***	0.0463***
	[0.0124]	[0.0141]
Observations	1,391,255	1,357,184
R-squared	0.099	0.097
Wave*grade*school f.e.	Yes	Yes
Wave*grade*year f.e.	Yes	Yes

Note: Each estimate represents a weighted average of wave and event-time effects. Robust standard errors in brackets are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1

Figure 3. Event-study on test scores in grade 9 in social study subjects and science studies pooled



Note: Each event-time estimate represents a weighted average of wave-specific effects. The reference year is -1. The 95-percent confidence intervals are based on cluster-robust standard errors clustered at the school level.

The estimated effects presented in Table 10 suggest that the Boost for Reading boosted student performance in social study subjects and science studies more than performance in the Swedish language. It is also interesting to note that the estimated impact is about the same in both social study subjects and science studies despite the fact that the share treated teachers at the school level is much higher in the former subjects (at the school level around on average 60 percent vs 40 percent, see Table 2). This may seem surprising but given the program's focus on integrating literacy broadly across all subjects and the participation of teachers from different disciplines, we have no a priori expectation on the relative effect sizes across subjects. Moreover, this pattern suggests that literacy is a general skill important for performing well in many different subjects, and this skill can be acquired in one subject and used in another.

6.5 Heterogeneous effects

In this section we examine heterogeneous effects of the program in various dimensions. We explore whether the program was particularly beneficial for certain types of students and schools. For precision reasons, we focus on pooled grade 9 test scores in Swedish, social study subjects and science studies– the subject-stage combination where we find positive effects on average. (In Appendix Table A7 to Table A10, we present the corresponding results for the overall effect on Swedish test scores in all grades.)

Table 11 shows the results separately for girls, boys, immigrants, and natives. All estimates are of about the same magnitude and there are no statistically significant differences in terms of the effect size across groups.

Table 11. Effects on pooled test scores in Swedish, social study subjects and science studies in grade 9, by student type

	(1) All	(2) Girls	(3) Boys	(4) Immigrants	(5) Natives
Pooled effect	0.0379***	0.0399***	0.0371***	0.0278	0.0385***
	[0.0097]	[0.0114]	[0.0117]	[0.0248]	[0.0096]
Observations	4,098,354	2,005,821	2,092,386	272,799	3,824,460
R-squared	0.088	0.098	0.104	0.131	0.088
Wave*school*subject f.e.	Yes	Yes	Yes	Yes	Yes
Wave*year*subject f.e.	Yes	Yes	Yes	Yes	Yes

Note: Each estimate represents a weighted average of wave and event-time effects. Robust standard errors in brackets are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1

Next, in Table 12 we examine heterogeneous effects by students' family background. We split the students into three equally sized groups based on their position in the distribution of predicted test scores. The lowest third (percentile 0 – percentile 33) is the group with the lowest expected performance based on their parents' socioeconomic status and the family's migration background. The two other groups are defined in the corresponding way. The results show similarly sized effects for students of all backgrounds.

Table 12. Effects on pooled test scores in Swedish, social study subjects and science studies in grade 9, by student's predicted test scores (three groups)

	(1) All	(2) Lowest third	(3) Middle third	(4) Highest third
Pooled effect	0.0379***	0.0317***	0.0408***	0.0429***
	[0.0097]	[0.0113]	[0.0116]	[0.012]
Observations	4,098,354	1,365,090	1,365,511	1,367,276
R-squared	0.088	0.065	0.071	0.078
Wave*school*subject f.e.	Yes	Yes	Yes	Yes
Wave*year*subject f.e.	Yes	Yes	Yes	Yes

Note: Each estimate represents a weighted average of wave and event-time effects. In columns 2,3 and 4, students are divided into three equally sized groups based on their predicted test scores. The lowest third are students whose predicted test scores are in percentiles 0–33 in the predicted test score distribution; the middle third are in percentiles 34–67 and the top third are in the percentiles 68–100. Robust standard errors in brackets are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1

To further understand the effectiveness of the program, we investigate differences across schools by teacher qualifications, teacher experience, school size and school municipality. These dimensions are relevant since they can shed light on whether the Boost for Reading compensates for lack of formal teacher qualifications and experience, and whether professional development is more effective in large schools where teachers might benefit from interactions with a more heterogeneous group of colleagues.[34] Table 13 presents effects separately for schools with high vs. low shares of certified teachers and schools whose teachers on average have high vs. low experience. In both cases, schools have been divided into two groups based on the median share/experience in the pre-treatment year 2013/14. The results show that the program had larger effects in schools where teachers have lower formal qualifications and lower teacher experience. Although the estimates are not statistically significantly different from each other, the results are in line with the hypothesis that the program improved the teaching more among unqualified and inexperienced teachers than among their trained and experienced colleagues. Table 14 presents the results by school size (where groups are defined by the median school-cohort size in 2013) and by urban/rural municipalities. The Boost for Reading was effective in all school types. However, similar to the Boost for Mathematics and in accordance with the previous literature, effects are larger in big schools than in small schools.

Table 13. Heterogeneous effects on pooled test scores in Swedish, social study subjects and science studies in grade 9, by teacher characteristics at the school

	(1) All	(2) High share certified	(3) Low share certified	(4) High experience	(5) Low experience
Pooled effect	0.0379***	0.0201	0.0544***	0.024*	0.051***
	[0.0097]	[0.0137]	[0.014]	[0.0134]	[0.0141]
Observations	4,098,354	1,667,733	1,789,837	1,741,719	1,715,866
R-squared	0.088	0.081	0.093	0.082	0.083
Wave*school*subject f.e.	Yes	Yes	Yes	Yes	Yes
Wave*year*subject f.e.	Yes	Yes	Yes	Yes	Yes

Note: Each estimate represents a weighted average of wave and event-time effects. In columns 2, 3, 4 and 5, schools have been divided into “high” and “low” based on the median share certified and median of average experience in the pre-treatment year 2013. Robust standard errors in brackets are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1

Table 14. Heterogeneous effects on pooled test scores in Swedish, social study subjects and science studies in grade 9, by school type

	(1)	(2)	(3)	(4)	(5)

[34] Murphy m.fl., (2021) find that a teacher peer-to-peer observation program in the U.K. had negative effects in small schools due to disruption effects, but positive effects in large schools. The latter is consistent with a “matching” effect, that is, in larger schools there is a wider heterogeneity in teacher quality and therefore a higher probability that low performing teachers are matched with and can learn from higher performing colleagues.

	All	Large school	Small school	Urban municipality	Rural municipality
Pooled effect	0.0379***	0.0428***	0.0222	0.0465***	0.0316***
	[0.0097]	[0.0131]	[0.0145]	[0.0155]	[0.0118]
Observations	4,098,354	1,676,366	1,797,335	1,402,272	2,689,605
R-squared	0.088	0.079	0.098	0.108	0.059
Wave*school*subject f.e.	Yes	Yes	Yes	Yes	Yes
Wave*year*subject f.e.	Yes	Yes	Yes	Yes	Yes

Note: Each estimate represents a weighted average of wave and event-time effects. Schools are divided into "large" and "small" based on median school-cohort size in the pre-treatment year 2013. Urban municipalities are the largest cities (Stockholm, Gothenburg and Malmo) including the suburban municipalities surrounding them. Robust standard errors in brackets are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1

7 Concluding discussion

The "Boost for Reading" program aimed to improve literacy among Swedish compulsory school students by offering in-service training to incumbent teachers. The training was in the form of collegial lesson preparations and evaluations, supervised by an experienced coach. We investigate whether this program was able to influence teachers' professional practice in the classroom, both during and after its formal ending, and whether this in turn affected student outcomes. Effective in-service training programs targeted at incumbent teachers could be a promising policy tool to raise teacher quality across the board. However, to make a difference, the intended program content must trickle down into the daily teaching activities and change the teaching beyond the end of the program.

By using data from a teacher survey, we show that the Boost for Reading was implemented in line with its aims: teachers increased their in-service training through collegial collaboration and through the support of a coach. When analyzing the program effects on self-assessed skills in literacy, we find no evidence of any effects among teachers in the lowest grades (1–3), while we do find larger and positive effects among teachers in higher grades (4–9). We find no evidence of changed teaching practices related to "reading strategies" (a central element of literacy) among teachers in grades 1–6, while we do find that grade 7–9 teachers changed their teaching as judged by the significant effects on two out of eight outcomes.

When we turn to student outcomes, we find positive but very small average effects on test scores in Swedish and mathematics, but only the estimated effect in Swedish is statistically significant. Estimating heterogeneous effects by educational stages suggest larger effects among students in higher grades. Among students observed in grade 9, test scorers in Swedish increased by about 2 percent of a standard deviation. The corresponding estimated effect is about 4 percent in both social study subjects and science studies, even though the average share among science teachers at the school level that participated in the Boost for Reading program was much lower than the share among teachers in social study subjects. This pattern suggests that literacy is a general skill important for performing well in many different subjects, and this skill can be acquired in one subject and used in another. Comparing our results to the previous literature, our findings echo those from the literature on "coaches" in in-service training programs, which generally show positive effects on student achievements (Kraft et al. 2018).

The different magnitudes of the estimated effects depending on stage are in line with the results from the teacher survey suggesting that teachers in lower grades did not change their teaching practices after the program to the same extent as teacher in higher grades. In Holmlund et al. (2021), we show that a larger share of the teachers in lower grades had studied literacy didactics at the university level and Carlbaum et al. (2019) conclude that the content of the Boost for Reading should already be known by certified teachers. Thus, an explanation for the smaller estimated effect among students in lower grades could be that teachers at those levels were not as affected by the program since they already were well-equipped to teach literacy. In this context it is interesting to note that the small fraction of teachers at stage 3 who lacks formal education in didactics also reported improved self-assessed competence in teaching literacy after taking part of the Boost for Reading program. Thus, a policy conclusion is that in-service training can be a means to compensate for the skills provided by regular teacher education in a setting where schools are left with only non-certified candidates for a teaching position.

To understand the different impact of the Boost for Reading across stages, it is important to take into account the following: At stage 1 and 2, we can only observe student performance in Swedish and mathematics, and not in reading and writing-heavy subjects such as social study subjects and science studies. Moreover, the outcome at stage 1 is less informative than the outcomes at higher stages, as it is primarily intended to capture whether the student surpasses the pass mark, while the outcomes at stages 2 and 3 capture the student's ability and skills in a more comprehensive manner. This in turn, may limit our ability to capture positive effects at the lowest stage. Finally, we cannot exclude the possibility that the module content differs across stages which may also play a role. Taken together, we cannot exclude that the Boost for Reading also has improved the teaching quality, and thereby student performance, at stage 1, although the estimated program effect in our evaluation is lower at this stage than it is at the higher ones.

From a policy perspective it is also interesting to compare the results of the Boost for Mathematics with those of Boost for Reading. Grönqvist et al. (2021) find that the Boost for Mathematics improved the test scores in mathematics (pooled across all grades) by on average

2.6 percent of a standard deviation. Our corresponding estimated effect in Swedish is about 1 percent of a standard deviation. When comparing the estimated effects of the two programs, it is important to remember that the Boost for Mathematics was focused on one subject, while the idea of the Boost for Reading was to target a broader group of teachers. This difference has implications both for evaluating the program and for the outcomes of interest. In the Boost for Mathematics, the outcome (test scores in mathematics) can be more closely related to the treatment (mathematics teaching practices), which in turn can be measured with higher intensity (a larger share of participating teachers at the school level), increasing the possibilities to find larger effects in an evaluation. In the Boost for Reading, we expect positive effects in many different subjects if the program is successful but attenuated by the low treatment intensity (lower share of participating teachers). In this context, it is interesting that we find effects of about 4 percent of a standard deviation in both social study subjects and science studies among grade 9 students.

Finally, we conclude that the lesson study format can yield positive effects without peer-to-peer observation. The findings from our study and from the evaluation of the Boost for Mathematics show that even without peer observation, lesson study can be successful (Grönqvist et al. 2021).

References

Angrist, Joshua D., and Victor Lavy. 2001. 'Does Teacher Training Affect Pupil Learning? Evidence from Matched Comparisons in Jerusalem Public Schools'. *Journal of Labor Economics 19* (2): 343–69. https://doi.org/10.1086/319564.

Briole, Simon, and Éric Maurin. 2022. 'There's Always Room for Improvement: The Persistent Benefits of a Large-Scale Teacher Evaluation System'. *Journal of Human Resources*, March. https://doi.org/10.3368/jhr.1220-11370R1.

Burgess, Simon, Shenila Rawal, and Eric S. Taylor. 2021. 'Teacher Peer Observation and Student Test Scores: Evidence from a Field Experiment in English Secondary Schools'. *Journal of Labor Economics* 39 (4): 1155–86. https://doi.org/10.1086/712997.

Callaway, Brantly, and Pedro H. C. Sant'Anna. 2021. 'Difference-in-Differences with Multiple Time Periods'. *Journal of Econometrics*, Themed Issue: Treatment Effect 1, 225 (2): 200–230. https://doi.org/10.1016/j.jeconom.2020.12.001.

Carlbaum, Sara, Anders Hanberger, Eva Andersson, Astrid Roe, Michael Tengberg, and Katarina Kärnebro. 2019. *Utvärdering av Läslyftet : Slutrapport från den nationella utvärderingen av Läslyftets genomförande och effekter i olika skolformer*. Umeå universitet. http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-172139.

Chen, Xiangming, and Yurong Zhang. 2019. 'Typical Practices of Lesson Study in East Asia'. *European Journal of Education* 54 (2): 189–201. https://doi.org/10.1111/ejed.12334.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. 'Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood'. *American Economic Review* 104 (9): 2633–79. https://doi.org/10.1257/aer.104.9.2633.

Cordingley, P., M. Bell, B. Rundell, and D. Evans. 2003. 'The Impact of Collaborative CPD on Classroom Teaching and Learning.' In *Research Evidence in Education Library. Version 1.1*. Vol. 2003. London: EPPI-Centre, Social Science Research Unit, Institute of Education.

Cunha, Flavio, and James Heckman. 2007. 'The Technology of Skill Formation'. *American Economic Review* 97 (2): 31–47.

Grönqvist, Erik, Lena Hensvik, and Anna Thoresson. 2022. 'Teacher Career Opportunities and School Quality'. *Labour Economics*, European Association of Labour Economists, World Conference EALE/SOLE/AASLE, Berlin, Germany, 25 – 27 June 2020, 77 (August): 101997. https://doi.org/10.1016/j.labeco.2021.101997.

Grönqvist, Erik, Björn Öckert, and Olof Rosenqvist. 2021. ‘Does the “Boost for Mathematics” Boost Mathematics?’ 2021:22. IFAU - Institute for Evaluation of Labour Market and Education Policy. https://www.ifau.se/Forskning/Publikationer/Working-papers/2021/does-the-boost-for-mathematics-boost-mathematics/.

Guryan, Jonathan, Jens Ludwig, Monica P. Bhatt, Philip J. Cook, Jonathan M.V. Davis, Kenneth Dodge, George Farkas, et al. 2021. ‘Not Too Late: Improving Academic Outcomes Among Adolescents’. Working Paper. Working Paper Series. National Bureau of Economic Research. https://doi.org/10.3386/w28531.

Holmlund, Helena, Josefin Häggblom, and Erica Lindahl. 2021. ‘Läs- och skrivutvecklande undervisning i den svenska grundskolan’. 2021:5. IFAU - Institute for Evaluation of Labour Market and Education Policy. https://www.ifau.se/Forskning/Publikationer/Rapporter/20212/las--och-skrivutvecklande-undervisning-i-den-svenska-grundskolan/.

Jackson, Kirabo, and Alexey Makarin. 2018. ‘Can Online Off-the-Shelf Lessons Improve Student Outcomes? Evidence from a Field Experiment’. *American Economic Journal: Economic Policy* 10 (3): 226–54. https://doi.org/10.1257/pol.20170211.

Jacob, Brian A., and Lars Lefgren. 2004. ‘The Impact of Teacher Training on Student Achievement: Quasi-Experimental Evidence from School Reform Efforts in Chicago’. *The Journal of Human Resources* 39 (1): 50–79. https://doi.org/10.2307/3559005.

Kirsten, Nils. 2020. *Svenska lärares deltagande i kompetensutveckling: En statistisk bearbetning av uppgifter om lärares kompetensutveckling i TALIS, TIMSS, PIRLS och PISA 2001-2018*. Uppsala universitet. http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-405426.

Kraft, Matthew A. 2020. ‘Interpreting Effect Sizes of Education Interventions’. *Educational Researcher* 49 (4): 241–53. https://doi.org/10.3102/0013189X20912798.

Kraft, Matthew A., David Blazar, and Dylan Hogan. 2018. ‘The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence’. *Review of Educational Research* 88 (4): 547–88. https://doi.org/10.3102/0034654318759268.

Leigh, Andrew. 2010. ‘Estimating Teacher Effectiveness from Two-Year Changes in Students’ Test Scores’. *Economics of Education Review* 29 (3): 480–88. https://doi.org/10.1016/j.econedurev.2009.10.010.

Liu, Jing, and Susanna Loeb. 2019. ‘Engaging Teachers: Measuring the Impact of Teachers on Student Attendance in Secondary School’. *Journal of Human Resources*, July, 1216. https://doi.org/10.3368/jhr.56.2.1216-8430R3.

Machin, Stephen, and Sandra McNally. 2008. ‘The Literacy Hour’. *Journal of Public Economics* 92 (5): 1441–62. https://doi.org/10.1016/j.jpubeco.2007.11.008.

Machin, Stephen, Sandra McNally, and Martina Viarengo. 2018. ‘Changing How Literacy Is Taught: Evidence on Synthetic Phonics’. *American Economic Journal: Economic Policy* 10 (2): 217–41. https://doi.org/10.1257/pol.20160514.

Mullis, I.V.S., M.O. Martin, P. Foy, and M. Hooper. 2017. ‘PIRLS 2016 International Results in Reading. Comprehension Skills and Strategies – PIRLS 2016’. https://timssandpirls.bc.edu/pirls2016/international-results/pirls/classroom-instruction/comprehension-skills-and-strategies/index.html.

Murphy, Richard, Felix Weinhardt, and Gill Wyness. 2021. ‘Who Teaches the Teachers? A RCT of Peer-to-Peer Observation and Feedback in 181 Schools’. *Economics of Education Review* 82 (June): 102091.

Regeringsbeslut U2013/7215/S. n.d. ‘Uppdrag Om Fortbildning i Läs- Och Skrivutveckling - Läslyftet’.

Rockoff, Jonah E. 2004. ‘The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data’. *The American Economic Review* 94 (2): 247–52.

Skolforskningsinstitutet. 2019. ‘Läsförståelse Och Undervisning Om Lässtrategier’. Systematisk översikt 2019:02. Solna: Skolforskningsinstitutet.

Skolverket. 2012. ‘PIRLS 2011 - Läsförmågan Hos Svenska Elever i Årskurs 4 i Ett Internationellt Perspektiv’. Rapport 381.

‘Skolverket 2020.Pdf’. n.d.

Sun, Liyang, and Sarah Abraham. 2021. ‘Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects’. *Journal of Econometrics*, Themed Issue: Treatment Effect 1, 225 (2): 175–99. https://doi.org/10.1016/j.jeconom.2020.09.006.

Taylor, Eric S., and John H. Tyler. 2012. ‘The Effect of Evaluation on Teacher Performance’. *American Economic Review* 102 (7): 3628–51. https://doi.org/10.1257/aer.102.7.3628.

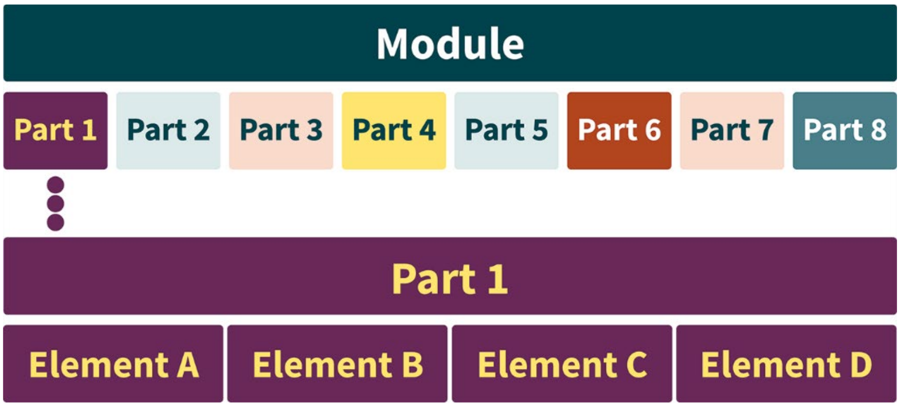
Timperley, Helen, Aaron Wilson, Heather Barrar, and Irene Fung. 2007. ‘Teacher Professional Learning and Development - Best Evicence Syntesis Iteration’. Ministry of Education, New Zealand.

Wiswall, Matthew. 2013. ‘The Dynamics of Teacher Quality’. *Journal of Public Economics* 100 (April): 61–78. https://doi.org/10.1016/j.jpubeco.2013.01.006.

Yoon, Kwang Suk, Teresa Duncan, Silvia Wen-Yu Lee, Beth Scarloss, and Kathy L. Shapley. 2007. ‘Reviewing the Evidence on How Teacher Professional Development Affects Student Achievement. Issues & Answers. REL 2007-No. 033’. *Regional Educational Laboratory Southwest (NJ1)*. Regional Educational Laboratory Southwest. https://eric.ed.gov/?id=ED498548.

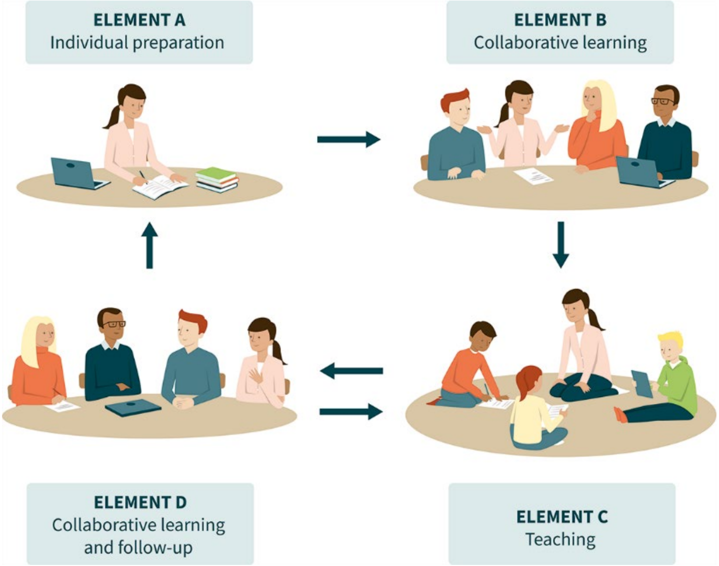
Appendix

Figure A1. Schematic description of the Boost for Reading



Source and illustration: Skolverket (National Agency for Education)

Figure A2. The planning and feed-back cycle of Boost for Reading



Source: Skolverket (National Agency for Education)

Illustration: Typoform

Table A1. Effects of the Boost for Reading on teacher characteristics among survey respondents

	Pooled	Implementation year	One year after	Two years after
Education in literacy didactic at university level (yes=1, 0 otherwise)				
Effect	0.02	0.0344	0.0259	0.0015
SE	[0.0126]	[0.0162]	[0.0202]	[0.0285]
Observations	19,773	19,773	19,773	19,773
Control group average	0.842	0.842	0.842	0.842
Years of experience teaching Swedish				
Effect	0.4586	0.7911	0.8153	−0.3834
SE	[0.394]	[0.506]	[0.609]	[0.822]
Observations	19,903	19,903	19,903	19,903
Control group average	12.41	12.41	12.41	12.41
Qualified to teach in Swedish (yes=1, 0 otherwise)				
Effect	0.0321	0.0533***	0.0313	0.0268
SE	[0.0131]	[0.0159]	[0.0200]	[0.0319]
Observations	18,791	18,791	18,791	18,791
Control group average	0.853	0.853	0.853	0.853
Years of experience teaching social study subjects				
Effect	−0.0128	-0.0113	−0.012	-0.039
SE	[0.0251]	[0.0343]	[0.0378]	[0.0560]
Observations	16,701	16,701	16,701	16,701
Control group average	1.520	1.520	1.520	1.520
Qualified to teach in social study subjects (yes=1, 0 otherwise)				
Effect	0.0193	0.0399	0.0075	0.0158
SE	[0.0152]	[0.0178]	[0.0226]	[0.0375]
Observations	18,418	18,418	18,418	18,418
Control group average	0.778	0.778	0.778	0.778

Note: All models include school-by-grade-by-wave fixed effects and time-by-grade-by-wave fixed effects. Cluster-adjusted standard errors in brackets at the school level and */**/*** refers to statistical significance at the 10/5/1 percent level.

Table A2. Effects of assigned treatment on actual treatment status

	(1)	(2)	(3)	(4)
	All grades pooled	Grade 3	Grade 6	Grade 9
Pooled effect	0.8243***	0.8877***	0.7683***	0.7866***
	[0.0054]	[0.0064]	[0.0134]	[0.0081]
Observations	4,704,115	1,701,206	1,673,220	1,329,689
R-squared	0.662	0.769	0.652	0.538
Wave*grade*school f.e.	Yes	Yes	Yes	Yes
Wave*grade*year f.e.	Yes	Yes	Yes	Yes

Note: Each estimate represents a weighted average of wave and event-time effects. Robust standard errors in brackets are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1

Table A3. Effects of the Boost for Reading on collegial collaboration

How often do you, with the aim to develop the students' literacy, together with another teacher at your school: Dependent variable is the probability that the teacher answered, "At least once every month" or "At least once every week" in comparison to "never", and "at least once every semester".

	Pooled	Implementation year	One year after	Two years after
Plan Leeson content				
Effect	0.0351	0.0579***	0.0222	0.0522
SE	[0.0177]	[0.0222]	[0.0295]	[0.0413]
Observations	19,808	19,808	19,808	19,808
Control group average	0.679	0.679	0.679	0.679
Conduct lessons and teaching activities				
Effect	−0.0058	0.014	−0.0207	−0.0388
SE	[0.0207]	[0.0256]	[0.0347]	[0.0441]
Observations	19,728	19,728	19,728	19,728
Control group average	0.481	0.481	0.481	0.481
Follow up the outcome of teaching activities				
Effect	0.0201	0.0413	0.0135	0.0074
SE	[0.0207]	[0.0255]	[0.0334]	[0.0473]
Observations	19,671	19,671	19,671	19,671
Control group average	0.625	0.625	0.625	0.625
Follow up the development of the students' reading and writing skills				
Effect	−0.0048	0.0011	−0.0202	0.0013
SE	[0.0213]	[0.0261]	[0.0350]	[0.0440]
Observations	19,733	19,733	19,733	19,733
Control group average	0.551	0.551	0.551	0.551
Discuss pedagogic and didactic issues				
Observations	19,733	19,733	19,733	19,733
Effect	−0.0048	0.0011	−0.0202	0.0013
SE	[0.0213	[0.0261]	[0.0350]	[0.0440]
Control group average	0.551	0.551	0.551	0.551
Visit colleagues' lessons to exchange teaching experiences (peer observations)				
Effect	0.0227	0.0351	0.0232	0.024
SE	[0.0147]	[0.0201]	[0.0246]	[0.0293]
Observations	19,754	19,754	19,754	19,754
Control group average	0.817	0.817	0.817	0.817

Note: The table continues on the next page.

Table A3. Effects of the Boost for Reading on collegial collaboration, cont.

	Pooled	Implementation year	One year after	Two years after
Exchange learning and teaching materials				
Effect	-0.0056	−0.0105	0.0084	−0.0274
SE	[0.0152]	[0.0200]	[0.0258]	[0.0302]
Observations	19,813	19,813	19,813	19,813
Control group average	0.167	0.167	0.167	0.167
Collaborate across teaching subjects				
Effect	−0.0155	−0.0268	−0.0021	−0.0341
SE	[0.0189]	[0.0250]	[0.0300]	[0.0384]
Observations	19,801	19,801	19,801	19,801
Control group average	0.693	0.693	0.693	0.693
Participate in competence development together with colleagues				
Effect	−0.0124	−0.0125	−0.005	−0.0445
SE	[0.0201]	[0.0252]	[0.0307]	[0.0439]
Observations	19,774	19,774	19,774	19,774
Control group average	0.464	0.464	0.464	0.464

Note: All models include school-by-grade-by-wave fixed effects and time-by-grade-by-wave fixed effects. Cluster-adjusted standard errors in brackets at the school level and */**/*** refers to statistical significance at the 10/5/1 percent level.

Table A4. Estimated average effect on teaching practices depending on formal education in literacy didactics

How often do you ask the students to do the following in your teaching about and with text? Dependent variable is the probability to report often" or "very often" in comparison to "never", "rarely", and "sometimes".				
	Pooled	Implementation year	One year after	Two years after
Panel A: teachers with formal education at university level in teaching literacy				
Average over all grades 1-9				
Effect	0.024	0.0318	0.0281	0.0315
SE	[0.0105]	[0.0133]	[0.0161]	[0.0245]
Observations	16,580	16,580	16,580	16,580
Control group average	0.637	0.637	0.637	0.637
Grade 1-3				
Effect	0.0114	0.036	−0.0165	−0.0024
SE	[0.0206]	[0.0207]	[0.0255]	[0.0441]
Observations	6,139	6,139	6,139	6,139
Control group average	0.574	0.574	0.574	0.574
Grade 4-6				
Effect	0.024	0.0076	0.0387	0.0439
SE	[0.0223]	[0.0219]	[0.0272]	[0.0487]
Observations	4,799	4,799	4,799	4,799
Control group average	0.657	0.657	0.657	0.657
Grade 7-9				
Effect	0.0569	0.049	0.0678	0.0576
SE	[0.0231]	[0.0242]	[0.0280]	[0.0367]
Observations	5,642	5,642	5,642	5,642
Control group average	0.689	0.689	0.689	0.689

Note: The table continues on the next page.

Table A4. Estimated average effect on teaching practices depending on formal education in literacy, cont.

Panel B: teachers without formal education at university level in teaching literacy

	Pooled	Implementation year	One year after	Two years after
Average over all grades 1-9				
Effect	0.1024***	0.1022***	0.2452***	0.0518
SE	[0.0284]	[0.0366]	[0.0474]	[0.0649]
Observations	2,633	2,633	2,633	2,633
Control group average	0.497	0.497	0.497	0.497
Grade 1-3				
Effect	0.1241	0.1169***	0.0547	0.2236
SE	[0.0577]	[0.0370]	[0.0840]	[0.147]
Observations	105	105	105	105
Control group average	0.483	0.483	0.483	0.483
Grade 4-6				
Effect	0.0286	0.057	0.1364	−0.1525
SE	[0.109]	[0.101]	[0.132]	[0.195]
Observations	407	407	407	407
Control group average	0.517	0.517	0.517	0.517
Grade 7-9				
Effect	0.1099***	0.0544	0.1943***	0.0977
SE	[0.0381]	[0.0413]	[0.0442]	[0.0703]
Observations	2,121	2,121	2,121	2,121
Control group average	0.495	0.495	0.495	0.495

Note: All models include school-fixed effects interacted with grade and wave and a year fixed effect interacted with grade and wave. Baseline is the average response reported in the control group. Cluster-adjusted standard errors at the school level are in brackets and */**/*** refers to statistical significance at the 10/5/1 percent level.

Table A5. Effects on test scores in Swedish, alternative levels of cluster-adjusted standard errors

	(1)	(2)	(3)
	Baseline Cluster at school-level	Cluster at municipality	Cluster at school-stage level
Pooled effect	0.013**	0.013**	0.013**
	[0.0061]	[0.0063]	[0.006]
Observations	4,774,825	4,772,758	4,774,825
R-squared	0.068	0.068	0.068
Wave*grade*school f.e.	Yes	Yes	Yes
Wave*grade*year f.e.	Yes	Yes	Yes

Note: Each estimate represents a weighted average of wave and event-time effects. Robust standard errors in brackets are clustered at the school/municipality/school-stage level. *** p<0.01, ** p<0.05, * p<0.1

Table A6. Specification test: Effects on predicted test scores in mathematics, social study subjects and science studies

	(1)	(2)	(3)	(4)	(5)	(6)
	All grades pooled	Grade 3	Grade 6	Grade 9	Grade 9	Grade 9
Predicted test score outcome:				*Mathematics*	*Social study subjects*	*Science studies*
Pooled effect	0.001	−0.0006	0.0029	0.0016	0.0011	0.0013
	[0.0015]	[0.0015]	[0.0029]	[0.0035]	[0.0034]	[0.0034]
Observations	4,677,232	1,735,260	1,702,778	1,239,194	1,391,254	1,357,180
R-squared	0.243	0.301	0.262	0.190	0.164	0.179
Wave*grade*school f.e.	Yes	Yes	Yes	Yes	Yes	Yes
Wave*grade*year f.e.	Yes	Yes	Yes	Yes	Yes	Yes

Note: Each estimate represents a weighted average of wave and event-time effects from a separate regression. Robust standard errors in brackets are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1

Table A7. Effects on Swedish test scores in grades 3, 6 and 9, by student type

	(1)	(2)	(3)	(4)	(5)
	All	Girls	Boys	Immigrants	Natives
Pooled effect	0.013**	0.0138**	0.0131*	0.0067	0.0123**
	[0.0061]	[0.0067]	[0.0073]	[0.0182]	[0.0061]
Observations	4,774,825	2,342,789	2,431,587	295,568	4,475,847
R-squared	0.068	0.086	0.086	0.134	0.069
Wave*grade*school f.e.	Yes	Yes	Yes	Yes	Yes
Wave*grade*year f.e.	Yes	Yes	Yes	Yes	Yes

Note: Each estimate represents a weighted average of wave and event-time effects from a separate regression. Robust standard errors in brackets are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1

Table A8. Effects on Swedish test scores in grades 3, 6 and 9, by students' predicted test scores (three groups)

	(1)	(2)	(3)	(4)
	All	Lowest third	Middle third	Highest third
Pooled effect	0.013**	0.0083	0.0151**	0.014*
	[0.0061]	[0.0088]	[0.0073]	[0.0072]
Observations	4,774,825	1,590,611	1,591,113	1,591,916
R-squared	0.068	0.067	0.060	0.080
Wave*grade*school f.e.	Yes	Yes	Yes	Yes
Wave*grade*year f.e.	Yes	Yes	Yes	Yes

Note: Each estimate represents a weighted average of wave and event-time effects from a separate regression. In columns 2,3 and 4, students are divided into three equally sized groups based on their predicted test scores. The lowest third are students whose predicted test scores are in percentiles 0–33 in the predicted test score distribution;

the middle third are in percentiles 34–67 and the top third are in the percentiles 68–100. Robust standard errors in brackets are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1

Table A9. Heterogeneous effects on test scores in Swedish in grades 3, 6 and 9, by teacher characteristics at the school

	(1) High share certified	(2) Low share certified	(3) High experience	(4) Low experience
Pooled effect	0.0074	0.0189**	0.0081	0.0191**
	[0.0081]	[0.0093]	[0.0082]	[0.0091]
Observations	2,023,060	2,065,001	2,052,824	2,035,237
R-squared	0.061	0.071	0.064	0.068
Wave*grade*school f.e.	Yes	Yes	Yes	Yes
Wave*grade*year f.e.	Yes	Yes	Yes	Yes

Note: Each estimate represents a weighted average of wave and event-time effects from a separate regression. Schools have been divided into "high" and "low" based on the median share certified and median of average experience in the pre-treatment year 2013. Robust standard errors in brackets are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1

Table A10. Heterogeneous effects on test scores in Swedish in grades 3, 6 and 9, by school type

	(1) Large school	(2) Small school	(3) Rural district	(4) Urban district
Pooled effect	0.012	0.0088	0.0078	0.0163
	[0.0086]	[0.009]	[0.0074]	[0.0105]
Observations	1,959,978	2,266,332	3,065,451	1,707,304
R-squared	0.066	0.066	0.053	0.081
Wave*grade*school f.e.	Yes	Yes	Yes	Yes
Wave*grade*year f.e.	Yes	Yes	Yes	Yes

Note: Each estimate represents a weighted average of wave and event-time effects from a separate regression. Schools are divided into "large" and "small" based on median school-cohort size in the pre-treatment year 2013. Urban municipalities are the largest cities (Stockholm, Gothenburg, and Malmo) including the suburban municipalities surrounding them. Robust standard errors in brackets are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1