Does the 'Boost for Mathematics' Boost Mathematics?[a]

A Large-Scale Evaluation of the 'Lesson Study' Methodology on Student Performance

Erik Grönqvist, Björn Öckert, Olof Rosenqvist[b]

December 22, 2021

Abstract

Students in East-Asian countries dominate international assessments. One possible explanation for this success is their use of 'Lesson study' to enhance teaching practices, but evidence on its effectiveness is still scant. We evaluate a national teacher development program in Sweden – the 'Boost for Mathematics' – containing core elements of Lesson study. Exploiting the gradual roll-out of the program across compulsory schools, we find that it improves teaching practices and boosts students' mathematics performance. The positive effect on student performance persists also long after the intervention has ended. We also show that the program passes a cost-benefit test.

Keywords: Teacher development, Student performance, Mathematics, Lesson study
JEL-codes: I21, J45, I28

[a] We have benefitted from comments from Adrian Adermon, Helena Holmlund, Kirabo Jackson, Eric Taylor, Anna Sjögren, Jonas Vlachos, Felix Weinhardt and from seminar participants at Université Clermont Auvergne, Frisch Center, Gothenburg University, IFAU, University of Essex, Stockholm-Uppsala Education Economics Workshop, Arne Ryde Workshop on Economics of Education, and University of Surrey. We also thank Swedish National Agency for Education for collaborating in the collection of teacher survey data, and Josefin Häggblom, Sara Martinson and Kristina Sibbmark for help with the teacher questionnaire.
[b] Grönqvist: Department of Economics, Uppsala University (UU); Öckert: Institute for Evaluation of Labour Market and Education Policy (IFAU) and Department of Economics UU; Rosenqvist: IFAU.

1 Introduction

What can be learned from other school systems? Hong Kong, Singapore, Shanghai, Taiwan and Japan consistently dominate league tables of international student performance, like TIMMS and PISA; in particular in mathematics and science (OECD 2019; Mullis et al. 2020). This is a concern for many Western economies, since the quality of schools, and the skills developed, is linked to labor market opportunities and earnings (Hanushek et al. 2015; Murnane, Willett, and Levy 1995; Murnane et al. 2000; Neal and Johnson 1996) and economic growth (Hanushek and Woessmann 2008; 2016).

One possible explanation for the success of East-Asian educational systems is the use of 'Lesson study' to support teachers' professional development (Lewis and Tsuchida 1999; Stigler and Hiebert 1999).[3] While details differ across countries, the common core of Lesson study entails a general school-based practice where teachers collaborate in learning cycles; plan and evaluate lessons together, and give each other feedback and critique; sometimes facilitated by outside experts (Chen and Zhang 2019; Rappleye and Komatsu 2017).[4] The Lesson study approach to teachers' professional development has been imported to schools in many countries around the world to improve students' learning outcomes (Lewis and Lee 2017; Quaresma et al. 2018).[5]

[3] Other possible explanations are selection of teachers to the profession, school curriculum, work ethic and discipline, and out-of-school tuition (see for example Jerrim 2015).

[4] Lesson study is generally considered to originate from Japan, but similar improvement strategies has also been developed in China, Singapore, Hong Kong and more recently in South Korea (Chen and Zhang 2019; Cheng and Yee 2012; Huang, Fang, and Chen 2017; Pang 2016).

[5] Lesson study communities are found in most European countries, the US and Canada, and in 2006 the World Association of Lesson Studies was formed (Lewis and Lee 2017; Quaresma et al. 2018). Rappleye and Komatsu (2017) report that about 1,500 US schools have active Lesson study communities, and since 2010 the Florida Department of Education has adopted Lesson study as a state-wide vehicle for teacher development (Akiba and Wilkinson 2016).

But does Lesson study work? The empirical evidence is still scant and comes mainly from small-scale trials. A systematic review on the effectiveness of Lesson study by Cheung and Wong (2014) is inconclusive due to remaining methodological challenges. More recently, Murphy, Weinhardt, and Wyness (2021) find no significant improvement on test scores (mathematics, science, reading, spelling and grammar) from a randomized intervention (89 treated schools and 92 controls) of a two-year Lesson study program for 4–6 graders in the UK.

In this paper we study the effectiveness of the 'Boost for Mathematics': A one-year national in-service professional development program for mathematics teachers in Sweden, introduced in 2013 as a response to the falling mathematics performance of Swedish students in TIMSS 2007 and PISA 2009 (Utbildningsdepartementet 2012). The program contains central elements of Lesson study. Teachers work collaboratively in learning cycles where they discuss a particular mathematical content in group, plan a lesson together, try out the planned lesson in their own classes, and then share their experiences in group. The teacher learning groups are supported by an external mathematics tutor, and the learning cycle is organized along educational modules with study material covering core mathematical areas (e.g. algebra, geometry and probability), where schools choose modules depending on their local needs (Skolverket 2016a). The modules promote a more active instructional practices, where teachers challenge students, discuss problem-solving strategies in class, and use assessments to learn about teaching outcomes. During the program the teacher groups meet about once a week for an entire school year. A distinguishing feature of the Boost for Mathematics is that teachers self-assess their performance, rather than being evaluated in the classroom by colleagues as is the case in traditional Lesson study.

In-service professional development programs can be an important policy tool as there is a large variation in the contribution of teachers to students' learning outcomes (e.g. Chetty, Friedman, and Rockoff 2014a; 2014b; C. K. Jackson 2018; Jacob and Lefgren 2008; Kane, Rockoff, and Staiger 2008; Rivkin,

Hanushek, and Kain 2005; Rockoff 2004; Rothstein 2010), and since teacher effectiveness improves with experience, also beyond the early parts of the career (Harris and Sass 2011; Papay and Kraft 2015; Wiswall 2013). This suggests that teacher skills are malleable through learning-by-doing. Teachers may advance their professional practice over time by improving how to give instruction, interact with students, manage the classroom, and organize the curriculum. Ost (2014) finds that both general teaching skills and content specific skills improve with experience. Hence, there is scope for in-service training to enhance this learning-by-doing process.

The Boost for Mathematics was organized by the Swedish National Agency for Education – by providing educational modules, training of tutors, and central funding – with an ambition that all mathematics teachers would participate in the training. The program was rolled out gradually across schools 2013/14–2015/16 and 60 percent of compulsory school mathematics teachers had participated by the end of the academic year 2015/16. The training spots (i.e., funding) were available to school districts in proportion to the number of mathematics teachers, and school districts had discretion over which and when schools participated (Skolverket 2012). We exploit the staggered implementation across schools in a difference-in-differences strategy suggested by Sun and Abraham (2020) and Callaway and Sant'Anna (2020) to evaluate the impact of the Boost for Mathematics, by comparing the change in student performance in each wave of schools participating in the program to the change for schools that never participated. We find no evidence that the intervention was targeted toward schools with declining (or increasing) student test scores, or that participation is related to changes in schools' student composition, thus lending support for a causal interpretation of results.

We show that the Boost for Mathematics improves student performance on standardized tests in mathematics, in particular in primary school.[6] On average, test scores increase by 2.6 percent of a standard deviation in treated schools. Student learning is boosted also in the longer run; at least 4–5 years after the intervention at the school. In fact, we find positive effects for students who had not yet started school when the program was implemented. However, the intervention does not benefit students from disadvantaged backgrounds; we find no effect for students in the lowest quartile of predicted test scores, but positive effects for those in other quartiles. We also show that the program passes a cost-benefit test.

We use a uniquely collected teacher survey panel data to explore the underlying mechanisms of the Boost for Mathematics, to find that teachers in participating schools receive more in-service training in the year of implementation. Participating teachers are also more satisfied with their work and believe they – and their colleagues – have improved their teaching practices. In participating schools, the Boost for Mathematics increases peer-to-peer interaction among teachers, but the effect peters out soon after the program has ended. Hence, we find no evidence of a long-lasting improvement of the collaborative culture in participating schools, as was intended. On the other hand, we find persistent changes in classroom practices. Consistent with the content of the educational modules, teachers in the program devote more time to discuss problem-solving strategies with students in class, and less time to let students solve standard problems.

This paper makes a number of contributions to the literature. A first contribution is that we – to the best of our knowledge – provide the first large-scale evaluation of Lesson study methodology, finding it to be an effective strategy to enhance student learning. We also show that it passes a cost-benefit test.

[6] Lindvall et al. (2021) find no significant difference in student performance for 208 teachers participating in the Boost for Mathematics compared to 145 untreated teachers, using TIMSS data for Sweden in 2015. In Appendix D, we re-analyze the Swedish TIMSS 2015 data.

A second contribution is that we study the impact of teacher in-service training in the longer run. Student performance in treated schools is found to be higher 4–5 years after the intervention has ended.[7] This finding is corroborated by persistent effects of the program on teacher's instructional practices.

A third contribution is that we provide real world evidence from a national implementation of a teacher training program. Effects found in small trials may not always generalize as it can be difficult to change the general teaching culture by national policy, especially in a decentralized school system, see e.g. Kraft, Blazar, and Hogan (2018). While central government policies can be effective in bringing innovations to schools, they need to be adaptive to local needs, and to change teaching practices, to be effective, and the Boost for Mathematics aims to strike this balance.

A fourth contribution is to use survey data to explicitly study how teachers respond to the program in terms of peer collaboration and classroom instructional practices, thus unveiling underlying mechanisms that are usually unobserved. A fifth contribution is that we address the endogeneity in the delivery of in-service teacher training by exploiting the staggered implementation of the program across schools.

Our paper relates closely to a growing literature on peer-to-peer learning which shows that teachers learn from their colleagues. As teachers typically do not interact with colleagues in the classroom, structured peer interaction for improved planning and preparation can be an important tool for professional development. Burgess, Rawal, and Taylor (2021) find a positive impact in a field experiment of 82 UK high schools on students' mathematics and English exams from a structured teacher feedback program where teachers observe each other in the classroom and provide advice and share strategies for improvement.

[7] Earlier evaluations of in-service teacher training on student achievement have either been restricted to the year(s) of implementation (Murphy, Weinhardt, and Wyness 2021; Burgess, Rawal, and Taylor 2021; Jacob and Lefgren 2004; Garet et al. 2010; 2011) or to the first year after the intervention (Papay et al. 2020a; Randel et al. 2016).

Similarly, Papay et al. (2020) provide experimental evidence on improved student achievements in mathematics and reading from an intervention pairing low-skilled teachers to a higher skilled teacher in the same school, and instructing the pair to work together on improving teaching skills.[8] Unlike these interventions, the Boost for Mathematics does not involve peer observation, with feedback and criticism of classroom practices. Instead, teachers self-assess their classroom performance as an input to teacher group discussions.

The paper furthermore relates to teacher observation programs with feedback from external classroom observers. In a meta-study, Kraft, Blazar, and Hogan (2018) find limited evidence that coaching programs with feedback on teachers' instructional practice improve students' mathematics achievement. Using quasi-experimental variation in the timing of exposure to inspections, both Taylor and Tyler (2012) and Briole and Maurin (2019) find that assessments from classroom observations by external experts, where assessments can have consequences for career advancement, have a positive impact on students' mathematics outcomes also in a longer run; thus suggesting that teacher evaluation can be a tool for improving teacher skill and effort.[9] A main difference to our context, however, is the role of the external experts. In the Boost for Mathematics, they coach teachers to improve their instruction practices through intrinsic motivation rather than through high stakes evaluations.

Our paper also relates to Jackson and Makarin (2018) who provides experimental evidence that high-quality online instructional material, available as a didactic support for mathematics teachers, improves students' achievement.

---

[8] More informally, Jackson and Bruegmann (2009) find positive peer spillovers using variation generated by job-to-job transitions of high-quality teachers. The achievement of a teacher's students improves when having more effective colleagues, and the improvement persists over time.

[9] Sojourner, Mykerezi, and West (2014) find that tying teacher bonuses in Minnesota (Q-Comp) to multiple performance measures, including high stakes classroom observations, improved student achievement, and Dee and Wyckoff (2015) find that students benefit from a program where teachers are rated by their performance, including detailed classroom observations.

More broadly the paper relates to the effectiveness of in-service professional development programs for teachers in general. Such programs are a prevalent feature in schools and vary in form and substance (Bill & Melinda Gates Foundation 2015), but evidence is still sparse (Yoon et al. 2007; Kennedy 2016). The literature on teacher in-service training finds modest or no effects on students outcomes, in particular when implemented at scale (Angrist and Lavy 2001; Jacob and Lefgren 2004; Garet et al. 2010; 2011; Harris and Sass 2011; Randel et al. 2011; 2016). For specific didactic interventions there is evidence for both positive (Machin and McNally 2005; 2008; Jerrim and Vignoles 2016; Cilliers et al. 2019) and limited (Machin, McNally, and Viarengo 2018; Dix, Hollingsworth, and Carslake 2018) or even negative (Haeck, Lefebvre, and Merrigan 2014) impacts on student achievement.

For in-service training to improve school quality and student achievement, programs must convey instructional innovation to schools, cater for local needs, and change teachers' professional practice. Our results suggest that the Boost for Mathematics manages to do just that.

The paper unfolds as follows. The next section describes the Boost for Mathematics and the context in which it was implemented, data, and descriptive statistics. Section 3 describes the empirical strategy. The main results on student performance in mathematics, validity checks and heterogeneous effects are provided in section 4 followed by results on teachers' peer activities and classroom practices from survey data in section 5. Section 6 provides cost-benefit calculations, and the paper is concluded in section 7.

2 Institutional setting and data

From the mid 1990's, and through to TIMSS 2007 and PISA 2009, the results of Swedish students, in particular in mathematics and science, were falling in international assessments both in absolute terms and relative to other countries (OECD 2014; Mullis et al. 2012). This led to a general concern about the

development in Swedish schools, resulting in the government introducing the Boost for Mathematics in 2013.[10]

2.1 The Swedish school system

The Swedish primary and secondary school system comprise three main components: pre-school, compulsory school and upper-secondary school. At age 6, all children start a one-year preparatory pre-school class, which is followed by 9 years of compulsory schooling (grades 1–9). Students can then apply for a 3-year theoretical or vocational upper-secondary school program, which is required for post-secondary education.

The compulsory school can be divided into three stages: lower and middle stages; grades 1–3 and 4–6 (primary school), and higher stage; grades 7–9 (lower secondary school). In primary school, students typically have a class teacher who teaches most subjects; a feature that is most salient in the lowest stage. In the middle stage there is more variation across schools and the same teacher may not necessarily cover all core subjects (mathematics, Swedish and English). In lower secondary school, students have specialized subject teachers in each subject. There are national standardized tests at the end of each stage (grade 3, 6 and 9). While the tests in primary school (grade 3 and 6) are mainly used to monitor progression, the national tests in lower secondary school are high-stakes and influence the school leaving GPA at the end of grade 9, which determines the set of opportunities for upper secondary school.

The school system is publicly financed and free from tuition. Municipalities are responsible for providing compulsory education, but there are also private voucher schools, following the same curriculum. Students are free to apply to any school – public or private – in the municipality. The allocation of students

[10] It also led to other policy initiatives during the same period: A merit-based 'Career teacher promotion program' in 2013 (Grönqvist, Hensvik, and Thoresson 2021), the 'Boost for reading' in 2015 and the 'Teachers' salary boost' in 2017 In Table 3, column 3, we show that our main results are stable to any cross-contamination in the take-up of these other policies across schools.

to compulsory schools – the school form we study – in not based on academic merit. If a public school is oversubscribed, students are allocated based on proximity as a main principle, and for voucher schools students are admitted mainly based on proximity and waiting lists (Skollag 2010). About 85 percent of compulsory school students attend a public school (in one of 290 municipalities) and 15 percent attend one of the more than 800 voucher schools (Skolverket 2020).

2.2 The Boost for Mathematics

The Boost for Mathematics is a one-year in-service professional development program in mathematics didactics for teachers in mathematics in Swedish compulsory and upper-secondary schools. It was developed and organized by the Swedish National Agency for Education. The program is based on peer-to-peer learning among teachers with support from an external mathematics tutor, with the goal to provide teachers with methods and tools to develop their teaching and instill a collaborative learning culture in the school, in order to improve student's proficiency in math. The program promotes more active instructional practices, where teachers engage students with challenging tasks, organize classroom discussions, and modify their instruction in response to students' questions and thoughts (Lindvall et al. 2021). The in-service training takes place locally at the schools and is based on peer-to-peer discussions about teaching situations and mathematical contents. Teachers exchange good teaching practices, highlight their difficulties, critically examine their own instruction, and receive feedback from colleagues.

2.2.1 Learning cycles

The program centers on teacher learning groups which are supported by an external mathematics tutor, who is an experienced and skilled mathematics teacher with special mentoring training.[11] Teachers work in learning cycles where they

[11] The appointment as mathematics tutor corresponds to 20 percent of full-time and entails responsibility for several teacher groups. Tutors receive 8–9 days of training at a teacher training

discuss a specific mathematical content, plan a lesson together, carry out the lesson in class, and then share their experiences in group. The learning cycles are organized along educational modules with tailored study material, such as scientific texts and videos. A module covers a specific mathematical content (such as algebra, geometry and problem solving) from different perspectives to provide teachers with tools for reflecting, planning, and carrying out teaching in different ways. All modules consist of 8 parts, highlighting different aspects, where teachers, in each part, work through a learning cycle of 4 steps as follows (see Appendix A for additional information on the content of the Boost for Mathematics):

A. *Individual preparation*: Teachers prepare individually by studying the didactic support material for that specific part (45–60 minutes).
B. *Collaborative learning*: Teachers meet in group to discuss the material that they have studied (step A) and plan a lesson together (90–120 minutes).
C. *Classroom activity*: Each teacher tries out the planned mathematics lesson in their own classroom.
D. *Collegial follow-up*: Teachers meet in group to discuss their lessons to reflect and learn what went well and what can be improved (45–60 minutes).

The collegial group discussions in steps B and D of the cycle are led by the tutor. In total, it takes a teacher 24–32 hours of learning activities, plus the regular classroom teaching activities, to work through a module.

The intention of the Boost for Mathematics is for teacher groups to work intensively with two modules during a school year (about 60 hours), which

college with emphasis on mentoring and group processes, and the content of the support material. Principals at participating schools also receive 4–5 days of training on how to strengthen their pedagogical leadership, on the content of the program, and how to organize the training.

means that teachers have collegial learning group meetings every week during the year.[12]

2.2.2 Assignment of treatment

Due to the Swedish decentralized school system, the central government cannot make in-service training programs mandatory, but it can provide recommendations and financial support, which they did for the Boost for Mathematics. The program was introduced in the academic year 2013/14 and rolled out to schools over three years through government grants providing financial resources to participating school districts. The grant covered the cost of mathematics tutors (20 percent of full time) and provided co-financing for all participating teacher (about 18 hours).

In each wave, the funds were restricted to one third of the mathematics teachers in the school district. The districts could apply for their reserved funding and were responsible for appointing tutors and allocating the available slots to the schools (Skolverket 2012). In total, 89 percent of the public school districts, and 28 percent of the private districts, decided to participate in the program. The main reasons for not taking part, as stated in interviews, were problems in adapting to the organizational model, problems for smaller school districts to participate due to scale properties of the program, and that school districts already were working with teacher professional development in other ways (Skolverket 2016a).

In participating school districts, on average 80 percent of the schools are treated (82 percent for public and 71 percent for private districts). The majority of schools (94 percent) participated with all stages in the same year, and we therefore define treatment at the school level. The principals were responsible for organizing the training, e.g., forming the teacher groups and making sure that sufficient time was available for the training. By 2016, about two thirds of

[12] The training takes place during regular working time, and schools have to repay the government grant if participating teachers must work overtime.

all compulsory schools, and 60 percent of all mathematics teachers had participated in the program.

In sum, the set of participating schools is determined by decisions at two levels; school districts choosing to participate, and then choosing which schools in the district to implement the program (and in which wave). In Section 2.4 we describe participating and non-participating schools.

2.3 Data

To analyze the effects of the Boost for Mathematics we combine data from different administrative sources held by Statistics Sweden and the National Agency for Education. In addition, we have collected survey data from mathematics teachers for a sample of compulsory schools. The underlying population for the analysis is the panel of Swedish primary and lower secondary schools (grades 1–9) for the years 2011–2019, and the students and mathematics teachers in these schools.

2.3.1 Administrative data

The school panel is based on information from the Swedish school registry listing all schools with a unique school identifier. We also retrieve information on school size, school district, and organizational form (i.e., municipal or voucher school) from the school registry.

To classify when (or if) a specific school participates in the Boost for Mathematics we first use information from the Swedish teacher registry on teachers' subject of teaching to identify the population of teachers in mathematics in all schools. The teacher register covers all educational personnel in Swedish schools measured is collected as a part of the official school statistics. The teacher register is also used to retrieve information on teachers' experience and certification. For each observed mathematics teacher, we then determine participation in the program (and when they participated) by linking them to a register on government grant payments for participating teachers, provided by the National Agency for Education. Using this data, we calculate the share of

mathematics teacher at each school that participate in the program. A school is defined as participating in the Boost for Mathematics a specific year, if at least 50 percent of the mathematics teachers received the government grant and is regarded as not participating if no grant is received that year. If some, but less than half, of the teachers receive grants, we regard participation as undetermined, and the school is dropped from the data.[13]

To all schools, we link individual level information on student performance at the end of each stage of compulsory school, using registry data on test scores from national tests in mathematics (and Swedish).[14] These exams are taken during the spring semester in grades 3, 6, and 9. We standardize student test scores (mean 0 and standard deviation 1) by year in the full population of test-takers.[15]

Using personal identifiers, we furthermore link each student to his or her parents using the population registry, and then to parents' socioeconomic and demographic characteristics using information from administrative records. This data includes information on parents' country of birth, level of education and income. To avoid that any of these variables are endogenously determined by the program (e.g., by parental responses) they are measured the year a child enters a specific stage (i.e., in grades 1, 4 or 7). We use predicted test scores as

[13] In most cases, the majority of mathematics teachers in the school participates in the training, but because of possible misclassification of teacher specialization or turnover, the share of participating teachers may include measurement errors. Appendix Figure B1 shows the distribution of participating teachers in the schools in the three waves. We can determine treatment status for 81 percent of the schools, and thus exclude 19 percent of schools. Results are, however, insensitive to changes of the treatment status threshold, see Appendix Table B1.

[14] The exams are typically marked by the student's own teacher, using centrally provided guidelines. In section 4.3 we provide evidence suggesting that the Boost for Mathematics is unlikely to affect teachers' grading standards.

[15] Each centralized exam consists of several sub-tests which are graded separately. We standardize each sub-test by year in the population of test-takers, and take the average of all sub-tests, which we, again, standardize. If a student is absent on one sub-test, we take the average of the sub-tests where the student participates.

a composite measure of students' demographic and socio-economic background.[16]

2.3.2 Teacher survey data

To gain information on professional development and teacher practices, we have collected yearly survey data (in collaboration with the Swedish National Agency for Education) from mathematics teachers for a sample of compulsory schools.

In 2013 we randomly sampled 560 schools, stratified to have an equal representation of all school stages, and we follow these schools through the years 2013–2016. In April each year we sent out a mail questionnaire to all mathematics teachers in the selected schools according to the teacher registry. The response rate of the survey varies between 42–55 percent across the waves, but there are no systematic differences across participating and non-participating schools in observable characteristics of responding teachers (see Table C1 for details). The survey data does not include personal identifiers of the teachers, so we can only link the yearly survey information at the school level.

From the teacher survey we retrieve information on different types of professional development practices as proof of treatment to check if the program affects teachers' in-service training. We additionally obtain information on teacher peer collaboration and classroom activities to assess how the program has changed teacher practices. The survey also provides assessments of own and colleagues' teaching skills and job-satisfaction.

[16] Specifically, using data for students who took the test before the reform, we regress students' test scores by grade on pre-determined student and parental characteristics and school fixed effects (R2 = 0.156), and use the estimated parameters to generate the predicted test score, similar to (Chetty, Friedman, and Rockoff 2014a). The variables used in the predictions are gender, birth month, income of mother, income of father, education of mother, education of father, indicators for whether the student and the parents are born in Sweden and indicators for having missing values on these variables.

2.4 Descriptive statistics

In our analysis data, we observe about 1,3 million unique students in 3,800 schools. The two first years of the Boost for Mathematics (academic years 2013/14 and 2014/15) about 900 schools per year participate, while only 624 schools participate the third year (2015/16). More than 1,300 schools did not participate in the program at all.

Table 1. Average characteristics of participating and non-participating schools

Column:	(1)	(2)	(3)	(4)
	Wave 1	Wave 2	Wave 3	
Sample:	(2013/14)	(2014/15)	(2015/16)	Never
Private school	0.0833	0.0417	0.0508	0.3530
	(0.2763)	(0.1920)	(0.2196)	(0.4779)
Located in major city	0.3382	0.3339	0.3810	0.4537
	(0.4731)	(0.4716)	(0.4856)	(0.4979)
School size	332	333	344	267
	(185)	(195)	(198)	(214)
Share certified teachers	0.7074	0.7296	0.7292	0.6629
	(0.2427)	(0.2305)	(0.2367)	(0.3002)
Teacher experience (years)	14.31	14.91	14.78	13.42
	(5.48)	(5.57)	(5.59)	(6.66)
Share of participating teachers	0.8181	0.8454	0.8105	0
	(0.1545)	(0.1422)	(0.1409)	(0)
Pre-reform test score	0.0119	0.0014	0.0046	-0.0085
	(0.3318)	(0.3250)	(0.3698)	(0.3861)
Predicted test scores	0.0084	-0.0087	-0.0003	0.0001
	(0.3265)	(0.3209)	(0.3418)	(0.3418)
Number of schools	959	886	624	1,331
Number of students	691,298	660,665	481,410	451,022

Note: The table shows student-weighted averages and standard deviations for schools participating in the Boost for Mathematics in different waves, and for schools that never participate. The teacher characteristics refer to mathematics teachers. All background variables are measured in the 2012/13 academic year.

Table 1 shows that participating schools are relatively similar, across waves, in pre-reform mathematics test scores measured in the academic year 2012/13. Schools that never participate are, however, slightly weaker on average with about 1–2 percent of a standard deviation lower mathematics scores. There are

also small differences in the background of students, measured as predicted scores, across schools.

Voucher schools are much less likely to participate in the Boost for Mathematics than are other schools. Only 24 percent of voucher schools implemented the program, compared to 73 percent among public schools. There are also differences in average teacher experience and certification rates between schools. Mathematics teachers in participating schools are more likely to be certified and have, on average, longer teaching experience (measured in 2012). Schools participating in the Boost for Mathematics are also larger and less likely to be located in a major city. In the next section we discuss our identification strategy to address these level differences between schools.

3 Empirical strategy and identification

The empirical challenge when evaluating the effectiveness of any in-service training program is to find a good estimate of the counterfactual outcome. We exploit the staggered implementation of the Boost for Mathematics across schools, and the fact that some schools never participated, to identify the effects of the intervention in a difference-in-differences design.

An emerging literature stresses the potential identification problems in difference-in-differences models with staggered rollout of treatment, since earlier treated cohorts are then used as controls for later treated cohorts (Goodman-Bacon 2018).[17] If there are heterogeneous treatment effects across cohorts, earlier treated cohorts are not accurate counterfactuals for later cohorts, and event study estimates will be biased (Sun and Abraham 2020). Therefore, we only use never-treated schools as controls and, thus, compare the change in outcomes for

[17] In addition, standard difference-in-differences models place more weight on cohorts in the middle of the panel, which can make it difficult to interpret the pooled treatment effects (de Chaisemartin and D'Haultfœuille 2020). This is, however, often a minor concern, in particular in our setting with only three treated cohorts (Baker 2019).

schools implementing the boost for Mathematics to the corresponding change for schools that never participated.

More specifically, for each implementation cohort g={2013, 2014, 2015} we retain only schools implementing the Boost for Mathematics that year and schools that never participate in the program. We then stack data for each cohort by event time and estimate separate effects by cohorts as suggested by Sun and Abraham (2020) and Callaway and Sant'Anna (2020). We estimate the following dynamic event study model for individual *i* (student or teacher) in school *s* in calendar year *t* and implementation cohort *g*:

$$y_{istg} = \sum_{g} \sum_{\tau \neq -1} \theta_{\tau g} D_s 1[\tau, g] + \gamma_{sg} + \lambda_{vtg} + \varepsilon_{istg} \qquad (1)$$

where y_{istg} is the outcome of interest, e.g., student test scores. Event time, τ, refers to time in relation to when the school implemented the Boost for Mathematics, and $\tau = 0$ represents the year of teacher training, $\tau = 1$ the first year after implementation, and so on. The effect of the program in event time τ = {-6,…, 5}, with τ = -1 as reference period, is estimated as a weighted average of the cohort-specific treatment effects, $\hat{\theta}_\tau = \sum_g p_{\tau g} \hat{\theta}_{\tau g}$, where the weights, $p_{\tau g}$, are the share of treated individuals in cohort g in event time τ, and $\hat{\theta}_{\tau g}$ are the corresponding treatment effects estimates. We obtain an estimate of the overall effect of the Boost for Mathematics by aggregating the effects for all years following (and including) program implementation, i.e., $\hat{\theta} = \sum_{\tau \geq 0} \sum_g p_{\tau g} \hat{\theta}_{\tau g}$. Cluster-adjusted standard errors at the school level are in parentheses to account for arbitrary correlation in outcomes between individuals within schools and over time.

We control for school-by-cohort fixed effects, γ_{sg}, to account for constant differences across schools (c.f. Table 1). In addition, the model includes cohort-specific calendar time effects, λ_{vtg}, to absorb any general time factors. We let the time effects differ between voucher and municipal schools *v* = {voucher, municipal}, since test scores have been shown to evolve differently in public

and private schools and voucher schools are less likely to participate in the program.[18] No other time varying controls are included in the main specification since they can endogenously be affected by the program.

A possible concern with evaluating the reform in the Swedish context is that parents may endogenously (de)select schools participating in the Boost for Mathematics. For this reason, students are sampled in the beginning of the stage (grades 1/4/6) and assigned the treatment status of the school they are expected to attend in the end of the stage (grades 3/6/9); that is, students are given the treatment status of the school they should have attended, had they followed the normal route.[19] This means that the earliest cohorts of students had already selected schools before the program was implemented, while parents to students in later cohorts potentially could have observed schools' treatment status at the time of school choice. In section 4.2, however, we show that student composition does not change differentially in treatment and control schools in response to the rollout of the program.

As some students change schools, not all students will take the centralized exam at the expected school. The estimates therefore capture a reduced form effect of students' expected exposure to the program. However, actual treatment corresponds to expected treatment for 84 percent of all students, which suggests

[18] Voucher schools tend to score higher on centralized exams in mathematics due to either a more selective student population, a more efficient teaching technology (Holmlund, Sjögren, and Öckert 2020), or more lenient grading practices (Tyrefors Hinnerich and Vlachos 2016; Hinnerich and Vlachos 2017). A possible diverging trend in mathematics scores for voucher schools can thus be due to voucher schools either becoming more selective in their student recruitment, innovative in teaching or in inflating the mathematics scores.

[19] Since school choice is more pronounced in the higher stage, than in the lower and middle stages of comprehensive school (Holmlund, Sjögren, and Öckert 2020), we assign students to a higher stage school based on their school in grade 6. Hence, we assign students to their school in grade 1/4/6 for the lower/middle/higher stage.

that the reduced form estimate is a good approximation of the treatment effect of interest.[20]

The identifying assumption for giving the difference-in-differences estimates a causal interpretation is that schools participating in the Boost for Mathematics would have had the same trend in outcomes, as schools that never participate, had the program not been implemented. Thus, while participating and non-participating schools may differ in average characteristics, the program must not be targeted towards schools with declining (or increasing) student test scores, or with deteriorating (or improving) student composition. Although this assumption cannot be tested formally, we show in section 4.1 that student test scores evolve similarly in participating and non-participating schools before the intervention.

4 Impact on student performance

We begin this section by presenting the main effects of the Boost for Mathematics on students' mathematics test scores. This is followed by discussions of possible threats to identification and the reliability of test scores. We then present heterogeneous effects with respect to student, teacher, and school characteristics.

4.1 Main results

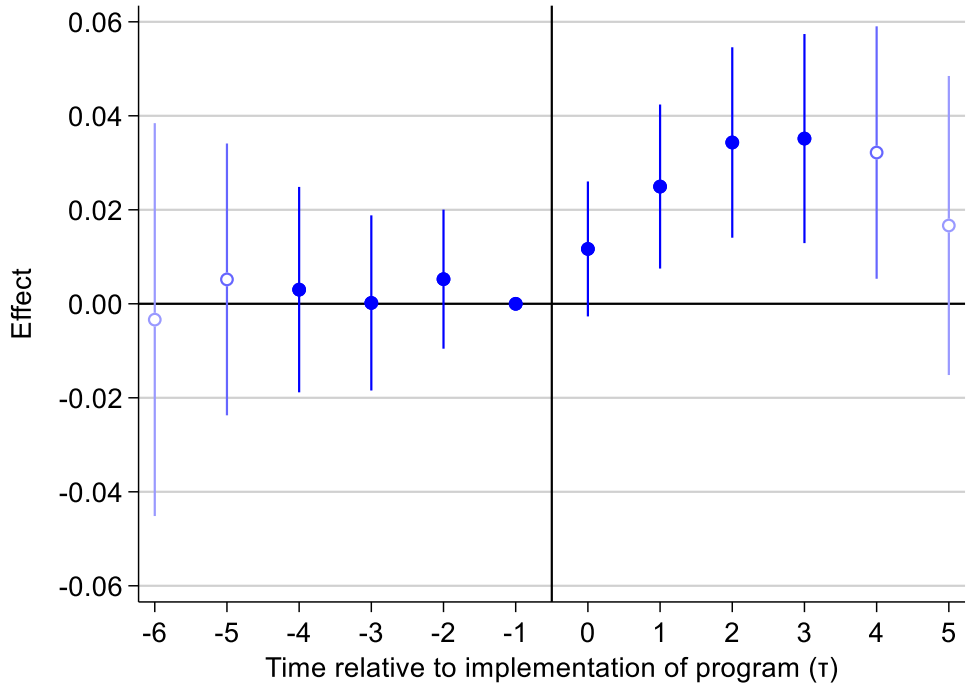
The effect of the Boost for Mathematics on student test scores in mathematics is illustrated in Figure 1. Before the program is implemented, student performance evolves similarly in treated and control schools. Thus, there is no 'effect' of schools' future treatment status and the placebo estimates (τ = -6, -5, -4, -3, -2) are all close to zero and not statistically significant.[21] This indicates that the

[20] Appendix Table B2 presents the 'first stage' estimate, i.e., the effect of the treatment status of the school that students enter in a given stage (grades 1/4/6) on the treatment status of the school they attend at the end of the stage (grades 3/6/9).

[21] This is confirmed by an F-test (p-value=0.918) of the joint hypothesis that all are zero.

program has not been targeted towards schools with falling (or improving) test scores, which lends support to the identifying assumption of our model that the program implementation was unrelated to underlying trends in test scores.

Figure 1. Effects of the Boost for Mathematics on test scores in mathematics



Note: The figure displays reduced form effects of the Boost for Mathematics on standardized test scores in mathematics along with 95-percent confidence bands. Estimates in a slightly lighter shade (τ = -5 and 4) are based only on schools in the first or second wave of the intervention and estimates in the lightest shade (τ = -6 and 5) only on schools in the first wave. The model includes school-by-cohort fixed effects and time-by-cohort fixed effects that vary by municipal and voucher schools. Test scores are measured in the end of lower/middle/higher stage (grade 3/6/9). Students are sampled in the beginning of the lower/middle/higher stage (grades 1/4/6) and assigned the treatment status of the school they are expected to attend in the end of the stage. Standard errors are clustered at the school level.

Once the Boost for Mathematics is implemented, student performance in participating schools increases (the estimates underlying Figure 1 are presented in the

first column of Appendix Table B3).[22] Already when teachers undergo training (τ = 0), student test scores rise by approximately 0.012 SD, but this estimate does not reach statistical significance (p-value = 0.105).[23] Student performance grow even further when teachers have completed the program; in the two years following implementation (τ = 1 and 2) test scores improve with 0.025–0.035 SD in participating schools. The boost in student performance persists also 3–4 years after the program is introduced (τ = 3 and 4), when new students have entered the stage. In the last follow-up period (τ = 5), which we can only observe for the schools in the first wave, the point estimate is smaller and no longer statistically significant. Due to the smaller sample size, however, the confidence band is too wide to rule out either a large positive, or even negative, effect. To gain precision, we therefore pool information from adjacent years to evaluate the longer run effects of the intervention.

The first column of Table 2 presents the impact of the Boost for Mathematics for different pairwise post-reform years. It shows that student performance is boosted in every period after program implementation. The effects are largest after 2–3 years, but the estimates for various post-reform years are not significantly different. Importantly, test scores are higher in participating schools also 4–5 years after the in-service training has ended. Thus, the Boost for Mathematics has long-lasting effects on student performance in mathematics.

On average, test scores in participating schools rise by about 0.026 SD. Due to student mobility, however, not all students attend their expected school in the end of the stage, and the reduced form (intention-to-treat) estimates therefore understate the effect of the program. Since the probability that students receive

[22] The national exams are mandatory, but students may be exempted due to illnesses, cognitive disorders, or weak language skills (immigrants). However, Appendix Table B6 shows that the Boost for Mathematics does not affect students' test-taking propensity.

[23] The remaining columns of Appendix

Table B3 show that inference is robust to clustering standard errors at the school district level or at the school×stage level (instead of the school level).

the expected treatment is 84 percent (see Appendix Table B2), the inferred IV-estimate of the program effect is 0.031 (0.026/0.84) SD.[24] Thus, there is a moderate but economically significant positive impact of the Boost for Mathematics on student learning in participating schools.

Table 2. Effects of the Boost for Mathematics on test scores in mathematics, by stage

Column:	(1)	(2)	(3)	(4)
Grades:	3, 6 and 9	3	6	9
	Panel A. Separately for different years			
0–1 years after implementation	0.0184**	0.0255*	0.0256**	0.0055
	(0.0074)	(0.0152)	(0.0120)	(0.0107)
2–3 years after implementation	0.0347***	0.0592***	0.0270*	0.0095
	(0.0102)	(0.0192)	(0.0163)	(0.0161)
4–5 years after implementation	0.0264**	0.0525**	0.0371*	-0.0081
	(0.0130)	(0.0244)	(0.0212)	(0.0195)
	Panel B. All years pooled			
All years	0.0263***	0.0447***	0.0286**	0.0044
	(0.0085)	(0.0165)	(0.0140)	(0.0123)
School×Wave FE	Yes	Yes	Yes	Yes
Year×Private×Wave FE	Yes	Yes	Yes	Yes
Observations	2,874,158	1,053,814	967,565	852,779

Note: The table shows reduced form effects of the Boost for Mathematics on standardized test scores in mathematics, divided by stage. All models include school-by-cohort fixed effects and time-by-cohort fixed effects that vary by municipal and voucher schools. The sample studied is indicated in the column heading. Test scores are measured in the end of lower/middle/higher stage (grade 3/6/9). Students are sampled in the beginning of the lower/middle/higher stage (grades 1/4/6) and assigned the treatment status of the school they are expected to attend in the end of the stage. Cluster-adjusted standard errors at the school level are in parentheses and */**/*** refers to statistical significance at the 10/5/1 percent level.

The remaining columns of Table 2 presents the effects of the Boost for Mathematics separately by stage.[25] It shows that the program only stimulates student

[24] In addition to the assumption that treated schools would have followed the same trend in outcomes as other schools in absence of the reform, the IV interpretation rests on the assumption that expected treatment status influences students' mathematics scores only through its effect on the probability to be exposed to the program by the end of the stage.

[25] Appendix Table B4 presents the yearly effects of the Boost for Mathematics separately by stage.

learning in primary school (grade 3 and 6), and that there is no significant impact in lower-secondary school. On average, the in-service training improves student performance in primary school by 0.035 SD (not shown in table) which is significantly higher than in lower-secondary school (p-value for test of difference is 0.083). This suggests that subject-specific in-service training may be more efficient for teachers with general teacher education, as is often the case for primary school class teachers.

The positive effect of the Boost for Mathematics in primary school persists in the longer run, and student test scores are higher in treated schools also 4–5 years after implementation. In particular, students entering the lower stage in the end of the follow-up period had not yet started school at the time of implementation, which suggests that the program changed teachers' instructional practices more permanently. Thus, the Boost for Mathematics successfully boosts mathematics performance, both for students who attended the school during the implementation, and for later incoming cohorts.

4.2 Exogeneity of treatment

A causal interpretation of the estimates crucially depends on the assumption that the rollout of the Boost for Mathematics is exogenous, i.e., that treated schools would have followed the same trend in outcomes, as the control schools, in absence of the program. As noted in Figure 1, student performance in participating and non-participating schools progress in a comparable way prior to the program. This is consistent with the assumption that schools in the Boost for Mathematics would have exhibited a similar pattern in outcomes as other schools, in the case the program had not been implemented.

To provide further support for the identifying assumption, we study changes in student composition between treated and control schools. We use predicted test scores – where students' pre-determined characteristics are summarized and weighted by their importance for mathematics performance – as outcome to describe how any changes in student composition is expected to translate into

outcome differences between participating and non-participating schools in the follow-up period in absence of the reform.

Table 3. Specification tests

Column:	(1)	(2)	(3)
Outcome:	Predicted test scores	Test scores	Test scores
	Panel A. Separately for different years		
0–1 years after implementation	0.0007	0.0176**	0.0175**
	(0.0014)	(0.0072)	(0.0074)
2–3 years after implementation	0.0017	0.0321***	0.0323***
	(0.0021)	(0.0097)	(0.0103)
4–5 years after implementation	0.0040	0.0212*	0.0246*
	(0.0031)	(0.0126)	(0.0133)
	Panel B. All years pooled		
All years	0.0018	0.0239***	0.0246***
	(0.0018)	(0.0081)	(0.0086)
School×Wave FE	Yes	Yes	Yes
Year×Private×Wave FE	Yes	Yes	Yes
Student controls	No	Yes	No
School intervention controls	No	No	Yes
Observations	2,874,158	2,874,158	2,874,158

Note: The table shows reduced form effects of the Boost for Mathematics on predicted test scores and test scores in mathematics. All models include school-by-cohort fixed effects and time-by-cohort fixed effects that vary by municipal and voucher schools. The outcome variable is indicated in the column heading. The student controls are gender, birth month, income of mother, income of father, education of mother, education of father, immigrant status and indicators for having missing values. The school intervention controls are dummy variables for the schools' participation in the Boost for Reading, Career teachers, Teachers' salary boost and the reintroduction of the Boost for Mathematics in 2017. Outcomes are measured in the end of lower/middle/higher stage (grade 3/6/9). Students are sampled in the beginning of the lower/middle/higher stage (grades 1/4/6) and assigned the treatment status of the school they are expected to attend in the end of the stage. Cluster-adjusted standard errors at the school level are in parentheses and */**/*** refers to statistical significance at the 10/5/1 percent level.

The first column of Table 3 presents the change in predicted test scores in schools introducing the Boost for Mathematics compared to schools that never

participate.[26] Any 'effect' of participation status on predicted test scores could be an indication of an endogenous roll-out of the program, or – for students that enter schools in the end of the follow-up period – parents' school choice responses. It is therefore reassuring that the point estimates are all close to zero and not statistically significant. This is, again, consistent with the assumption that student performance would have evolved similarly in the treated and control schools in absence of the reform. Not surprisingly, adding controls for pre-determined student characteristics has very limited impact on the estimated test score effects, see column 2.

As discussed in section 2.1, the Boost for Mathematics was the first in a series of national school initiatives to improve student performance. Although the implementation of other school development programs was not contingent on participation in the Boost for Mathematics, it opens up the concern that the estimated treatment effects may partly reflect the impact of other interventions. As a final specification check, we therefore add controls for three national school development programs implemented during the period studied.[27] Column 3 in Table 3 shows that the estimated program effects are only marginally affected when adding these school-level controls, suggesting that the estimated effects of the Boost for Mathematics do not to pick up the impact of other concurrent reforms.

4.3 Reliability of test scores

In Swedish schools, mathematics teachers grade their own students' national exams. A relevant question is therefore whether the estimated effects of the Boost for Mathematics on test scores reflect changes in teachers' grading standards rather than improved student performance. In this section, we provide three

[26] Appendix Table B5 presents the corresponding specification tests separately for every post-reform year.

[27] The initiatives are the Boost for reading, Career teachers, Teachers' salary boost and the re-introduction of the Boost for Mathematics in 2017.

pieces of evidence suggesting that the estimated effects are likely to reflect improved student learning rather than changes in teachers' grading leniency (for more details see Appendix D).

First, there is little room for teachers' subjective judgement of students' answers to questions that can be characterized as being either 'right' or 'wrong', as is often the case in mathematics. This is confirmed by the re-assessments of national exams conducted by the Swedish Schools Inspectorate for a sample of schools every year (see e.g. Skolinspektionen 2021), which shows that the teachers' judgement of their own students' mathematics performance does not differ much from that of external examiners.

Second, using data from the TIMSS 2015 survey for Sweden, we compare the effect of teachers (intraclass correlations) for student performance in mathematics on the national tests (internally graded) and on the TIMSS test (externally graded). The teacher effects are of the same order of magnitude for both tests (0.267 for the national tests and 0.249 for TIMSS), which suggests that there is little room for teachers' subjective grading in mathematics.

Third, as a final check for any impact of the Boost for Mathematics on teachers' grading standards, we exploit information on program participation obtained from the Swedish version of the TIMSS 2015 school questionnaire. The estimated difference in student performance between schools participating in the Boost for Mathematics, and schools not participating, is very similar if we use the internally graded national exams or the externally graded TIMSS test (0.047 for the national tests and 0.045 for TIMSS). This indicates that the program had very minor, if any, effects on teachers' grading standards in mathematics.

4.4 Heterogeneous effects

Having established that the Boost for Mathematics improves performance for the average student, has a longer run impact at schools, and assessed threats against identification, we next turn to heterogeneous effects of the program by

student, teacher, and school characteristics. To gain precision, we restrict attention to the overall effect of the program in the years following implementation.

First, we analyze heterogeneities effects of the program by student background. In Table 4 we estimate the effects separately for students in each quartile of the predicted test score distribution. The effect of exposure to the Boost for Mathematics is concentrated to the three highest quartiles. For students in the lower tail of the distribution, however, we find no effect (p-value of difference between the lowest and the other quartiles is 0.122).[28] This may partly be explained by immigrant students being overrepresented in the lowest quartile of predicted test scores, and that they generally gain less from the program (see Appendix Table B7).[29] However, even if we restrict the analysis to natives, the program fails to help students in the lowest quartile (see Appendix Table B8). This suggests that the Boost for Mathematics is less effective for weaker students in general, and not only for those with lower language proficiency. The more active teaching practices promoted by the program may thus not be well suited for low-performing students. Hence, the intervention has contributed to a widening of the differences in mathematics performance across students of different backgrounds, and potentially reinforced inequalities in the educational system.

Second, we investigate if the effectiveness of the program is related to the teachers' formal qualifications. Since we cannot directly link teachers to their students, we instead divide schools by the median share of certified and experienced mathematics teachers, respectively. Overall, we find quite small differences in the effects (see Appendix Table B9). Students in schools with more certified teacher do not seem to gain more (or less) from the program. While we find that effects are slightly larger in schools with a higher share of experienced teachers, but this difference is not significant. Thus, we fail to find important

[28] The program does not affect the probability to take the test for students in different quantiles of the predicted test score distribution (see Appendix Table B6).

[29] We find no significant heterogeneity by the gender of students (see Appendix Table B7).

heterogeneities of the program for different types of teachers, possibly because we cannot link teachers to their students at the individual level.

Table 4. Effects of the Boost for Mathematics on test scores, by quartiles of students' predicted test scores

Column:	(1)	(2)	(3)	(4)	(5)
Sample:	All	P0–P25	P25–P50	P50–P75	P75–P100
All years pooled	0.0263***	0.0084	0.0290***	0.0366***	0.0207**
	(0.0085)	(0.0150)	(0.0105)	(0.0105)	(0.0101)
School×Wave FE	Yes	Yes	Yes	Yes	Yes
Year×Private×Wave FE	Yes	Yes	Yes	Yes	Yes
Observations	2,874,158	699,144	720,617	725,428	728,969

Note: The table shows reduced form effects of the Boost for Mathematics on standardized test scores in mathematics, divided by quartiles of students' predicted test scores. All models include school-by-cohort fixed effects and time-by-cohort fixed effects that vary by municipal and voucher schools. The sample studied is indicated in the column heading. Test scores are measured in the end of lower/middle/higher stage (grade 3/6/9). Students are sampled in the beginning of the lower/middle/higher stage (grades 1/4/6) and assigned the treatment status of the school they are expected to attend in the end of the stage. Cluster-adjusted standard errors at the school level are in parentheses and */**/*** refers to statistical significance at the 10/5/1 percent level.

Third, we assess heterogeneities by characteristics of the schools and the environment in which they operate (see Appendix Table B10). Effects are greater in large schools (p-value for difference is 0.172) and greater in schools located in larger metropolitan areas (Stockholm, Gothenburg, Malmö) compared to schools in medium sized cities and rural areas (p-value for difference is 0.034). These results echo the findings in Murphy, Weinhardt, and Wyness (2021) who report that Lesson study was more effective in larger schools. However, when we study the effects of the Boost for Mathematics by school size and region simultaneously, we find that the heterogeneous effects are driven mainly by geographical area rather than school size.[30] The greater effects in metropolitan

[30] The average difference in effects of the Boost for Mathematics between large and small schools in the same areas (metropolitan or other) is 0.0166 (standard error 0.0186), while the

areas could potentially be explained by these larger school markets being more competitive.

Finally, we study spillover effects from the Boost for Mathematics on test scores in Swedish (see Appendix Table B11). We find that the program improves student performance also in Swedish, in particular in primary school, although the magnitude of the effect is smaller. This suggesting that the program may affect the instructional practices of class teachers in all subjects. Important spillover effects from reading to mathematics are also found by Machin and McNally (2008) in evaluating the 'Literacy Hour'.

5 Impact on teaching practices

In order for in-service training programs to have an impact on student performance it must change instructional practices in the classroom. In this section, we use teacher survey information to analyze how the Boost for Mathematics has affected teachers' peer collaboration and classroom teaching practices, and, thus, explore some of the underlying mechanisms behind the effects of the program. To study the dynamics of the reform, we present effect by year after program implementation. Teachers are here assigned to the school where they work (i.e., the school for which they answered the survey) so results should be interpreted as average treatment effects on the treated. The panel with teacher survey data spans four years (2013–2016) and we, therefore, only estimate effects for $\tau = 0, 1, 2$ and let all pre-reform years define the baseline.

To validate that teachers in treated schools in fact receive additional in-service training as proof of treatment, we first estimate the effects of the program on teachers' training activities. Table 5 shows that mathematics teachers receive significantly more in-service training when the program is implemented ($\tau = 0$), than teachers in other schools. The training covers core mathematics,

average difference in effects between schools of the same size (small or large) in metropolitan and other areas is 0.0356 (standard error 0.0187).

mathematics didactics, and assessment of students' mathematics skills. During the implementation phase, teachers also more often report to participate in peer collaboration and coaching activities, which are two core elements of the Boost for Mathematics. The effects on these training activities add up to 65 hours, which can be compared to the expected time-use of two training modules being about 60 hours. However, the categories in the survey question are not mutually exclusive, so the net additional hours of training during the implementation year is likely lower.

Table 5. Effects of the Boost for Mathematics on teachers' training activities (hours per school year)

Column:	(1)	(2)	(3)	(4)	(5)
Outcome:	Mathe-matics	Didactics	Coaching	Collabo-ration	Assess-ment
Implementation year	14.82***	23.33***	10.95***	12.89***	3.13***
	(0.95)	(1.09)	(0.77)	(1.05)	(1.04)
1 year after implementation	3.46***	4.92***	1.65**	2.61*	-2.02
	(1.16)	(1.43)	(0.65)	(1.44)	(1.30)
2 years after implementation	1.58	1.08	0.04	1.86	-2.42
	(1.53)	(1.87)	(0.82)	(1.94)	(1.77)
School×Wave FE	Yes	Yes	Yes	Yes	Yes
Year×Private×Wave FE	Yes	Yes	Yes	Yes	Yes
F-test for $\theta_0=\theta_1=\theta_2=0$	0.0000	0.0000	0.0000	0.0000	0.0000
Observations	8,376	8,376	8,376	8,376	8,376
Pre-reform mean	4.32	5.53	1.88	13.20	10.81

Note: The table shows effects of the Boost for Mathematics on teachers' self-reported training activities. All models include school-by-cohort fixed effects and time-by-cohort fixed effects that vary by municipal and voucher schools. The outcome variable indicated in the column heading is the answer to the survey question: "This academic year, how many hours have you participated in in-service training or other activities that involved; (1) subject knowledge in mathematics, (2) didactics of mathematics, (3) support by a coach, (4) peer collaboration, or (5) student assessment?". Answers are reported as hours per school year. The table reports the p-value of the F-test for the hypothesis that all model parameters are zero. Cluster-adjusted stand-ard errors at the school level are in parentheses and */**/*** refers to statistical significance at the 10/5/1 percent level.

The intensive training phase of the Boost for Mathematics lasts only a year. Even if there are some additional training activities also in the year after

implementation ($\tau = 1$), the initial effect tapers off quickly. Two years after the program was introduced, teachers in participating schools are not more (or less) likely to undertake training than teachers in other schools (the effect in $\tau = 0$ is significantly different from the overall effect in $\tau = 1–2$ for all outcomes). Thus, the Boost for Mathematics has no long-lasting effects on formal in-service training activities, which is not surprising given that the government grant only covered one year of professional development.

An intermediate goal of the program was to instill a collaborative learning culture among teachers, where they continuously learn from each other. Therefore, we have asked teachers to report on their peer collaboration activities. Table 6 shows that teachers are more likely to work together in many different ways during the year of implementation ($\tau = 0$). Teachers more often plan and follow-up their mathematics teaching together with colleagues, and they also discuss didactics more often. These activities are likely to reflect the weekly teacher group meetings as part of the learning cycles. There is no effect on peer observation in the classroom, which is to be expected since the program focuses on self-assessment of classroom performance.

The higher prevalence of peer collaboration is, however, not maintained over time; after the implementation phase ($\tau = 1, 2$) teachers in participating schools are not more likely to collaborate with colleagues than teachers not exposed to the program. We only find a lingering impact on didactic discussions between teachers (the effect in $\tau = 0$ is significantly different from the effect in $\tau = 1–2$ for collegial planning, following-up of lectures and discussions). The results thus suggest that organized peer collaboration needs to be actively promoted by school management to be maintained over time, at least in a Swedish context.[31]

[31] Appendix Table C2, shows that teachers continue to find inspiration from colleagues in improving their teaching also in the year following implementation, potentially through maintained discussions about teaching methods. However, we find no effect for later years. This corroborates the finding that much of the collegial interactions that were spurred in the initial phase of the Boost for Mathematics do not persist.

Table 6. Effects of the Boost for Mathematics on teacher peer collaboration activities (frequency per term)

Column:	(1)	(2)	(3)	(4)	(5)
Outcome:	Plan teaching	Follow-up teaching	Assess students	Discuss didactics	Classroom visits
Implementation year	3.55***	2.64***	0.76	4.55***	-0.02
	(0.66)	(0.61)	(0.58)	(0.60)	(0.42)
1 year after implementation	1.30	0.75	0.61	1.54*	0.02
	(0.92)	(0.81)	(0.77)	(0.81)	(0.61)
2 years after implementation	1.67	1.55	1.59	1.76	-0.38
	(1.20)	(1.17)	(1.08)	(1.09)	(0.83)
School×Wave FE	Yes	Yes	Yes	Yes	Yes
Year×Private×Wave FE	Yes	Yes	Yes	Yes	Yes
F-test for $\theta_0=\theta_1=\theta_2=0$	0.0000	0.0000	0.3935	0.0000	0.9175
Observations	8,370	8,359	8,347	8,384	8,381
Pre-reform mean	10.28	8.92	9.59	12.27	2.57

Note: The table shows effects of the Boost for Mathematics on teachers' self-reported peer collaboration activities. All models include school-by-cohort fixed effects and time-by-cohort fixed effects that vary by municipal and voucher schools. The outcome variable indicated in the column heading is the answer to the survey question: "How often do you, together with another mathematics teacher; (1) plan teaching, (2) follow up on teaching, (3) follow up students' knowledge, (4) discuss instructional practices, or (5) visit each other's lessons to exchange experiences?". Answers are reported as frequency per term. The table reports the p-value of the F-test for the hypothesis that all model parameters are zero. Cluster-adjusted standard errors at the school level are in parentheses and */**/*** refers to statistical significance at the 10/5/1 percent level.

In order for the Boost for Mathematics to influence students' mathematics performance, the program must change teachers' classroom practices. In Table 7 we find that teachers in the program on average spend more time in the classroom discussing problem-solving strategies with the students as well as organizing other types of teaching activities. They also allocate less time in class for students to work standard problems (alone or in groups), which is otherwise a common instructional practice in Swedish schools (Mullis, Martin, and Foy 2008). This is in line with the program's stronger focus on active instructional practices. There is no impact on the time teachers spend lecturing on mathematical material, or on time assigned for tests and homework quizzes.

Table 7. Effects of the Boost for Mathematics on teachers' classroom practices (share of lecture time)

Column:	(1)	(2)	(3)	(4)	(5)
Outcome:	Teacher lectures	Teacher and students discuss	Students solve problems	Students take tests	Other activities
Implementation year	0.30	1.91**	-3.32***	-0.38	1.49**
	(0.55)	(0.75)	(0.96)	(0.31)	(0.60)
1 year after implementation	0.23	2.69**	-4.05***	-0.31	1.44*
	(0.70)	(1.02)	(1.24)	(0.43)	(0.87)
2 years after implementation	0.47	1.86	-3.35**	-0.77	1.78
	(1.02)	(1.44)	(1.59)	(0.56)	(1.09)
School×Wave FE	Yes	Yes	Yes	Yes	Yes
Year×Private×Wave FE	Yes	Yes	Yes	Yes	Yes
F-test for $\theta_0=\theta_1=\theta_2=0$	0.9522	0.0476	0.0040	0.4267	0.1039
Observations	7,819	7,819	7,819	7,819	7,819
Pre-reform mean	18.22	18.81	50.03	5.42	7.51

Note: The table shows effects of the Boost for Mathematics on teachers' self-reported classroom practices. All models include school-by-cohort fixed effects and time-by-cohort fixed effects that vary by municipal and voucher schools. The outcome variable indicated in the column heading is the answer to the survey question: "In a typical week, what percentage of the lesson time in mathematics do students spend on each of the following activities; (1) listening to lecture-style presentations, (2) discussing problem-solving strategies together with the teacher, (3) working problems on their own or in group, (4) taking tests or quizzes, or (5) other student activities?" Answers are reported as percent of time. The table reports the p-value of the F-test for the hypothesis that all model parameters are zero. Cluster-adjusted standard errors at the school level are in parentheses and */**/*** refers to statistical significance at the 10/5/1 percent level.

Unlike the training activities, and most of the teacher peer group interactions, the Boost for Mathematics has persistent effects on instructional practices. The initial reduction in the share of lesson time that students work with standard problems is maintained throughout the period. There are also remaining positive effects on the discussion of problem-solving strategies and other activities. The program, thus, seems to be successful in implementing more active teacher practices more permanently (we do not find significant differences between the effects in $\tau = 0$ and in $\tau = 1$–2 for any of the outcomes). This may explain why the

Bost for Mathematics is able to improve student test scores also in a longer-run, even though the increased formal training and collegial learning activities during the initial phase of the program fade out.

The main obstacle to the adoption of Lesson study in American schools, as argued by Rappleye and Komatsu (2017), is for teachers to take criticism from their peers, which, disrupts or even breaks the learning cycles. To avoid this problem, the Boost for Mathematics instead relies on self-assessments, and, in a national follow-up of the intervention (Ramböll Management Consulting 2014), a majority of the participating teachers characterized the atmosphere of the peer group meetings as open-minded, and also reported to have received constructive feedback from their colleagues. The teachers were also very satisfied with the program in general, as it made them feel more self-confident and engaged in their instruction of mathematics.

Consistent with the national follow-up of the Boost for Mathematics, we find that teachers exposed to the program are boosted in their confidence, and to a larger extent believe they have sufficient competence in mathematics instruction and in assessing the results of their teaching (see Table C3). They also believe that their colleagues have improved in subject knowledge in mathematics and in their didactic competences (see Appendix Table C4). Overall, teachers seem to be positive to the program, which indicates that it was well-designed to meet the needs at the local level.

6 Costs and benefits of the program

The results show that the Boost for Mathematics changed teacher practices and improved student performance. For the program to be a worthwhile investment, however, the benefits must outweigh the costs. In this section, we therefore discuss the societal costs and benefits of the intervention.[32]

[32] See Appendix E for a detailed discussion of the cost-benefit calculations.

One important cost of the Boost for mathematics is the time teachers devote to training, since this is expected to crowd out other out-of-class teacher activities. In the cost-benefit calculations, we assume that half of these activities are directly (or indirectly) related to students' human capital production and, thus, captured by the estimated test score effects. The other half of the reduced teacher activities is instead assumed to produce other societal goods, which we value by the market price of teacher time. For the external tutor, we on the other hand account for the full opportunity cost of the time the tutors devote to the program, since it is unlikely to affect student performance in treated schools. The program also had some direct costs, such as expenditures for training of tutors and principals, setting up the web-portal, and administration. Taken together, we estimate the total costs of the program at about €51.2 million, or €80 per student on average.

The major benefit of the Boost for mathematics is the students' improved mathematics skills. We translate the short-run learning effects to permanent earnings gains using auxiliary data on mathematics performance in grade 6 and life-cycle earnings for a sample of individuals born 1953, and a sample of twins born 1953–82. When we control for differences in both observed and unobserved family background, and adjust for measurement error in observed test scores, we find that 1 SD better mathematics skills is associated with about 9 percent higher life-time earnings. Based on this estimate, we translate the effects of the Boost for Mathematics on performance to life-cycle earnings gains and multiply by the number of students to arrive at an estimated benefit of the program. This back-on-the-envelope calculation yields a benefit of about €1,395 million, or € 2,158 per student on average.

The benefit-to-cost ratio for the Boost for Mathematics is about 27, meaning that the program generates €27 in savings for every €1 spent. It should be stressed, however, that the estimated societal benefits and costs are uncertain, and the effectiveness of the program may change under alternative assumptions. But even if we double the costs and cut the benefits in half, we arrive at a benefit-

to-cost ratio of more than 6. Thus, also under more restrictive assumptions, the Boost for Mathematics appears to be a profitable investment to society.

7 Conclusion

The challenge for teacher professional development programs to successfully enhance student performance, is to influence the interaction between students and teachers in the classroom. Successful small-scale trials may not be generalizable to other settings, and it can be difficult to change the teachers' professional practice by national policy, especially in a decentralized school system. For a national policy to successfully bring new innovations to schools, it must be relevant and adaptive to local needs and motivate teachers to alter their classroom practices. Our results suggest that the Boost for Mathematics manages to do just this.

In 2013 the Swedish government introduced the Boost for Mathematics—a one-year in-service training program for mathematics teachers in compulsory and upper-secondary school—as a response to the deteriorating results of Swedish students in TIMMS 2007 and PISA 2009. The program centers on teacher learning groups supported by a mathematics tutor, in which teachers work in learning cycles. Based on educational modules, with tailored study material, teachers exchange good practices, highlight their difficulties, critically examine their own teaching, and receive feedback from colleagues.

We find that the program improved student performance in mathematics, in particular in primary school. Test scores for the average compulsory school student increased by 2.6 percent of a standard deviation in treated schools during the follow-up period. This indicates that the intervention helps teachers implement more active and effective classroom practices. Importantly, the impact of the program persists also in the longer run, and performance is enhanced also for students who had not yet entered the school when the program was introduced. This suggests that teachers maintain their new instructional practices, something that we also confirm using teacher survey data. However, the effect

sizes are too small to explain the substantial improvement of Swedish students in mathematics (and in other subjects) in PISA 2015 and TIMSS 2015.[33]

The positive effect of the program is concentrated to students in the three top quartiles of predicted mathematics test scores. For the weakest students in the lowest quartile we find no effect. Possibly, the active instructional practices recommended in the program, such as providing challenging tasks and orchestrating group discussions, may have been too advanced for low-performing students. Even if mathematics performance for the average student is boosted by the program, there is a risk that the weakest students are left behind, with reduced equality of opportunity as a result. In addition, the larger effects of the program in primary school suggests that there is an important scope for improving the mathematics instruction of class teachers, who have a more general teacher training than the mathematics subject teachers in lower-secondary school.

We find that it is possible to change teacher behavior in the classroom through national policies in a decentralized schooling system. The program led to lasting changes in classroom practices. Participating teachers devote more time in class to discuss problem-solving strategies together with students, and less time for students to work standard problems. Participating teachers believe they have improved their instructional practices in mathematics, and also that their colleagues have become more skilled. The program also led to an increased peer-to-peer interaction between teachers, but this largely petered out after the program ended, suggesting that peer learning needs to be actively promoted by school management for a collaborative learning culture to be sustained over time.

The Boost for Mathematics contains central elements of the Lesson study methodology. A key difference, however, is that the Boost for Mathematics

[33] The program may explain about 10–15 percent of the overall improvement among Swedish students in TIMSS and PISA.

aims to facilitate information flows between peers focusing on intrinsic motivation where teachers self-assess their classroom performance, rather than on peer observation, with feedback and criticism on classroom practices from colleagues. This is potentially a success factor; Rappleye and Komatsu (2017) argues that an obstacle in introducing Lesson study in the US is an inability of teachers to take criticism from their peers.

We show that the Boost for Mathematics passes a cost-benefit test. Even though the impact on student performance is moderate, the cost of the intervention is even smaller. We estimate that the program generates €27 in return for every €1 invested.

Our results is consistent with recent experimental evidence on the effects of Lesson study on student performance in mathematics. Murphy, Weinhardt, and Wyness (2021) evaluates a teacher peer-to-peer observation and feedback program in 89 English primary schools and find that test scores in mathematics improves by 0.033 (standard error 0.042) standard deviations. Similarly, Burgess, Rawal, and Taylor (2021) report that a teacher peer evaluation program in 41 English upper secondary schools boosts overall test scores, while the impact on mathematics achievement is only 0.044 (standard error 0.031) standard deviations and not statistically significant. We evaluate a Lesson study program at scale, 2,469 Swedish compulsory schools, and find it to have significant and persistent effects on student learning of about the same magnitude as in earlier papers. In fact, our larger sample size enables us to find significant effects of the Lesson study on test scores in mathematics, even though the boost in performance is moderate.

More generally, we conclude that we can learn from other school systems, and that educational strategies of Asian countries can be successfully modified and adapted to Western contexts by national policy.

References

Akiba, Motoko, and Bryan Wilkinson. 2016. “Adopting an International Innovation for Teacher Professional Development: State and District Approaches to Lesson Study in Florida.” *Journal of Teacher Education* 67 (1): 74–93.

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.

Baker, Andrew C. 2019. “Difference-in-Differences Methodology.” *Andrew Baker* (blog). September 25, 2019.

Bill & Melinda Gates Foundation. 2015. “Teachers Know Best: Teachers’ Views on Professional Development.” *Bill & Melinda Gates Foundation*. Bill & Melinda Gates Foundation.

Briole, Simon, and Eric Maurin. 2019. “Does Evaluating Teachers Make a Difference?” SSRN Scholarly Paper ID 3390297. Rochester, NY: Social Science Research Network.

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* In press.

Callaway, Brantly, and Pedro H. C. Sant’Anna. 2020. “Difference-in-Differences with Multiple Time Periods.” *Journal of Econometrics*, December.

Chaisemartin, Clément de, and Xavier D’Haultfœuille. 2020. “Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects.” *American Economic Review* 110 (9): 2964–96.

Chen, Xiangming, and Yurong Zhang. 2019. “Typical Practices of Lesson Study in East Asia.” *European Journal of Education* 54 (2): 189–201.

Cheng, Lu Pien, and Lee Peng Yee. 2012. “A Singapore Case of Lesson Study.” *The Mathematics Educator* 21 (2).

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014a. “Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates.” *American Economic Review* 104 (9): 2593–2632.

———. 2014b. “Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood.” *American Economic Review* 104 (9): 2633–79.

Cilliers, Jacobus, Brahm Fleisch, Cas Prinsloo, and Stephen Taylor. 2019. “How to Improve Teaching Practice? An Experimental Comparison of Centralized Training and in-Classroom Coaching.” *Journal of Human Resources*, February, 0618-9538R1.

Dee, Thomas S., and James Wyckoff. 2015. “Incentives, Selection, and Teacher Performance: Evidence from IMPACT.” *Journal of Policy Analysis and Management* 34 (2): 267–97.

Dix, Katherine, Hilary Hollingsworth, and Toby Carslake. 2018. “Thinking Maths Learning Impact Fund Evaluation Report.” Australian Council for Educational Research.

Garet, Michael S., Andrew J. Wayne, Fran Stancavage, James Taylor, Marian Eaton, Kirk Walters, Mangli Song, Seth Brown, and Steven Hurlburt. 2011. “Middle School Mathematics Professional Development Impact Study: Findings After the Second Year of Implementation.” National Center for Education Evaluation and Regional Assistance 2011–4024. U.S. Department of Education.

Garet, Michael S., Andrew J. Wayne, Fran Stancavage, James Taylor, Kirk Walters, Mangli Song, Seth Brown, and Steven Hurlburt. 2010. “Middle School Mathematics Professional Development Impact Study: Findings After the First Year of Implementation.” National Center for Education Evaluation and Regional Assistance 2010–4009. U.S. Department of Education.

Goldschmidt, Pete, and Geoffrey Phelps. 2010. “Does Teacher Professional Development Affect Content and Pedagogical Knowledge: How Much and for How Long?” *Economics of Education Review* 29 (3): 432–39.

Goodman-Bacon, Andrew. 2018. “Difference-in-Differences with Variation in Treatment Timing.” w25018. National Bureau of Economic Research.

Grönqvist, Erik, Lena Hensvik, and Anna Thoresson. 2021. “Teacher Career Opportunities and School Quality.” *Labour Economics*, May, 101997.

Haeck, Catherine, Pierre Lefebvre, and Philip Merrigan. 2014. “The Distributional Impacts of a Universal School Reform on Mathematical Achievements: A Natural Experiment from Canada.” *Economics of Education Review* 41 (August): 137–60.

Hanushek, Eric A., Guido Schwerdt, Simon Wiederhold, and Ludger Woessmann. 2015. “Returns to Skills around the World: Evidence from PIAAC.” *European Economic Review* 73 (January): 103–30.

Hanushek, Eric A., and Ludger Woessmann. 2008. “The Role of Cognitive Skills in Economic Development.” *Journal of Economic Literature* 46 (3): 607–68.

———. 2016. “Knowledge Capital, Growth, and the East Asian Miracle.” *Science* 351 (6271): 344–45.

Harris, Douglas N., and Tim R. Sass. 2011. “Teacher Training, Teacher Quality and Student Achievement.” *Journal of Public Economics* 95 (7–8): 798–812.

Hinnerich, Björn Tyrefors, and Jonas Vlachos. 2017. “The Impact of Upper-Secondary Voucher School Attendance on Student Achievement. Swedish Evidence Using External and Internal Evaluations.” *Labour Economics*, EALE conference issue 2016, 47 (August): 1–14.

Holmlund, Helena, Anna Sjögren, and Björn Öckert. 2020. “Jämlikhet i möjligheter och utfall i den svenska skolan.” IFAU Rapport 2020:7.

Huang, Rongjin, Yanping Fang, and Xiangming Chen. 2017. “Chinese Lesson Study: A Deliberate Practice, a Research Methodology, and an

Improvement Science." *International Journal for Lesson and Learning Studies* 6 (4): 270–82.

Jackson, C. Kirabo. 2018. "What Do Test Scores Miss? The Importance of Teacher Effects on Non–Test Score Outcomes." *Journal of Political Economy* 126 (5): 2072–2107.

Jackson, C. Kirabo, and Elias Bruegmann. 2009. "Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers." *American Economic Journal: Applied Economics* 1 (4): 85–108.

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.

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.

Jacob, Brian A., and Lars Lefgren. 2008. "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education." *Journal of Labor Economics* 26 (1): 101–36.

Jerrim, John. 2015. "Why Do East Asian Children Perform so Well in PISA? An Investigation of Western-Born Children of East Asian Descent." *Oxford Review of Education* 41 (3): 310–33.

Jerrim, John, and Anna Vignoles. 2016. "The Link between East Asian 'Mastery' Teaching Methods and English Children's Mathematics Skills." *Economics of Education Review* 50 (February): 29–44.

Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger. 2008. "What Does Certification Tell Us about Teacher Effectiveness? Evidence from New York City." *Economics of Education Review* 27 (6): 615–31.

Kennedy, Mary M. 2016. "How Does Professional Development Improve Teaching?" *Review of Educational Research* 86 (4): 945–80.

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.

Lewis, Catherine C., and Ineko Tsuchida. 1999. “A Lesson Is Like a Swiftly Flowing River: How Research Lessons Improve Japanese Education.” *Improving Schools* 2 (1): 48–56.

Lewis, Catherine, and Christine Lee. 2017. *The Global Spread of Lesson Study*. Routledge Handbooks Online.

Lindvall, Jannika, Ola Helenius, Kimmo Eriksson, and Andreas Ryve. 2021. “Impact and Design of a National-Scale Professional Development Program for Mathematics Teachers.” *Scandinavian Journal of Educational Research* 0 (0): 1–16.

Machin, Stephen, and Sandra McNally. 2005. “Gender and Student Achievement in English Schools.” *Oxford Review of Economic Policy* 21 (3): 357–72.

———. 2008. “The Literacy Hour.” *Journal of Public Economics* 92 (5): 1441–62.

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.

Ming Cheung, Wai, and Wing Yee Wong. 2014. “Does Lesson Study Work? : A Systematic Review on the Effects of Lesson Study and Learning Study on Teachers and Students.” *International Journal for Lesson and Learning Studies* 3 (2): 137–49.

Mullis, Ina V. S., Michael O. Martin, and Pierre Foy, eds. 2008. *TIMSS 2007 International Mathematics Report: Findings Form IEA’s Trend in International Mathematics and Science Study at the Fourth and Eighth Grades*. Chestnut Hill, Mass: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Mullis, Ina V. S., Michael O. Martin, Pierre Foy, and Alka Arora. 2012. *Timss 2011 International Results in Mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Mullis, Ina V. S, Michael O. Martin, Pierre Foy, Dana Kelly, and Bethany Fishbein. 2020. "TIMSS 2019 International Results in Mathematics and Science."

Murnane, Richard J., John B. Willett, Yves Duhaldeborde, and John H. Tyler. 2000. "How Important Are the Cognitive Skills of Teenagers in Predicting Subsequent Earnings?" *Journal of Policy Analysis and Management* 19 (4): 547–68.

Murnane, Richard J., John B. Willett, and Frank Levy. 1995. "The Growing Importance of Cognitive Skills in Wage Determination." *The Review of Economics and Statistics* 77 (2): 251–66.

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* In press.

Neal, Derek A., and William R. Johnson. 1996. "The Role of Premarket Factors in Black-White Wage Differences." *Journal of Political Economy* 104 (5): 869–95.

Öckert, Björn. 2021. "School Absenteeism during the COVID-19 Pandemic – How Will Student Performance Be Affected?" In *Swedish Children and Youth during the COVID-19 Pandemic. Evidence from Research on Childhood Environment, Schooling, Educational Choice and Labour Market Entry*. IFAU Working Paper, 2021:3.

OECD. 2014. "PISA 2012 Results: What Students Know and Can Do : Student Performance in Mathematics, Reading and Science (Volume I)."

———. 2019. "PISA 2018 Insights and Interpretations."

Ost, Ben. 2014. "How Do Teachers Improve? The Relative Importance of Specific and General Human Capital." *American Economic Journal: Applied Economics* 6 (2): 127–51.

Pang, JeongSuk. 2016. “Improving Mathematics Instruction and Supporting Teacher Learning in Korea through Lesson Study Using Five Practices.” *ZDM* 48 (4): 471–83.

Papay, John P., and Matthew A. Kraft. 2015. “Productivity Returns to Experience in the Teacher Labor Market: Methodological Challenges and New Evidence on Long-Term Career Improvement.” *Journal of Public Economics* 130 (October): 105–19.

Papay, John P., Eric S. Taylor, John H. Tyler, and Mary E. Laski. 2020a. “Learning Job Skills from Colleagues at Work: Evidence from a Field Experiment Using Teacher Performance Data.” *American Economic Journal: Economic Policy* 12 (1): 359–88.

———. 2020b. “Learning Job Skills from Colleagues at Work: Evidence from a Field Experiment Using Teacher Performance Data.” *American Economic Journal: Economic Policy* 12 (1): 359–88.

Quaresma, Marisa, Carl Winsløw, Stéphane Clivaz, João Pedro da Ponte, Aoibhinn Ní Shúilleabháin, and Akihiko Takahashi, eds. 2018. *Mathematics Lesson Study Around the World: Theoretical and Methodological Issues*. ICME-13 Monographs. Springer International Publishing.

Ramböll Management Consulting. 2014. “Delutvärdering. Matematiklyftets första år.”

Randel, Bruce, Helen Apthorp, Andrea D. Beesley, Tedra F. Clark, and Xin Wang. 2016. “Impacts of Professional Development in Classroom Assessment on Teacher and Student Outcomes.” *The Journal of Educational Research* 109 (5): 491–502.

Randel, Bruce, Andrea D. Beesley, Helen Apthorp, Tedra F. Clark, Xin Wang, Louis F. Cicchinelli, and Jean M. Williams. 2011. “Classroom Assessment for Student Learning: Impact on Elementary School Mathematics in the Central Region.” National Center for Education Evaluation and Regional Assistance 2011–4005. U.S. Department of Education.

Rappleye, Jeremy, and Hikaru Komatsu. 2017. “How to Make Lesson Study Work in America and Worldwide: A Japanese Perspective on the onto-Cultural Basis of (Teacher) Education:” *Research in Comparative and International Education*, November.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. “Teachers, Schools, and Academic Achievement.” *Econometrica* 73 (2): 417–58.

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

Rothstein, Jesse. 2010. “Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement.” *The Quarterly Journal of Economics* 125 (1): 175–214.

Skolinspektionen. 2012. “Lika För Alla? Omrättning av nationella prov i grundskolan och gymnasieskolan under tre år.”

———. 2021. “Ombedömning av Nationella Prov 2019.” Diarienummer 2019:503.

Skollag. 2010. *SFS 2010:400, Utbildningsdepartementet, Stockholm*.

Skolverket. 2012. “Information om Matematiklyftet.” Mimeo Dnr 2012:1958.

———. 2016a. “Slutredovisning av Uppdrag att svara för utbildning.” Mimeo Dnr 2011:643.

———. 2018. “The Boost for Mathematics. Slides from Presentation in Karlstad.”

———. 2020. “Ubildningsstatistik.” Text. December 20, 2020.

Sojourner, Aaron J., Elton Mykerezi, and Kristine L. West. 2014. “Teacher Pay Reform and Productivity Panel Data Evidence from Adoptions of Q-Comp in Minnesota.” *Journal of Human Resources* 49 (4): 945–81.

Stigler, James W., and James Hiebert. 1999. *The Teaching Gap: Best Ideas from the World’s Teachers for Improving Education in the Classroom*. The Free Press, A Division of Simon & Schuster Inc.

Sun, Liyang, and Sarah Abraham. 2020. “Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects.” *Journal of Econometrics*, December.

Taylor, Eric S., and John H. Tyler. 2012. “The Effect of Evaluation on Teacher Performance.” *American Economic Review* 102 (7): 3628–51.

Tyrefors Hinnerich, Björn, and Jonas Vlachos. 2016. “Skillnader i resultat mellan gymnasieelever i fristående och kommunal skolor.” IFAU Rapport 2016:10. Institutet för arbetsmarknads- och utbildningspolitisk utvärdering (IFAU).

Utbildningsdepartementet. 2012. “Uppdrag Att Svara För Utbildning.” Regeringsbeslut I:44.

Wiswall, Matthew. 2013. “The Dynamics of Teacher Quality.” *Journal of Public Economics* 100 (April): 61–78.

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.” 2007–033. Issues & Answers. REL. Regional Educational Laboratory Southwest.

For online publication

Appendix

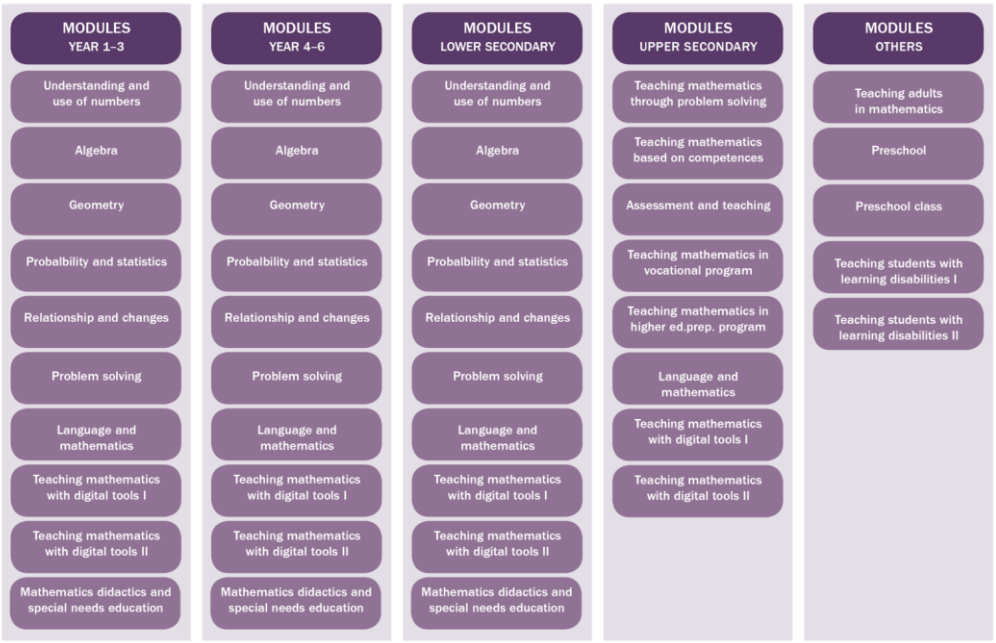
A. Content of the Boost for Mathematics

The Boost for Mathematics is based on educational modules with didactic support material (available online) covering different mathematical content. For each stage of compulsory school and upper secondary school there are separate modules, which are adapted to the didactic challenges at the specific level of schooling. Compulsory school has 10 different educational modules at each stage covering different mathematical themes; see Figure A1 for a full list of modules. All modules address the theme from the didactic perspectives: formative assessment or assessment for learning; competencies in the Swedish curriculum; classroom norms and socio-mathematical norms; interaction in the classroom (for details see Lindvall et al. 2021). There can also be additional didactic perspectives in the modules e.g., ICT, a historical perspective, or variation theory of learning.

The support material (e.g., texts, articles, films, and mathematics problems) in the modules is based on courses and syllabi, research on learning and teaching mathematics, and analyses of Swedish students' performance in national and international assessments. To ensure the quality and relevance of the didactic support material, each module is developed by two universities or teacher training colleges in collaboration, where the content is assessed by independent researchers in a peer review process. Focus groups of teachers have also been involved in this process. All modules consist of 8 parts, with each working through a learning cycle of 4 steps; see Figure A2 for a typology.

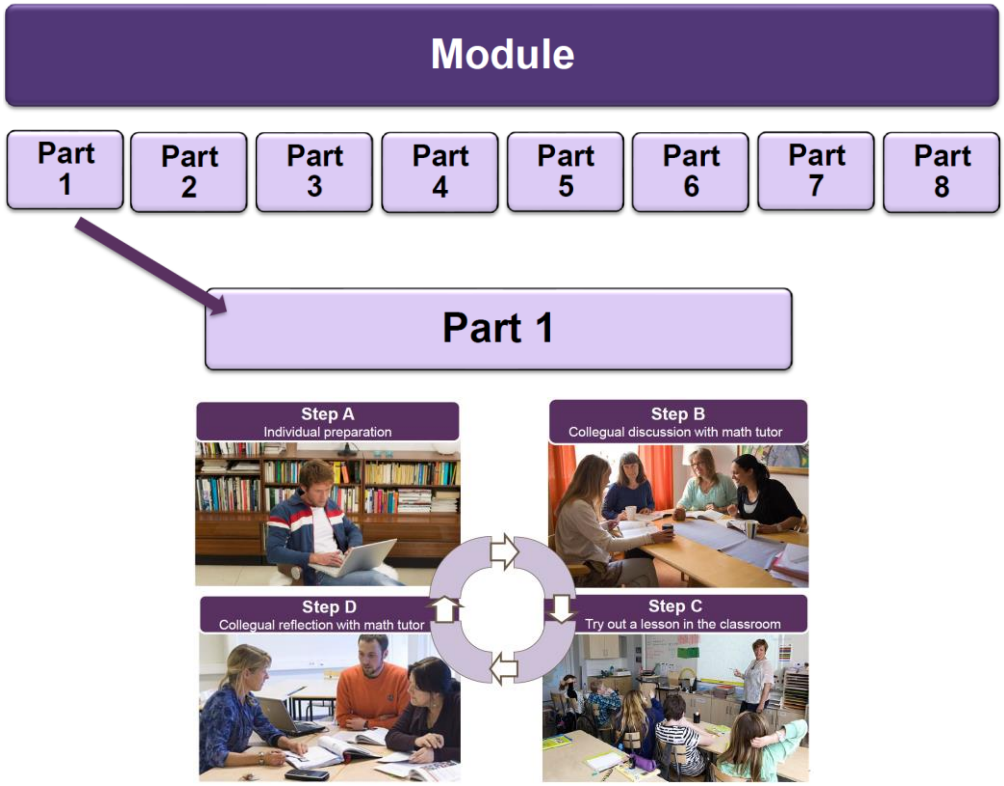
The set-up of the program is based on the local needs of the school and it is the principal together with the tutor and teacher group – and in collaboration with the school district – that decides on which two modules to work with. The local principal is responsible for organizing the teacher groups and allocating time for training activities within the regular working hours.

Figure A1. Content of modules



Source: (Skolverket 2018)

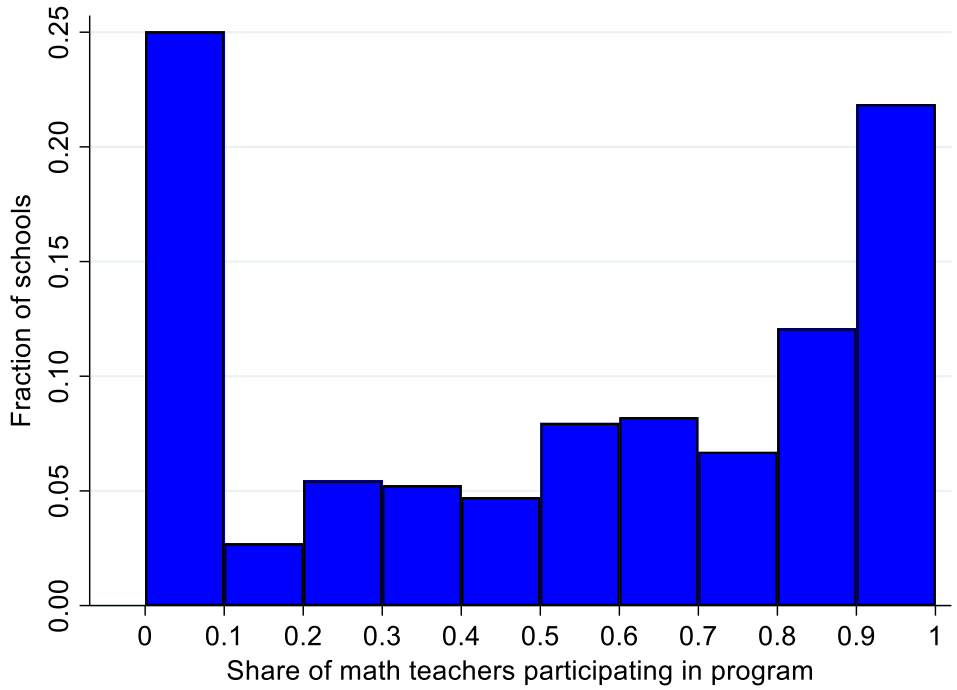
Figure A2. Illustration of the learning cycle in the module



Source: (Skolverket 2018)

B. Additional results

Figure B1. Distribution of the share of participating teachers in schools



Note: The figure shows the distribution of schools with different share of mathematics teachers that receive the government grant for participating in the Boost for Mathematics.

Table B1. Effects of the Boost for Mathematics on mathematics test scores. Alternative treatment definitions

Column:	(1)	(2)	(3)
Treatment cutoff:	0.50	0.20	0.80
All years pooled	0.0263***	0.0248***	0.0245**
	(0.0085)	(0.0084)	(0.0096)
School×Wave FE	Yes	Yes	Yes
Year×Private×Wave FE	Yes	Yes	Yes
Observations	2,874,158	3,259,047	2,253,493

Note: The table shows reduced form effects of the Boost for Mathematics on standardized test scores in mathematics for different definitions of schools' treatment status. The treatment cutoff values indicated in the column heading is the lowest share of mathematics teachers participating in the program required for the school to be defined as treated. All models include school-by-cohort fixed effects and time-by-cohort fixed effects that vary by municipal and voucher schools. Test scores are measured in the end of lower/middle/higher stage (grade 3/6/9). Students are sampled in the beginning of the lower/middle/higher stage (grades 1/4/6) and assigned the treatment status of the school they are expected to attend in the end of the stage. Cluster-adjusted standard errors at the school level are in parentheses and */**/*** refers to statistical significance at the 10/5/1 percent level.

Table B2. Effects of the Boost for Mathematics on actual exposure to the program (first stage)

Column:	(1)	(2)	(3)	(4)
Grades:	3, 6 and 9	3	6	9
All years pooled	0.8444***	0.9012***	0.8395***	0.7697***
	(0.0039)	(0.0043)	(0.0061)	(0.0077)
School×Wave FE	Yes	Yes	Yes	Yes
Year×Private×Wave FE	Yes	Yes	Yes	Yes
Observations	2,874,158	1,053,814	967,565	852,779

Note: The table shows reduced form effects of the student's expected exposure to the Boost for Mathematics on actual exposure. The outcome variable is years of exposure to the program in the school the student attends in the end of lower/middle/higher stage (grade 3/6/9). Students are sampled in the beginning of the lower/middle/higher stage (grades 1/4/6) and assigned the years of exposure to the program in the school they are expected to attend in the end of the stage. All models include school-by-cohort fixed effects and time-by-cohort fixed effects that vary by municipal and voucher schools. The sample studied is indicated in the column heading. Cluster-adjusted standard errors at the school level are in parentheses and */**/*** refers to statistical significance at the 10/5/1 percent level.

Table B3. Effects of the Boost for Mathematics on test scores in mathematics. Alternative levels of cluster-adjusted standard errors.

Column:	(1)	(2)	(3)
	Panel A. Separately for different years		
Implementation year	0.0117	0.0117	0.0117
	0.0073	(0.0072)	(0.0075)
1 year after implementation	0.0249***	0.0249***	0.0249***
	(0.0089)	(0.0084)	(0.0090)
2 years after implementation	0.0343***	0.0343***	0.0343***
	(0.0103)	(0.0115)	(0.0105)
3 years after implementation	0.0352***	0.0352***	0.0352***
	(0.0113)	(0.0130)	(0.0113)
4 years after implementation	0.0322**	0.0322**	0.0322**
	(0.0137)	(0.0148)	(0.0134)
5 years after implementation	0.0167	0.0167	0.0167
	(0.0162)	(0.0174)	(0.0165)
	Panel B. All years pooled		
All years	0.0263***	0.0263***	0.0263***
	(0.0085)	(0.0092)	(0.0086)
School×Wave FE	Yes	Yes	Yes
Year×Private×Wave FE	Yes	Yes	Yes
Cluster level	School	School district	School × stage
Observations	2,874,158	2,874,158	2,874,158

Note: The table shows reduced form effects of the Boost for Mathematics on test scores in mathematics using cluster-adjusted standard errors at different levels. All models include school-by-cohort fixed effects and time-by-cohort fixed effects that are allowed to vary by municipal and voucher schools. Test scores are measured in the end of lower/middle/higher stage (grade 3/6/9). Students are sampled in the beginning of the lower/middle/higher stage (grades 1/4/6) and assigned the treatment status of the school they are expected to attend in the end of the stage. */**/*** refers to statistical significance at the 10/5/1 percent level.

Table B4. Effects of the Boost for Mathematics on test scores in mathematics, by stage

Column:	(1)	(2)	(3)	(4)
Grades:	3, 6 and 9	3	6	9
Implementation year	0.0117	0.0104	0.0185	0.0069
	0.0073	(0.0153)	(0.0117)	(0.0105)
1 year after implementation	0.0249***	0.0403**	0.0325**	0.0041
	(0.0089)	(0.0177)	(0.0146)	(0.0129)
2 years after implementation	0.0343***	0.0603***	0.0225	0.0175
	(0.0103)	(0.0197)	(0.0168)	(0.0173)
3 years after implementation	0.0352***	0.0582***	0.0318*	0.0002
	(0.0113)	(0.0208)	(0.0174)	(0.0190)
4 years after implementation	0.0322**	0.0497**	0.0416**	0.0144
	(0.0137)	(0.0244)	(0.0207)	(0.0255)
5 years after implementation	0.0167	0.0583*	0.0282	-0.0286
	(0.0162)	(0.0316)	(0.0280)	(0.0241)
School×Wave FE	Yes	Yes	Yes	Yes
Year×Private×Wave FE	Yes	Yes	Yes	Yes
Observations	2,874,158	1,053,814	967,565	852,779

Note: The table shows reduced form effects of the Boost for Mathematics on standardized test scores in mathematics, by stage. All models include school-by-cohort fixed effects and time-by-cohort fixed effects that vary by municipal and voucher schools. The sample studied is indicated in the column heading. Test scores are measured in the end of lower/middle/higher stage (grade 3/6/9). Students are sampled in the beginning of the lower/middle/higher stage (grades 1/4/6) and assigned the treatment status of the school they are expected to attend in the end of the stage. Cluster-adjusted standard errors at the school level are in parentheses and */**/*** refers to statistical significance at the 10/5/1 percent level.

Table B5 Specification tests

Column:	(1)	(2)	(3)
Outcome:	Predicted test scores	Test scores	Test scores
Implementation year	0.0019	0.0101	0.0110
	(0.0014)	(0.0071)	(0.0073)
1 year after implementation	-0.0005	0.0250***	0.0239***
	(0.0018)	(0.0086)	(0.0089)
2 years after implementation	0.0010	0.0321***	0.0319***
	(0.0020)	(0.0099)	(0.0104)
3 years after implementation	0.0024	0.0322***	0.0327***
	(0.0024)	(0.0108)	(0.0115)
4 years after implementation	0.0036	0.0267**	0.0300**
	(0.0031)	(0.0132)	(0.0139)
5 years after implementation	0.0048	0.0120	0.0155
	(0.0041)	(0.0158)	(0.0165)
School×Wave FE	Yes	Yes	Yes
Year×Private×Wave FE	Yes	Yes	Yes
Student controls	No	Yes	No
School intervention controls	No	No	Yes
Observations	2,874,158	2,874,158	2,874,158

Note: The table shows reduced form effects of the Boost for Mathematics on predicted test scores and test scores in mathematics. All models include school-by-cohort fixed effects and time-by-cohort fixed effects that vary by municipal and voucher schools. The outcome variable is indicated in the column heading. The student controls are gender, birth month, income of mother, income of father, education of mother, education of father, immigrant status and indicators for having missing values. The school intervention controls are dummy variables for the schools' participation in the Boost for Reading, Career teachers, Teachers' salary boost and the reintroduction of the Boost for Mathematics in 2017. Outcomes are measured in the end of lower/middle/higher stage (grade 3/6/9). Students are sampled in the beginning of the lower/middle/higher stage (grades 1/4/6) and assigned the treatment status of the school they are expected to attend in the end of the stage. Cluster-adjusted standard errors at the school level are in parentheses and */**/*** refers to statistical significance at the 10/5/1 percent level.

Table B6. Effects of the Boost for Mathematics on test-taking, by quartile of predicted test scores

Column:	(1)	(2)	(3)	(4)	(5)
Sample:	All	P0–P25	P25–P50	P50–P75	P75–P100
All years pooled	0.0011	0.0028	0.0009	0.0032	-0.0063
	(0.0050)	(0.0060)	(0.0057)	(0.0055)	(0.0062)
School×Wave FE	Yes	Yes	Yes	Yes	Yes
Year×Private×Wave FE	Yes	Yes	Yes	Yes	Yes
Observations	3,079,940	783,804	765,007	762,572	768,557
Mean of outcome	0.9332	0.8920	0.9420	0.9513	0.9485

Note: The table shows reduced form effects of the Boost for Mathematics on the probability to take the standardized test in mathematics, divided by quartile of students' predicted test scores. All models include school-by-cohort fixed effects and time-by-cohort fixed effects that are allowed to vary by municipal and voucher schools. The sample studied is indicated in the column heading. The outcome is measured in the end of lower/middle/higher stage (grade 3/6/9). Students are sampled in the beginning of the lower/middle/higher stage (grades 1/4/6) and assigned the treatment status of the school they are expected to attend in the end of the stage. Cluster-adjusted standard errors at the school level are in parentheses and */**/*** refers to statistical significance at the 10/5/1 percent level.

Table B7. Effects of the Boost for Mathematics on test scores in mathematics, by immigration status and gender

Column:	(1)	(2)	(3)	(4)
Sample:	Native	Immigrant	Girls	Boys
All years pooled	0.0276***	0.0096	0.0324***	0.0208**
	(0.0085)	(0.0212)	(0.0093)	(0.0101)
School×Wave FE	Yes	Yes	Yes	Yes
Year×Private×Wave FE	Yes	Yes	Yes	Yes
Observations	2,657,401	543,227	1,405,750	1,468,408

Note: The table shows reduced form effects of the Boost for Mathematics on standardized test scores in mathematics, divided by immigration status and gender. All models include school-by-cohort fixed effects and time-by-cohort fixed effects that are allowed to vary by municipal and voucher schools. The sample studied is indicated in the column heading. Test scores are measured in the end of lower/middle/higher stage (grade 3/6/9). Students are sampled in the beginning of the lower/middle/higher stage (grades 1/4/6) and assigned the treatment status of the school they are expected to attend in the end of the stage. Cluster-adjusted standard errors at the school level are in parentheses and */**/*** refers to statistical significance at the 10/5/1 percent level.

Table B8. Effects of the Boost for Mathematics on test scores in mathematics for natives, by quartile of predicted test scores

Column:	(1)	(2)	(3)	(4)	(5)
Sample:	All	P0–P25	P25–P50	P50–P75	P75–P100
All years pooled	0.0276***	0.0080	0.0300***	0.0372***	0.0211**
	(0.0085)	(0.0160)	(0.0108)	(0.0106)	(0.0102)
School×Wave FE	Yes	Yes	Yes	Yes	Yes
Year×Private×Wave FE	Yes	Yes	Yes	Yes	Yes
Observations	2,657,401	562,095	687,497	699,246	708,563

Note: The table shows reduced form effects of the Boost for Mathematics on standardized test scores in mathematics for natives, divided by quartile of students' predicted test scores. The quartiles are defined in the full population (natives and immigrants). All models include school-by-cohort fixed effects and time-by-cohort fixed effects that are allowed to vary by municipal and voucher schools. The sample studied is indicated in the column heading. Test scores are measured in the end of lower/middle/higher stage (grade 3/6/9). Students are sampled in the beginning of the lower/middle/higher stage (grades 1/4/6) and assigned the treatment status of the school they are expected to attend in the end of the stage. Cluster-adjusted standard errors at the school level are in parentheses and */**/*** refers to statistical significance at the 10/5/1 percent level.

Table B9. Effects of the Boost for Mathematics on test scores in mathematics, by teacher characteristics

Column:	(1)	(2)	(3)	(4)
Characteristic:	Certified		Experienced	
Sample:	High share	Low share	High share	Low share
All years pooled	0.0270**	0.0294**	0.0316***	0.0248*
	(0.0119)	(0.0135)	(0.0118)	(0.0137)
School×Wave FE	Yes	Yes	Yes	Yes
Year×Private×Wave FE	Yes	Yes	Yes	Yes
Observations	1,283,925	1,381,908	1,289,374	1,376,459

Note: The table shows reduced form effects of the Boost for Mathematics on standardized test scores in mathematics, divided by the schools' average teacher characteristics. All models include school-by-cohort fixed effects and time-by-cohort fixed effects that are allowed to vary by municipal and voucher schools. The sample studied is indicated in the column heading. Test scores are measured in the end of lower/middle/higher stage (grade 3/6/9). Students are sampled in the beginning of the lower/middle/higher stage (grades 1/4/6) and assigned the treatment status of the school they are expected to attend in the end of the stage. Cluster-adjusted standard errors at the school level are in parentheses and */**/*** refers to statistical significance at the 10/5/1 percent level.

Table B10. Effects of the Boost for Mathematics on test scores in mathematics, by school characteristics

Column:	(1)	(2)	(3)	(4)
Characteristic:	School size		School market	
Sample:	Small	Big	Big city	Smaller city
All years pooled	0.0164	0.0415***	0.0508***	0.0124
	(0.0112)	(0.0143)	(0.0141)	(0.0104)
School×Wave FE	Yes	Yes	Yes	Yes
Year×Private×Wave FE	Yes	Yes	Yes	Yes
Observations	1,477,846	1,271,349	1,121,352	1,752,806

Note: The table shows reduced form effects of the Boost for Mathematics on standardized test scores in mathematics, divided by school characteristics. All models include school-by-cohort fixed effects and time-by-cohort fixed effects that are allowed to vary by municipal and voucher schools. The sample restriction is indicated in the column heading. Test scores are measured in the end of lower/middle/higher stage (grade 3/6/9). Students are sampled in the beginning of the lower/middle/higher stage (grades 1/4/6) and assigned the treatment status of the school they are expected to attend in the end of the stage. Cluster-adjusted standard errors at the school level are in parentheses and */**/*** refers to statistical significance at the 10/5/1 percent level.

Table B11. Effects of the Boost for Mathematics on test scores in Swedish, by stage

Column:	(1)	(2)	(3)	(4)
Grades:	3, 6 and 9	3	6	9
All years pooled	0.0188**	0.0236**	0.0153	0.0207
	(0.0077)	(0.0118)	(0.0136)	(0.0130)
School×Wave FE	Yes	Yes	Yes	Yes
Year×Private×Wave FE	Yes	Yes	Yes	Yes
Observations	2,923,080	1,054,511	983,587	884,982

Note: The table shows reduced form effects of the Boost for Mathematics on standardized test scores in Swedish, divided by stage. All models include school-by-cohort fixed effects and time-by-cohort fixed effects that are allowed to vary by municipal and voucher schools. The sample studied is indicated in the column heading. Test scores Outcomes are measured in the end of lower/middle/higher stage (grade 3/6/9). Students are sampled in the beginning of the lower/middle/higher stage (grades 1/4/6) and assigned the treatment status of the school they are expected to attend in the end of the stage. Cluster-adjusted standard errors at the school level are in parentheses and */**/*** refers to statistical significance at the 10/5/1 percent level.

C. Additional teacher survey results

Table C1. Effects of the Boost for Mathematics on pre-determined characteristics among teachers who responded to the survey

Column:	(1)	(2)	(3)	(4)
Outcome:	Years of experience	Hours of teaching per week	Teacher diploma	University semesters in mathematics
	Panel A. Separately for different years			
Implementation year	0.86	0.02	0.01	0.01
	(0.58)	(0.21)	(0.01)	(0.07)
1 year after implementation	-0.52	0.16	0.00	-0.02
	(0.79)	(0.26)	(0.02)	(0.10)
2 years after implementation	-0.27	0.11	-0.01	-0.06
	1.07	(0.36)	(0.03)	(0.15)
		Panel B. All years pooled		
All years	0.18	0.08	0.003	-0.01
	(0.67)	(0.22)	(0.015)	(0.08)
School×Wave FE	Yes	Yes	Yes	Yes
Year×Private×Wave FE	Yes	Yes	Yes	Yes
Observations	8,363	8,166	8,385	8,314
Mean of dependent var	15.26	5.60	0.95	1.76

Note: The table shows effects of the Boost for Mathematics on pre-determined characteristics among teachers who responded to the survey. All models include school-by-cohort fixed effects and time-by-cohort fixed effects that vary by municipal and voucher schools. Cluster-adjusted standard errors at the school level are in parentheses and */**/*** refers to statistical significance at the 10/5/1 percent level.

Table C2. Effects of the Boost for Mathematics on teachers' sources of inspiration for improving their instruction

Column:	(1)	(2)	(3)	(4)	(5)
Outcome:	School management	Colleagues	Educational web-platforms	Literature	Seminars and conferences
Implementation year	0.38	4.00***	0.93	2.08***	1.98***
	(0.28)	(0.57)	(0.58)	(0.54)	(0.34)
1 year after implementation	0.01	2.41***	1.09	0.90	0.07
	(0.37)	(0.81)	(0.78)	(0.70)	(0.41)
2 years after implementation	0.25	1.69	1.68	0.13	0.45
	(0.49)	(1.19)	(1.10)	(0.98)	(0.50)
School×Wave FE	Yes	Yes	Yes	Yes	Yes
Year×Private×Wave FE	Yes	Yes	Yes	Yes	Yes
F-test for $\theta_0=\theta_1=\theta_2=0$	0.3348	0.0000	0.4013	0.0000	0.0000
Observations	8,089	8,315	8,250	8,295	8,260
Pre-reform mean	1.66	13.64	9.11	7.83	2.38

Note: The table shows effects of the Boost for Mathematics on teachers' self-reported sources of inspiration for improving their instruction. All models include school-by-cohort fixed effects and time-by-cohort fixed effects that vary by municipal and voucher schools. The outcome variable indicated in the column heading is the answer to the survey question: "How often do you get inspiration and knowledge to improve your instruction from; (1) the school management, (2), colleagues, (3) educational web-platforms, (4) literature (e.g. books and research papers), or (5) seminars and conferences?" Answers are reported as: "At least once a week" (25 times per semester); "At least once a month" (12); "At least once per semester" (3); "More rarely/never" (0). The table reports the p-value of the F-test for the hypothesis that all model parameters are zero. Cluster-adjusted standard errors at the school level are in parentheses and */**/*** refers to statistical significance at the 10/5/1 percent level.

Table C3. Effects of the Boost for Mathematics on teachers' self-assessment of their knowledge and competences

Column:	(1)	(2)	(3)
Outcome:	Subject knowledge in Mathematics	Mathematics didactics	Assessing the results of teaching
Implementation year	0.13**	0.21***	0.21***
	(0.06)	(0.06)	(0.05)
1 year after implementation	0.13*	0.06	0.15**
	(0.08)	(0.08)	(0.07)
2 years after implementation	0.22**	0.11	0.24**
	(0.11)	(0.11)	(0.09)
School×Wave FE	Yes	Yes	Yes
Year×Private×Wave FE	Yes	Yes	Yes
F-test for $\theta_0=\theta_1=\theta_2=0$	0.1235	0.0003	0.0010
Observations	8,393	8,374	8,365

Note: The table shows effects of the Boost for Mathematics on teachers' self-assessment of their knowledge and competences. All models include school-by-cohort fixed effects and time-by-cohort fixed effects that vary by municipal and voucher schools. The outcome variable indicated in the column heading is the answer to the survey question: "To what extent do you think you have sufficient knowledge and competence in; (1) mathematics, (2) methodology and didactics of mathematics, or (3) following up the results of your mathematics teaching?" Answers are reported as: "To a very high degree" (5); "To a high degree" (4); "To neither a high nor a low degree" (3); "To a low degree" (1); "To a very low degree" (1). The outcome variable has been standardized. The table reports the p-value of the F-test for the hypothesis that all model parameters are zero. Cluster-adjusted standard errors at the school level are in parentheses and */**/*** refers to statistical significance at the 10/5/1 percent level.

Table C4. Effects of the Boost for Mathematics on teachers' opinion of their school

Column:	(1)	(2)	(3)
Outcome:	Principals' pedagogical leadership	Colleagues' subject knowledge in Mathematics	Colleagues' knowledge in didactics of Mathematics
Implementation year	0.01	0.12*	0.19***
	(0.08)	(0.07)	(0.07)
1 year after implementation	-0.05	0.07	0.20**
	(0.13)	(0.09)	(0.09)
2 years after implementation	-0.02	0.15	0.22*
	(0.16)	(0.15)	(0.13)
School×Wave FE	Yes	Yes	Yes
Year×Private×Wave FE	Yes	Yes	Yes
F-test for $\theta_0=\theta_1=\theta_2=0$	0.8113	0.2765	0.0401
Observations	8,076	8,023	7,894

Note: The table shows effects of the Boost for Mathematics on teachers' self-reported opinion of their school. All models include school-by-cohort fixed effects and time-by-cohort fixed effects that vary by municipal and voucher schools. The outcome variable indicated in the column heading is the answer to the survey question: "How do you think the following is at your school; (1) the principal's pedagogical leadership, (2) the mathematics teachers' subject knowledge in mathematics, and (3) the mathematics teachers' knowledge of methodology and didactics in mathematics?" Answers are reported as: "Very good" (5); "Good" (4); "Neither good nor bad" (3); "Bad" (2); "Very bad" (1). The outcome variable has been standardized. The table reports the p-value of the F-test for the hypothesis that all model parameters are zero. Cluster-adjusted standard errors at the school level are in parentheses and */**/*** refers to statistical significance at the 10/5/1 percent level.

D. Reliability of test scores

In Swedish schools, mathematics teachers grade their own students' national exams. A relevant question is therefore whether the estimated effects of the Boost for Mathematics on test scores reflect changes in teachers' grading standards rather than improved student performance. Even though the Swedish National Agency for Education provides detailed guidelines on how to assess different answers, and promotes co-grading, it is still possible that participating teachers adopt less (or more) stringent grading standards.[34] In this section, we provide three pieces of evidence suggesting that the estimated effects are likely to reflect improved student performance rather than changes in teachers' grading standards.

First, there is little room for teachers' subjective judgement of students' answers to questions that can be characterized as being either 'right' or 'wrong', as is often the case in mathematics. This is confirmed by the re-assessments of national exams conducted by the Swedish Schools Inspectorate for a sample of schools every year (see e.g. Skolinspektionen 2021). Teachers are often found to be more lenient when judging their own students than are the external graders, but the magnitudes differ considerably across subjects. In Swedish, the deviation in test scores between the school teacher and the external examiner is on average more than 20 percent of a standard deviation (of the externally graded test score). The corresponding number for the national exams in mathematics is about 5 percent of a standard deviation (Skolinspektionen 2012). Thus, the teachers' judgement of their own students' mathematics performance does not differ much from that of external examiners.

Second, to further investigate the subjectiveness of teachers' assessment, we make use of data from TIMSS, which is an international assessment of student performance in mathematics and science in grades 4 and 8, conducted by the

[34] To the extent that the program helps teachers to make more reliable (less noisy) assessments of student performance, this would not bias the estimates (but rather make them more precise).

International Association for the Evaluation of Educational Achievement (IEA). We have access to the TIMSS 2015 survey for Sweden, matched to the students' national exams in grades 3, 6 and 9. This enables us to compare the effect of teachers (intraclass correlations) for student performance in mathematics on the national tests (internally graded) and on the TIMSS test (externally graded). We find that the teacher effects are of the same order of magnitude for both tests; 0.267 (0.013) for the national tests and 0.249 (0.013) for TIMSS, which, again, suggests that there is little room for teachers' subjective grading in mathematics.

Third, as a final check for any impact of the Boost for Mathematics on teachers' grading standards, we exploit information on program participation obtained from the Swedish version of the TIMSS 2015 school questionnaire. The schools were asked to state the share of mathematics teachers who participated in the Boost for mathematics (in the 2013/14 or in the 2014/15 school years). Similar to our main analysis, we define schools where at least half of the teachers participate in the program as treated, and schools with no participating teachers as untreated. We can, thus, compare the difference in student performance in mathematics between participating and non-participating schools using both the internally graded national exams (in grades 3 and 9) and the externally graded TIMSS test (in grades 4 and 8).[35]

Appendix Table D1, column 1, shows that students in schools participating in the Boost for Mathematics perform on average about 0.05 SD better on the national tests in mathematics than students in schools that do not participate. However, the difference is not significant. In column 2, we attempt to adjust for some of the selection to the program by adding pre-determined student characteristics, which reduces the differences between schools slightly. In the last two columns of Appendix Table D1, we repeat the same exercise using the externally graded TIMSS test. Column 3 reveals that students in treated schools score

[35] Since TIMSS 2015 is a cross-sectional data set we are unable to control for fixed differences between schools, and the difference in performance between treated and untreated schools may therefore not be given a causal interpretation.

on average about 0.05 SD higher than other students on the TIMSS test (not significant). Again, adding pre-determined student characteristics reduces the estimates somewhat.[36] Thus, the estimated difference in student performance between schools participating in the Boost for Mathematics, and schools not participating, is very similar if we use the internally graded national exams or the externally graded TIMSS test, which indicates that the program had very minor, if any, effects on teachers' grading standards in mathematics.

Table D1. Descriptive differences in student performance in mathematics between schools participating in the Boost for Mathematics and schools that do not, using data on national tests and the TIMSS test

Column:	(1)	(2)	(3)	(4)
Outcome:	Test scores	Test scores	TIMSS test	TIMSS test
Grades:	3 and 9	3 and 9	4 and 8	4 and 8
All years pooled	0.047	0.037	0.045	0.026
	(0.068)	(0.058)	(0.063)	(0.041)
Student controls	No	Yes	No	Yes
Number of students	7,142	7,142	7,581	7,581
Number of schools	270	270	270	270

Note: The table shows differences in student performance in mathematics for schools participating in the Boost for Mathematics and schools that do not participate, using data on national tests (grades 3 and 9) and the TIMSS 2015 test (grades 4 and 8), respectively. The data have been provided by the National Agency of Education. Schools are defined as being treated if at least half of the mathematics teachers in the school participate in the program, and untreated if no teacher participate. All stages have been pooled and the models include a dummy variable for grade level. The student controls consist of dummy variables for month of birth, gender, first- and second-generation immigrant, age at immigration, and mother's and father's highest educational level. The outcome variable studied is indicated in the column heading. Cluster-adjusted standard errors at the school level are in parentheses and */**/*** refers to statistical significance at the 10/5/1 percent level.

[36] This result is consistent with Lindvall et al. (2021) that also use data from TIMSS 2015. They do not find any significant performance differences between students taught by teachers participating in the Boost for Mathematics and other students.

E. Cost and benefit calculations

The evaluation of the Boost for Mathematics captures the short-run effects on mathematics skills. To inform policy about the efficiency of the program, however, it is necessary to also take the costs and long-run benefits of the intervention into account. In this section, we attempt to attach a monetary value to the societal costs and benefits of the program compared to the situation had it not been introduced.

Costs

During the implementation phase of the Boost for Mathematics, participating teachers devote about 60 hours of their time to the learning cycles. The training is required to take place during regular working hours, and is, thus, expected to crowd out other out-of-class teacher activities. We lack time-use data for participating teachers but assume that half of their non-teaching activities – such as preparation, interaction with students and parents, and other types of professional development – are directly (or indirectly) related to students' human capital production, whereas the other half – such as school management, administration, and extracurricular activities – produce other outputs valuable to society.

To the extent that the program infringes on out-of-class activities that matter for skill formation, this will be captured by the estimated test score effects. This is, however, not the case for other types of teacher activities, and we therefore value the lost production of other societal goods by the market price of teacher time. The gross hourly wage (including payroll taxes) for participating teachers is €28.6 (in 2020 prices). Since we assume that half of the 60 training hours crowd out production of other societal goods, we estimate the cost of training at €858 per teacher (€28.6×60×0.50). In all, 23,209 teachers participated in the program in the schools covered by the evaluation, and the total cost for all teachers is, thus, about €19.9 million (€858 ×23,209).

The external tutors are expected to spend 20 percent of their time to prepare and coach teachers, which corresponds to about 400 hour per school year. We take the full opportunity cost of the time the external tutors spend on the program into account, since it is not likely to affect student performance in treated schools. Assuming that tutors have the same wage as the average participating teacher, the cost is estimated at €11,440 per tutor (€28.6×400). There were 1,360 tutors hired in the schools covered by the evaluation, adding up to a cost of about €15.6 million (€11,440×1,360).

The program also involved other costs, such as the training of tutors and principals, setting up the web-portal, administration, etc., amounting to €15.7 million (Skolverket 2016a).[37] The grand total cost of the program is, thus, estimated to be about €51.2 million (€19.9 + €15.6 + €15.7 million). In all, 646,267 unique students were exposed to the program at some point (in the schools covered by the evaluation), yielding an average cost per student of about €80 (€51.2 million / 646,267 students).

Benefits

The major benefit of the Boost for mathematics is the value of the students' improved mathematics performance. We translate the short-run learning effects to life-time earning gains using data from the 'Evaluation-through-follow-up' (ETF) project. The ETF data includes information on, among other things, mathematics performance and cognitive abilities in grade 6 for a 10 percent sample of cohorts born 1953, 1967, 1977 (5 percent), and 1982. The individuals are matched to their earnings records for the 1968–2015 period, making it possible to follow the earliest cohort throughout most of their labor market careers. We calculate the present value of life-cycle earnings by discounting the real annual earnings (including payroll taxes) in the period 1968–2015 at 3 percent (in 2020 prices). For ease of interpretation, we divide the life-cycle earnings by the mean

[37] We assume that the accounting cost corresponds to the value of lost production.

(separately by cohort and gender), and the estimates should be interpreted as a percentage change associated with 1 SD better mathematics skills.

Table E1. The life-cycle earnings associated with mathematics skills in grade 6

Column:	(1)	(2)	(3)	(4)	(5)	(6)
Model:	OLS	OLS	Twin FE	IV	IV	Twin FE-IV
			Panel A. All in 1953 cohort			
Mathematics test score (SD)	0.086***	0.076***		0.100***	0.089***	
	(0.004)	(0.005)		(0.007)	(0.007)	
Individual controls	No	Yes		No	Yes	
R^2	0.047	0.082		0.046	0.081	
Observations	8,090	8,090		8,090	8,090	
			Panel B. Twins in 1953–82 cohorts			
Mathematics test score (SD)	0.071***	0.079***	0.057***	0.068***	0.083**	0.099*
	(0.017)	(0.019)	(0.027)	(0.028)	(0.037)	(0.058)
Individual controls	No	Yes	Yes	No	Yes	Yes
R^2	0.057	0.1171	0.680	0.057	0.171	0.680
Observations	468	468	468	468	468	468

Note: The table shows the association between the present value of real life-cycle earnings and standardized mathematics test scores in grade 6. The present value of life-cycle earnings has been obtained by discounting real annual earnings in the period 1968–2015 (in 2020 prices) at 3 percent. The life-cycle earnings have been divided by the mean in the population (by cohort and gender), and the estimates should be interpreted as a percentage change associated with 1 SD better mathematics skills. All models control for gender and cohort. The individual controls are dummy variables for month of birth, indicators for first and second generation immigrant, dummy variables for age at immigration, dummy variables for mother's and father's highest level of education, mother's and father's percentile rank mid-age (35–45 years) earnings in levels and squared, and indicator variables for having missing information on mother's or father's earnings. Columns (4)–(6) attempt to adjust for measurement error using the individual's logical-inductive ability in grade 6 as an instrument for mathematics test scores in grade 6. */**/*** refers to statistical significance at the 10/5/1 percent level.

Table E1 shows the life-cycle earnings associated with 1 SD higher mathematics test score in grade 6. All models control for gender and cohort fixed effects. Panel A shows the estimates for individuals born in 1953, for whom we can observe earnings for ages 16–62. The first column shows that 1 SD better

mathematics performance in grade 6 is associated with about 9 percent higher life-cycle earnings. The second column accounts for differences in observed demographic characteristics and family background, which leads to slightly lower estimates. Columns 4 and 5 attempt to adjust for measurement error in the observed mathematics scores by using the individual's logical-inductive ability in grade 6 as an instrument. This increases the estimates slightly, and 1 SD better test scores is associated with about 9 percent higher discounted real life-cycle earnings.

The ETF-data includes a sample of twins which allows us to also account for unobserved family characteristics. Panel B of Table E1 shows the estimates for twins born 1953, 1967, 1972, 1977 or 1982, for whom we observe parts of their labor market career. Columns 1–2 and 4–5 replicate the models used in Panel A for the twin sample. Column 3 shows that the estimates are substantially reduced when adding twin FE to the model, which indicates that the association between test scores and earnings partly reflects difference in unobserved family background. An alternative explanation, however, is that the potential bias arising from measurement errors in observed test scores is exacerbated when exploiting the within-twin variation. This is supported by the results presented in the last column, where we use logical-inductive ability as an instrument for mathematics test scores in an attempt to adjust for attenuation bias. This leads to an association of about 10 percent but it is rather imprecisely estimated. Thus, unobserved (or observed) family background does not seem to drive much of the correlation between test scores and earnings.

Based on these estimates we assume that the return to 1 SD better mathematics performance over the life-cycle amounts to 9 percent.[38] In our data, the average real gross life-cycle earnings (including employer contributions),

[38] (Öckert 2021) reviews papers attempting to estimate causal effects of educational attainment on skills and earnings and finds that, on average, one year of schooling improves test scores by about 0.25 SD and earnings by 2.5 percent. This leads to an earnings-to-skill-effects-ratio of 10 percent.

discounted at 3 percent to age 16, for men born 1952–53 is about €900,000 (in 2020 prices). We arrive at an estimated benefit of the Boost for Mathematics by first dividing the reduced form effect for different years of exposure (first column of Table 2) by the share of treated students (first column of Table B2) and then multiplying by the estimated return to test scores (Table E1), the discounted life-cycle earnings (discounted back to the age when students are first exposed to the program) and, finally, the number of students. This yields an estimated benefit of about €1395 million, or about €2,158 per student (€1395 million/646,267 students). The benefit-to-cost ratio is about 27.23 (€1395 million/€51.2 million), meaning that the program generates €27.23 in savings for every €1 spent.

The calculations suggest that the Boost for Mathematics passes a cost-benefit test. It should be stressed, however, that the estimated societal benefits and costs are uncertain, and the effectiveness of the program may change under alternative assumptions. For instance, we base the benefit calculations only on students who have taken the final exams by year 2019, while the results show that test scores improve also for students who enter school after program implementation. Thus, if we were to extrapolate the effects of the program also for future incoming cohorts, the benefits of the Boost for Mathematics would increase even further.

On the other hand, our calculations may overstate the productivity gains of the program. Some of the estimated return to mathematics skills in Table E1 could reflect sorting of individuals in the education system – along with the corresponding return to schooling – as well as signaling on the labor market. In addition, the program could generate general equilibrium effects on the labour market, which would dampen the productivity gains. However, even if only half of the estimated return to skills is due to improved productivity, the benefit-cost-ratio would still be more than 13. Thus, also under more restrictive assumptions, the Boost for Mathematics appears to be a profitable investment.