CESifo
AREA
CONFERENCES 2022

Economics of Education

Munich, 2 – 3 September 2022

Small and Large-Group Instruction and Didactic Methods in Mathematics for Low-Performing Adolescents: Results from a Randomized Field Experiment

Lars J. Kirkebøen, Trude Gunnes, Lena Lindenskov, and Marte Rønning

CESifo

Small and large-group instruction and didactic methods in mathematics for low-performing adolescents: Results from a randomized field experiment*

Lars J. Kirkebøen[†], Trude Gunnes[‡], Lena Lindenskov[§], and Marte Rønning[¶]

Abstract: We conducted a randomized experiment combining teacher training and 8-12 weeks of targeted math instruction for low-performing 8th-graders. The low-performing students received instruction from newly trained teachers in small, homogeneous groups or larger, more heterogeneous groups. We randomized 24 schools to treatment and 24 schools to control. For low-performing students in small groups, test scores increase by 0.06 SD. We find no impact on treated students in large groups. Classroom observations indicate higher implementation fidelity of the didactic method among small-group teachers, suggesting that instruction in small, homogeneous groups combined with targeted pedagogy can reduce achievement gaps in lower secondary education.

Keywords: low-performing students, didactics of mathematics, high-dosage tutoring, ability grouping, classroom observations, randomized controlled trials, stratified randomization, cost-benefit of interventions

JEL codes: I21, I24, I28

*We thank the school authorities in Oslo (UDE) for making the experiment possible and researchers at Fafo for qualitative evaluation and administrating surveys to teachers. We further thank Gaute Eielsen and Susann Strømsvåg for excellent research assistance and Martin Eckhoff Andresen and seminar participants at the EEA virtual 2020 conference for comments. Kirkebøen is the first author due to his role as a project leader, spanning over several years, and the administrative workload related to the execution of the RCT in schools. Financing from the Norwegian Ministry of Education is much appreciated. All remaining errors are our own. The data used come from Norwegian national administrative registers and administrative registers of Oslo Municipality. Statistics Norway provides Norwegian micro data for research projects. More information is available at: https://www.ssb.no/en/omssb/tjenester-og-verktoy/data-til-forskning. Inquiries about access to data from Statistics Norway should be addressed to mikrodata@ssb.no. The authors are willing to assist with inquiries, and with contact information for Oslo Municipality. The intervention was registered in the AEA RCT Registry in July 2017 (AEARCTR-0002308).

[†]Statistics Norway, research department, Oslo, Norway. Email: kir@ssb.no

[‡]Statistics Norway, research department, Oslo, Norway.

[§]Danish School of Education, DPU, Aarhus University, Denmark.

[¶]Statistics Norway, research department, Oslo, Norway.

1 Introduction

Adolescents from low socioeconomic status (SES)-families are over-represented among those who perform poorly in school and have lower prospects for labor market careers. Increasing educational attainment among low-SES students and reducing achievement gaps among socioeconomic groups is an important policy target in many countries. Scholars point to the importance of math skills for succeeding in compulsory education and completing upper secondary education (e.g., Duncan et al., 2007). Although previous research concludes to a large degree that early investment is more beneficial than later investment (Carneiro and Heckman, 2003; Heckman, 2013), recent findings indicate high returns from programs that target adolescents (Cook et al., 2014; Clotfelter et al., 2015; Cortes et al., 2015; Fryer and Howard-Noveck, 2020; Guryan et al., 2021).[1]

We designed and implemented an intervention that targeted 8th-graders with low numeracy skills. The intervention started the school year 2016/17, lasted for three school years, with three consecutive cohorts being treated for one year each. The treatment combined extra training for qualified math teachers with targeted instruction in two periods (each lasting 4-6 weeks) for low-performing students in either small or large groups. Small-group instruction largely corresponded to what Fryer (2017) defines as high-dosage tutoring and involved substantial extra funding for the school.[2] Large-group instruction was based on the same pedagogical principles and tools. The selection of low-performing students into small and large groups followed explicit assignment rules and the targeted instruction replaced regular math classes during the intervention period. In practice, the large groups often were regular classes minus students assigned to small-groups.

The teacher training program was based on well-known didactic methods and focused on how specific didactic principles and tools can boost the achievement of low-performing students (Torgerson et al., 2012; Harder et al., 2020; Pellegrini et al., 2021). The didactic method (see Appendix A) has proven to be successful in lower grades. The idea is to apply these didactic principles and tools to improve the achievement among low-performing students in higher grades. The teacher training program lasted around 26 hours per school per year.

[1] Effective programs include accelerating algebra, charter school practices, and high-dosage tutoring.

[2] Fryer (2017) defines high-dosage tutoring as being instructed in groups of 6 or fewer for more than three days per week or being tutored at a rate that would equate to 50 hours or more over 36 weeks. While the size of our small groups aligns with Fryer (2017) in the second and third intervention years, the total extent of instruction (three hours per week for 8-12 weeks, i.e., 27-36 hours) may be somewhat less than what Fryer (2017) defines as high-dosage.

We randomly selected 24 out of 48 lower secondary schools in Oslo (the capital of Norway) to participate in the intervention, one from each of 24 matched pairs (following the recommendations of Bruhn and McKenzie, 2009). We find no evidence of significant pre-treatment differences between the treatment and control schools. Due to capacity/time constraints, full treatment was not implemeted the first year (Kirkebøen, 2017). Our main analyses therefor rely on the two last years of the intervention. [3]

We find that low-performing students predicted to receive small-group instruction by newly-trained teachers increase their average test scores by about 6 percent of a standard deviation in the year following the intervention. Our baseline estimate controls for students' earlier performance and is significant at the 5% level, both when evaluated with a wild bootstrap test and with randomization inference (based on re-sampling). Incomplete data on the actual small-group assignment suggest that 89 percent of students predicted to get instruction in small groups does get it, implying a treatment effect of 0.067 SD on those treated. The share of low-performing students is reduced by about 3 percentage points (significant at the 10% level), corresponding to a reduction of 5-25 percent for different measures of low performance. Using other studies to evaluate our results, we conclude that small-group intervention is cost-effective, with an estimated cost per small-group student of USD 1200-1800 and estimated benefits of USD 3700. We find no impact on low-performing students in large, heterogeneous groups who receive instruction from newly-trained teachers.

Supplementary data allow us to expand upon the findings from the effect analyses. Classroom observations and surveys show higher satisfaction and higher implementation fidelity of the didactic principles and tools among small-group teachers than large-group teachers, suggesting that the lack of targeted instruction is likely to explain the non-impact of the large-group treatment.

Our paper contributes to the literature on teaching interventions in schools, e.g., Guryan et al. (2021). However, despite similarities, our intervention and context differ from Guryan et al. (2021).[4]

[3]The education authorities in Oslo municipality informed quite early that only a limited amout of schools could be selected to full treatment the first school year. Hence, in 2016/17 only 8 of the 24 (treatment) schools got full treatment (both intruction in small or large groups and teacher training). Maximum group size the first year was 8 students as opposed to 6 students in the remaining two years. The remaining 16 schools only got funding and instruction how to spend the extra resources, but were in principle free to spend the money as they wanted. For more information, see Kirkebøen (2017).

[4]Guryan et al. (2021) carried out a randomized controlled trial among 9th- and 10th-graders in 12 public high schools in Chicago located in economically disadvantaged neighborhoods. Students received one-on-one/two-on-one math tutoring after school by instructors carefully selected through a screening process (pedagogical background not required). Tutoring hours could be up to 140 per year. They found that personalization of the instruction increased math test scores by 0.16 percent of a standard deviation. They did not implement any particular didactic methods.

First, we use small and large groups of students, requiring fewer teachers than more individualized one-on-one tutoring. Second, in our case, the targeted instruction replaces regular math instruction for limited periods. While possibly reducing the effect on students, it lowers the cost of the intervention. Guryan et al. (2021) rely on relatively low-cost tutors. Elsewhere, such tutors may not be available.[5] We demonstrate that we can achieve effects per dollar similar to Guryan et al. (2021) with regular teachers.

Our paper also adds to the literature on teacher fidelity in implementing new didactic principles and tools (e.g., Durlak et al., 2011). Implementing effective teaching strategies has often proved difficult (e.g., Forgasz, 2010; Rønning et al., 2013; Jacob, 2017), although our results point in the direction of high implementation fidelity of the didactic method in small, homogeneous groups but low implementation fidelity in large, heterogeneous groups.

Furthermore, this paper is tangible with the literature on teaching practices (e.g., Kane et al., 2011; Bietenbeck, 2014; Lavy, 2016; and Aucejo, 2018). These scholars find no effect of a teaching method on the class level - taking together all types of students. However, when studying heterogeneous effects by student subgroups, they find that a teaching style boosts student achievement for a certain sub-group, indicating that the same teaching practice is not optimal for every student type. Lavy (2016), for instance, finds that "instilment of knowledge and enhancement of comprehensions" have a positive effect on test scores of students from low socioeconomic backgrounds, while techniques that endow students with "analytical and critical skills" positively affect students from educated families.[6]

The paper suggests that ability grouping (i.e., the small and homogeneous groups) simplifies the implementation of the didactic method of the intervention: Teachers need to spend less time on differentiated instruction and classroom management and are left with more time to concentrate their effort on teaching low-performing target students (see Connor et al., 2013). Thus, our paper also relates to the literature on ability tracking (e.g., Duflo et al., 2011).[7]

However, half of each session focused on re-mediating skill deficits and the other half on what students were learning in their regular math classrooms.

[5] Andersen et al. (2020) find that, in Denmark, the cost of 14.5 hours of instruction by an assistant without teacher training is the same as for 10.5 hours by a trained teacher.

[6] Van Klavern (2011) fails to account for effect heterogeneity by student subgroups. Thus, he argues there is no impact of teaching practices on student achievement.

[7] The evidence on ability tracking is mixed (Cortes and Goodman, 2014). The effect depends mainly on the extent to which the teaching is tailored and matches the ability group. Duflo et al. (2011) and Guryan et al., (2021) suggest that more homogeneous groups make it easier for teachers to provide more targeted instruction (i.e., personalization), which can more than counteract the negative peer effect on low-achieving students.

Finally, we contribute to the literature on the practical design and implementation of moderate-scale randomized controlled trials (RCTs). RCTs have a large and increasing role in educational research (Fryer, 2017; Jacob, 2017; Styles and Torgerson, 2018; Andersen et al., 2020; Haaland et al., 2021; Bonesrønning et al., 2021). While the key virtue of RCTs is the expected balancing of treatment and control groups, treatment and control may not be balanced ex-post (Bruhn and McKenzie, 2009; Athey and Imbens, 2017). We investigated how our population of 48 schools might give randomly imbalanced treatment and control groups and the extent to which this could be mitigated ex-ante by stratifying on different variables. As the number of units randomized in our study is typical for the studies in Fryer (2017), our inquiry is likely to be relevant for future RCTs.

The paper is organized as follows: Section 2 presents the institutional setting. Section 3 describes the didactic method, organization of the teacher training program and targeted instruction, and implementation of the intervention. Section 4 presents the data and empirical strategy, investigates the similarity of the treatment and control schools, and analyzes alternative approaches to randomization. Section 5 presents our effect estimates, and section 6 discusses teachers' adherence to the didactic method based on classroom observations. Section 7 presents a cost-benefit analysis whereas section 8 provides a conclusion.

2 Institutional setting

Compulsory education in Norway consists of seven years of primary school and three years of lower secondary school. Children start primary school the year they turn six. Schools at the primary or secondary level are almost all public and have a local catchment area.[8] Early/late starting and grade retention are rare, such that nearly everybody starts lower secondary school the year they turn fourteen. Ability tracking is controversial, and persistent ability tracking is not permitted. There are standardized national tests in numeracy, literacy, and English in the 5th, 8th and 9th grades. In the 10th and final year, students sit exit exams.

Each municipality is in charge of its school policy. However, several explicit and implicit national standards exist, such as a national curriculum and a fixed number of teaching hours per subject. Oslo is the largest municipality and the capital of Norway. The student composition in Oslo is

[8]Parents can apply to transfer to another school. The request will be subject to available capacity at the receiving school. Less than 5 percent of students attend private schools.

heterogeneous in terms of parents' education and ethnic background. There are substantial differences in student composition across schools, reflecting residential segregation. Within municipalities, school funding is compensatory, such that schools with students of less advantageous backgrounds get more funding.

Upper secondary school is not compulsory, but students are entitled to three years of upper secondary school. Almost all students start upper secondary school directly after completing lower secondary education. However, about 25 percent do not complete it within five years. For many students, passing mathematics is a binding constraint for completing upper secondary education. Thus, better numeracy skills may enable more students to graduate. Moreover, an improved understanding of mathematics may foster motivation and create a greater sense of mastery, which low-performers may be in a deficit of (ex-ante). Low completion rates are a policy concern and the backdrop for the intervention we study.

3 The intervention

The intervention ran during the school years 2016/17, 2017/18, and 2018/19 in 24 treatment schools. The treatments was a combination of extra training for already qualified math teachers and targeted small- or large-group instruction for students in 8th grade with low proficiency in mathematics. In the remainder of the paper, students with poor numeracy are denoted 'target students'. Qualified teachers from the treatment schools attended a training program that provided them with didactic principles and tools adapted for low-performing students in mathematics. We describe this training and the underlying didactic method in subsection 3.1.

The target students in the treatment schools received two periods (5-6 weeks during October-November and 4-6 weeks around April) of targeted instruction, either in small or large groups. In the first intervention year some treatment schools did not receive teacher training, only funding for small- and large-group instruction. We describe the organization of the small- and large-group instruction in more detail and the funding-only treatment in subsection 3.2.

Before each intervention year, the school authorities in Oslo (UDE) informed the treatment schools about the intervention and what it entailed in terms of extra funding, teacher training, student selection, implementation of targeted instruction, and reporting. The 24 control schools only received

information (at the management level) about the experiment. Both the teacher training and the organization of the targeted instruction were substantially changed after the first year. We will mostly focus on the final form of the intervention, as implemented in the last two years. We describe the changes (in the teacher training program and the size of the small groups) in the two subsections below. The intervention was registered in the AEA RCT Registry in July 2017 (Kirkebøen, 2017).

3.1 The didactic method and organization of the teacher training

According to Valenta (2015), conceptual understanding, calculation, application strategies, rational thinking, and commitment are crucial for understanding numerical reasoning. Previous tests and analyses by UDE show that target students have poor comprehension of these five components, suffer from misconceptions, and derive little learning benefit from ordinary teaching. Without (basic) knowledge from primary education, the target students lack the prerequisites for mastering mathematics at the lower secondary level (Borg et al., 2014). Identified shortcomings and misconceptions have influenced the mathematics content and the didactic method used in the intervention.

UDE was responsible for the teacher training program whereas the Danish School of Education (DPU) provided professional guidance. DPU has extensive experience with research on students with low math skills, including several interventions (Jankvist and Niss, 2015; Lindenskov and Tonnesen, 2020; Harder et al., 2020). Designing the didactic content was an essential part of the project. The didactic method is evidence-based, meaning it is supported by research establishing its effectiveness. The didactic method is based on internationally acknowledged teaching practices and supplemented with experience based on other Norwegian teacher training programs. DPU and UDE embedded six principles into the teacher training program. (i) Create a link between learning sessions to activate students' memory of mathematical concepts and help them form mathematical connections. (ii) Use low threshold and high ceiling tasks to ensure that all students can get started and simultaneously make sure that the instruction is sufficiently differentiated for everybody to utilize their potential to the full. (iii) Foster motivation leading to improved performance and acknowledging that affection and cognition are aspects of learning mathematics. (iv) Initiate conversations with and among students on mathematical processes and concepts to support mathematical understanding. (v) Set realistic but high expectations to support student motivation and engagement. (vi) Create a logbook to activate students' concentration, reflections, and long-term memory. See details in Appendix A.

Teachers can endorse these six principles in the classroom by using four didactic tools: (a) the Singapore thinking blocks method, (b) persistent pairing of students (learning partner), (c) organization of instruction and learning at three levels: individual, group, plenary, and (d) linguistic expressions to enrich students' oral communication.

UDE implemented the teacher training program with assistance from DPU. The teacher training program took place before and in parallel with the instruction of target students. Treatment schools selected qualified math teachers for the intervention. Extra teachers not planning to teach as part of the intervention also attended the training, to have a pool of qualified teachers that could step in as substitutes (for instance, in case of illness), and to further embed the didactic method in the professional community. Also, head teachers and school managers attended the training.

The teacher training program separated small- and large-group teachers. The six didactic principles and four didactic tools were the same for both small- and large-group teachers. However, teachers selected to teach small groups got additional instruction materials, including concrete lesson plans and exercises. Large-group teachers did not receive any. The rationale was to let them adapt standard materials when appropriate. Large-group teachers faced mixed-ability classrooms and needed to adjust the classroom instruction to the variance among learners. They were therefore not given teaching material tailored to low-performers only.

In designing the teaching material for the small, homogeneous groups of low-performing students, DPU and UDE (re)used many elements from Numbers count.[9] This program traditionally targets students in the lowest grades, where it has proved effective (Torgerson et al., 2011), but there is less evidence on how the program affects adolescents. As poorly performing students in the 8th grade in Oslo have challenges related to curriculum goals for much lower grade levels, DPU and UDE nevertheless used Numbers count when designing the learning materials for the small groups. Numbers count can be applied in many ways if tailored to the students' age, specific conditions, and motivation structure.

A survey following the first intervention period in 2016 showed that teacher fidelity to the didactic method was low among trained teachers. This was mainly due to a shortage of information and course material (see more in Appendix B.1), which probably was related to capacity constraints of the education authorities in Oslo municipality, as already mentioned above. Based on experiences

[9]See, for instance, https://everychildcounts.edgehill.ac.uk/mathematics/numbers-count/.

from the first year, changes were made in the execution of the teacher training program (keeping the content constand). From being theoretically oriented to a large extent, the program became more practically oriented. For instance workshops where teachers could practice the didactic method and receive feedback, were included. (as discussed in Appendix B.1). There were no indications of similar problems in subsequent years, nor with the simpler first-year funding-only treatment. Thus, we will mostly analyze the first year separately. However, we will also discuss effects by year and wave. See Table A1 in the Appendix for an overview of the treatment characteristics of the different intervention years.

The teacher training program started with a meeting at the beginning of the school year explaining the background and aim of the intervention. The teachers then attended lectures and participated in workshops. More precisely, there was a total of 7 workshops (in the last two intervention years 2017/18 and 2018/19). Of these, 5 were in the autumn and 2 in the spring. There were full-day workshops in the autumn. Otherwise, a duration of 3-4 hours. In total: 26 hours per school per year. The instructors had at least a master's degree in mathematics didactic, mathematics subjects, or pedagogy and were international or national: Singapore, DPU, the Norwegian National Center for Mathematics in Education, or UDE.

3.2 Organization and funding of the small and large groups

The targeted instruction replaced regular instruction in mathematics, typically three hours per week, during the intervention period.[10] Small groups consisted of up to six students. This group size fits Fryer's (2017) definition of high-dosage tutoring (see footnote 2). The remaining target students received instruction in large groups, in practice, mainly in their regular classes (minus the small group students).

The 24 treatment schools received funding for small-group instruction. Schools that had 18 or fewer target students received funding to form up to three small groups. Schools with more than 18 target students received funding to create two small groups for the 12 lowest-performing students and a smaller amount of funding to adapt the teaching to the didactic method in large groups for the

[10]Most schools have three math sessions of 60 minutes or four math sessions of 45 minutes per week in 8th grade. There are 38 school weeks a year, so there will be 114 sessions of 60 minutes or 152 sessions of 45 minutes. The intervention thus replaced 25-30 percent of the math instruction during the 8th grade.

remaining target students.[11]

This organization meant that all treatment schools had small-group instruction, while a substantial number of students also received large-group instruction, although in fewer schools. The clear assignment rules for small and large groups mean that we can identify students in control schools that would have received small- or large-group instruction. The fact that the lowest-performing target students were taken out of regular math classes during the treatment implies that non-target students also experienced a change in class size and class composition during the treatment periods. Moreover, since large-group target students were often grouped with non-target students, non-target students may also have experienced a change in the didactic method.

The students take the 8th-grade numeracy test in late September/early October. The results were available shortly after and were used to identify target students. The selection of target students into small or large groups followed explicit assignment rules as mentioned above. The targeted instruction started in mid-October each year, and UDE followed up with the schools during the treatment years.

As already mentioned above (footnote 3) theintervention the first year differs from the intervention in the remaining two years in the sense that only 8 of 24 schools in the first year were selected to full treatment (both funding for small and large group instruction and teacher training), whereas the remaining 16 schools only got funding for group instruction. These differences across years can be exploited to shed light on important mechanisms. More precicely it

allows us to say something about the importance of group instruction versus the importance of teacher training. However, the limited number of students and schools assigned to full treatment and funding only respectively, in the first intervention year, involves that these analyses do not provide very precise results.

4 Data and empirical strategy

In this section, we describe our data, the student population, randomization and balancing across treatment and control schools as well as how we analyze the intervention effect.

[11] Small groups were funded with NOK 60 000 (USD 7 000) per group, and schools with any number of target students in large groups received an additional NOK10 000 per school. Both small and large group instruction was fully funded.

4.1 Data and target students

The data were obtained mainly from national registers or registers from Oslo Municipality. In addition, we used self-collected data from teacher surveys and classroom observations to shed light on mechanisms, notably implementation fidelity of the didactic method. From the National Education Database (NUDB), we had detailed information on students' previous results from the standardized national tests at the beginning of 8th grade (NP8). NUDB also provided information on birth year, sex, and family background, i.e., parents' highest educational attainment and immigration status. From UDE, we obtained individual-level data on students enrolled in special-needs education and the results of national tests in the 9th grade (NP9). UDE also supplied data on the assignment to groups in treatment schools. The national employer-employee register allowed us to track teachers across schools and employers.

Our complete student sample included all students in the 8th grade in Oslo in the school years 2016/17, 2017/18, and 2018/19, about 5500 students per year. We excluded students receiving special-needs education from the analysis, as they were already receiving small-group instruction and were not eligible for targeted instruction in the intervention. Furthermore, we excluded students with no data from the 8th-grade numeracy test, as we were not able to determine whether these students belonged to the target group or not. In total, our main estimation sample excluded about 10 percent of the original full sample.[12]

We defined target students as those who scored at the two lowest proficiency levels (out of five) on the standardized national test in the 8th grade, NP8. Figure A1 in Appendix C shows the distribution of test scores on NP8 for 2017 (the other years have a very similar distribution). The target group constituted about 20 percent of the students.

Table 1 presents descriptive statistics for our main estimation sample of students, where we also differentiated between target and non-target students. Overall, 49 percent were female, 36 percent had parents without higher education, and 31 percent had two foreign-born parents. As expected, among the target students, there was an over-representation of males and adolescents with lower-educated and foreign-born parents.

To facilitate interpretation of the estimated effects, test scores are normalized relative to the

[12] 4.4 percent lacked NP8 while 8.1 percent received special-needs education, with some overlap between these two groups.

national mean and standard deviation. Compared to the national average test score, students in Oslo scored about 0.37 standard deviation better on the 8th-grade numeracy tests. Target students, selected on their low 8th-grade performance, scored almost 0.8 standard deviation below the national average in the 8th grade. We estimate treatment effects on the numeracy test score in the 9th grade, which is directly comparable to the 8th-grade score. The average progress from 8th to 9th grade corresponded to about 0.32 standard deviation. However, the average improvement of the students belonging to the target group was only about 0.17 standard deviation. While 20 percent of all students in the sample performed at proficiency level one or two in the 8th grade, only 12 percent did so in the 9th grade. Of the target students, 10 percent performed at the lowest proficiency level in the 9th grade and an additional 44 percent at the second lowest. Few non-target students performed at the two lowest levels in the 9th grade.

Table 1: Descriptive statistics, main estimation sample

	Estimation sample	Target students	Non-target students
Student background			
Female	0.492	0.408	0.513
Low parental education	0.355	0.671	0.276
Foreign-born parents	0.312	0.576	0.246
Pre-determined test scores			
Grade 8 numeracy (y^8)	0.37	-0.79	0.61
Outcomes			
Grade 9 numeracy (y^9)	0.69	-0.62	0.99
Proficiency level 1, grade 9 (D^{L1})	0.020	0.103	0.001
Proficiency level 2, grade 9 (D^{L2})	0.123	0.540	0.025
Number of students	9930	1977	7953

Note: The sample consisted of students who sat the 8th-grade numeracy test in 2017 or 2018 in Oslo and who did not receive special-needs education. Test scores are normalized.

4.2 Randomization and implementation of the different treatments

We conducted the RCT at the school level.[13] Principals of all lower secondary schools in Oslo were informed about the project in February 2016. In May, the randomization took place. Shortly after,

[13]By conducting the randomization at the school level, we avoided spillover effects between treatment and control groups within the same school. This is the same motivation as, for instance, in Andersen et al. (2020).

schools knew whether they were in the treatment or control group, and the treatment schools started to make plans for teacher training and small- and large-group instruction.

Schools in Oslo are heterogeneous, with the number of target students ranging from six to 64 in 2015/16 (the year before the intervention and the most recent available test results at the time of randomization). To increase the likelihood of treatment and control schools being similar, the 48 lower secondary schools were matched on the number and shares of students in the target group in 2015/16 and divided into 24 pairs (strata). From each stratum, we randomly selected one school for treatment.[14] This method of stratifying schools before randomization follows the recommendations of Bruhn and McKenzie (2009).[15] For the first interventions in 2016/17, we randomly selected eight of the 24 treatment schools for full treatment in the following way: after sorting the strata, we pooled them into eight groups of three and selected one treatment school from each group for full treatment the first year. The remaining 16 treatment schools received the funding-only treatment in 2016/17. In 2017/18 and 2018/19, all 24 treatment schools received the full intervention, including teacher training and funding for small and large groups.

According to the assignment rules and administrative data, 560 target students in treatment schools were taught in small groups and 400 in large groups in the school years 2017/18 and 2018/19. In the first year, 2016/17, about 130 target students in the eight full-treatment schools received small-group instruction, and another 50 target students were taught in large groups. The 16 funding-only schools had 375 target students, of whom 234 were eligible for small-group instruction and 141 for instruction in large groups.

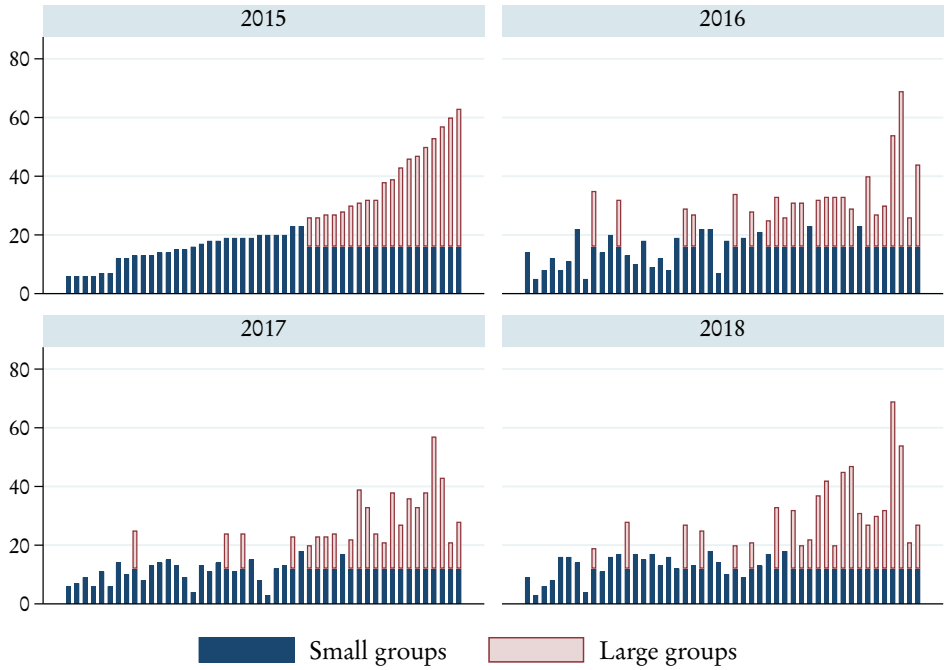
The upper left-hand panel of Figure 1 shows the number of target students who would have been assigned to small or large groups in the school year 2015/16, that is, the year used for stratifying schools for randomization. The remaining panels show how we distributed target students into small and large groups from 2016/17 to 2018/19. The number of target students varies, partly due to differences in school size (ranges from 37 to 203 8th grade students) and partly due to differences in

[14] We matched schools by constructing a distance measure based on standardized numbers and shares of target students. The number of target students has a fundamental effect on the implementation of the intervention in that it determines the number of small groups and the number of target students in large groups, while the share of low-performing students is a measure of the average performance level at the school. To ensure that there was a sufficient number of target students in large groups, in both control and treatment schools, the number of target students was given twice the weight f the share of target students when schools were matched. Randomization was carried out by writing a script that randomized schools. After testing, a random seed was set, and the program ran once.

[15] Athey and Imbens (2017) recommend having at least two treated and two control units in each stratum.

test scores (school average test scores range from 0.68 SD below the national mean to 1.03 SD above). Figure A2 in Appendix C is identical to Figure 1, except that it reports shares of target students instead of numbers.

Figure 1: Number of target students by school and year



Note: Each bar represents the number of target students in one school and year. The bars distinguish between target students predicted to get instruction in small and large groups if the school participates in the intervention. In 2015 (the school year 2015/16, and the year used as the basis for stratifying schools) and 2016 (the first intervention year), we use the 2016 maximum small-group size of eight students, while in 2017 and 2018, the reduced group size of six students. Schools are sorted by the number of target students in 2015.

For the 2017/18 students, we have data from Oslo Municipality on the actual assignment of students to small and large groups. Of 466 target students in treatment schools, 299 received small-group instruction and 154 large-group instruction. Only 13 target students were not recorded as receiving treatment. In Figure A3 in Appendix C, we compare the predicted and observed numbers. For the lowest-performing students, there is a large overlap between observed and predicted treatment. 89 percent of the lowest-performing students, who should have received small-group instruction according to the assignment rule, did so. However, about 1/3 of the target students predicted to get

large-group instruction are reported to have received small-group instruction.[16] [17]

Small groups differ from large groups in within-group student heterogeneity. The within-group standard deviation of the 8th-grade numeracy score was approximately 30 percent of the overall SD in the small groups and 70 percent in the large groups (for both predicted and reported small-group students).

4.3 Empirical strategy

As students were assigned to small and large groups based on observed test scores, we were able to identify the corresponding groups of students in control schools. Hence, we could identify the effects for the following three groups. (i) The (lowest-performing) target students in small groups, (ii) the remaining target students in large groups, and (iii) spillovers to non-target students. As specified in the pre-registration, our primary outcomes are effects on 9th grade test score for groups (i) and (ii). Secondary outcomes are test scores for group (iii) and for different subgroups. We focus on 2017/18 and 2018/19 for our main analysis of the intervention and study the treatments in the first year separately. However, we will also discuss heterogeneous effects by treatment years.

We estimated intention-to-treat effects (ITT) using the following equation:[18]

$$y_{ist} = \beta_0 + \theta T_s + \gamma_t + \delta_s + \mu X_i + \varepsilon_{ist} \tag{1}$$

y_{ist}, is the outcome of student i at school s in year t. T_s equals 1 if school s is a treated school, otherwise 0. We control for differences between cohorts (γ_t) and the 24 strata from the randomization (δ_s), as well as student characteristics X_i (gender, family background, and previous achievements such as 8th-grade test scores). As the intervention is at the school level, we need to account for clustering of the residuals (ε_{ist}). The number of schools (48) is sufficient for cluster-robust estimation to usually be considered reliable. However, with heterogeneous cluster sizes, the effective number of clusters is

[16] About half of them came from three schools which reported having 22-26 students in small groups. We do not know if these schools had more groups or larger groups than stipulated or misreported the number of students being taught in small groups.

[17] In some schools, a substantial number of non-target students were reported to have received large-group instruction. In total, 329 non-target students are reported to get large-group instruction. All these students, apart from 16, belong to seven schools indicating that all their students get small- or large-group instruction. Likely, this is due to mixing large-group target students with non-target students.

[18] The comparison of predicted and actual assignment in the previous sub-section suggests a minor attenuation bias due to mismeasurement of the small-group treatment. We will briefly comment on the treatment effect on the treated (ATT) when presenting the results.

smaller (Cameron and Miller, 2015), and in some analyses, there are fewer clusters. Therefore, we base our inference on wild bootstrap tests and comment on these tests when presenting the results. For some specifications, we also estimate regular cluster-robust standard errors and p-values based on randomization inference (resampling the assignment of the treatments) to evaluate the performance of the wild bootstrap test. Crucially, wild bootstrap performs better than cluster-robust standard errors when the clusters are few or imbalanced, while randomization inference does not depend on asymptotic properties and thus is a useful robustness check.[19]

Our parameter of interest, θ, indicates the difference between treatment and control schools and can be estimated separately for target students in small and large groups and non-target students (spillovers). We can use the same model framework for pre-determined student and school characteristics to investigate whether the treatment and control schools are similar, as expected from the randomization. If they are similar, we interpret θ as a causal effect of the intervention for post-intervention outcomes. If the treatment and control groups are not similar, we will still get an unbiased effect estimate if we manage using γ, δ, and X to control for all differences between the treatment and control groups that are not effects of the intervention. Lin (2013) justifies such OLS adjustments to experimental data if samples are sufficiently large.[20] Adjusting for individual baseline outcomes has an impact on precision and our ability to handle (random) imbalances.

4.4 Balancing of treatment and control schools

The basic idea behind stratified randomization is to ensure balance across schools belonging to the treatment and control groups. However, as we only have a limited number of schools, we may still get imbalances by chance.

Table A2 in Appendix C compares treatment and control schools. There is little evidence of systematic differences between treatment and control schools. The only significant difference (and

[19]For the wild bootstrap tests, we use the boot-test procedure (Roodman, 2015). As the wild bootstrap is sampling-based, p-values and confidence sets will vary between replications. We fixed the random seed to make the results reproducible. Regular cluster-robust standard errors are estimated using Stata's cluster option.

[20]Regression methods are not developed for analyzing data from randomized experiments. There is a disconnect between conventional assumptions in regression analyses and the implications of randomization. Athey and Imbens (2017) caution against studying RCTs with regression models and recommend using re-sampling methods. Freedman (2008a; 2008b) argues that OLS adjustment can lead to worsened asymptotic precision, invalid measures of precision, and small-sample bias. Lin (2013), however, shows why and when OLS adjustment for random differences between the baseline characteristics of the treatment and control groups is unproblematic, notably with sufficiently large samples. Following Lin (2013), we thus use control variables to balance our treatment and control groups.

Table 2: Balancing - check of randomization, all students 2017/18 and 2018/19

	(1) Dummy main sample	(2) 8th grade score (y^8)	(3) Dummy target group	(4) Small-group instruction	(5) Large-group instruction
Estimates from specification with					
No controls	0.001	0.076	-0.022	-0.011	-0.011
	[-0.030, 0.032]	[-0.056, 0.207]	[-0.058, 0.011]	[-0.034, 0.012]	[-0.041, 0.020]
Family controls		-0.005	0.004	0.006	-0.002
		[-0.077, 0.076]	[-0.017, 0.025]	[-0.017, 0.028]	[-0.030, 0.027]
N	11106	9930	9930	9930	9930
$\bar{y}$	0.894	0.363	0.199	0.115	0.084

Note: Each cell gives an estimate of θ from equation (1) for a given outcome (column) and set of controls (rows). Outcomes are (1) dummy for being in the main sample (i.e., observed 8th-grade numeracy and no special needs education), (2) 8th-grade numeracy score, (3) dummy for being in the target group (i.e., low 8th-grade numeracy score), (4) dummy for getting small-group instruction if treated and (5) dummy for getting large-group instruction if treated. The sample in column (1) consists of all students in 8th grade, whereas the sample in other columns consists of the students belonging to the main sample. The specifications in the first row only control for student cohort and strata (group in randomization), while the second row adds controls for family background. Wild bootstrap 95 percent confidence sets in brackets. Statistical significance (from wild bootstrap test): ** 5 percent level and * 10 percent level.

only at the 10 percent level) is the share of female teachers when weighted by the number of students. However, there are statistically non-significant differences in student composition. Students in treated schools are more likely to have parents with tertiary education, less likely to have immigrant parents, and have higher average 8th-grade numeracy scores.

In Table 2, we analyze pre-determined characteristics according to the design specified in section 4.3. Each cell represents a separate regression. The columns indicate the outcome variable studied, while the rows indicate which control variables we include. We start by looking at the first-row specifications, controlling only for strata in the randomization and cohort.

In the first column, we investigate whether there is a difference across treatment and control schools in the number of students from the full sample versus in the main sample. We find no such difference. In both treatment and control schools, just under 90 percent of students have test scores for 8th grade and are not receiving special needs education, and thus are included in the analyses (see outcome means in the last row). In the subsequent columns, we investigate treatment-control differences in student characteristics within the estimation sample and find no significant differences.

Column (2) shows the difference in score on the 8th-grade numeracy test, which amounts to 7.6

percent of a standard deviation. This is a sizeable difference, however, the wild bootstrap confidence sets (in brackets) span zero, and the p-value of the corresponding wild bootstrap test is 0.25. Thus, the difference is insignificant and in line with what we could expect, given the clustered design, and not an indication of failed randomization. Regular cluster-robust standard errors not using wild bootstrap provide a narrower confidence interval and indicate that the difference in 8th-grade numeracy is significant at the 10 percent level. Thus, not taking the effective number of clusters into account, e.g., through wild bootstrap, would lead us to over-reject.

As a result of the better earlier performance of the students in the treatment schools, fewer students belong to the target group in treatment schools than in control schools, but not significantly so. This difference amounts to 2.2 percentage points (column (3)) and can be compared to the sample average of 20 percent target students. Finally, columns (4) and (5) decompose the target students into those that would get small-group and those that would receive large-group instruction if treated. For both groups, the share of target students is 1.1 percentage point lower in the treatment schools, but the differences are not significant.

In the second row, we add controls for family background. Family background explains the differences in both test scores and the share of target students, as the differences in the second row are much smaller than in the first row.

In Table A3 in Appendix C, we show differences in 8th-grade numeracy scores separately for the (predicted) small-group students, large-group students, and non-target students. The treatment-control difference is insignificant for every group. The difference is largest for the small-group students, both unconditionally and conditional on family background. We address this random imbalance by controlling for pre-determined variables when analyzing the effects, as discussed in the previous sub-section, primarily 8th-grade numeracy.

With the random assignment of schools, it may seem surprising to see substantial, although insignificant, differences between treatment and control schools. However, while we expect schools to be similar on average (across many randomizations), the limited number of schools combined with the variation in school size and heterogeneity of the student population make differences such as those we observe fairly likely. This is visible from the wild bootstrap results, and also from our investigation of re-randomization in Table A4 in Appendix C, where we re-randomized 10,000 times, and in 33 percent of the replications, we get at a difference at least as large as in Table 2. This is qualitatively

different from the results using regular cluster-robust standard errors, but similar to the results using wild bootstrap, and thus gives further credibility to the wild bootstrap analyses.

We stratify schools before randomization to increase the likelihood of balanced treatment and control groups. In Figure A4 in Appendix C, we show how 8th-grade numeracy scores in the main estimation sample (consisting of 2017/18 and 2018/19 students) vary with treatment status and strata (based on the 8th-grade scores of the 2015/16 students). There is a clear tendency for average scores to be lower in strata with lower 2015/16 scores, but the relationship between strata and scores in later years is not monotonic. The within-strata score differences are also often substantial. This is not entirely unexpected. Figure 1 is sorted by the number of target students in 2015/16 such that a school retains its position along the horizontal axis in subsequent years. We see that the number of target students (and the share of target students in Figure A2) does not increase monotonically with rank in later years; while the number of target students correlates over the years, the ranking of schools changes.

Given the imperfect sorting of schools into strata, it is reasonable to ask if we could have done better with respect to stratifying and randomization of schools. In Table A4 in Appendix C, we compare the performance of alternative stratification methods. We discuss this in more detail in our working paper (Kirkebøen et al., 2021). In short, our findings highlight the tension between a desire to balance several characteristics and a desire to better balance one (see Bruhn and McKenzie, 2009). In our case, we seek to balance baseline outcomes, but also school size, because we want sufficiently many large-group students in a sufficient number of treatment and control schools. Our method of stratification balances these two, but a small increase in the expected balance of school size comes at the cost of a substantially reduced expected balance in baseline outcomes.

5 Results

In this section, we present our effect estimates of the different treatments. We first investigate the effects on target students of receiving small- and large-group instruction of newly-trained teachers in the two last intervention years. We then study the funding-only treatment, and the two waves of the full treatment, separately for the first year and pooled for all three intervention years. Finally, we study spillovers to non-target students.

5.1 Effects on the target students

Small-group instruction and teacher training

In Table 3, we report the estimated effects of being taught by newly-trained teachers in small groups. Each cell represents a separate regression. We study different outcome variables (columns) and include various control variables (rows). As shown in section 4.4, we find no evidence of significant random pre-intervention differences between treatment and control schools. However, although insignificant, the pre-existing differences are sufficiently large that they may impact post-intervention differences, and we will take them into account when estimating treatment effects, as justified by Lin (2013).

We start by establishing (in column (1)) that the difference in test-taking across treatment and control schools is essentially zero, irrespective of controls. This is reassuring, as marginal test-takers will typically be low-performing students. If the intervention affected test-taking, this could mask or exacerbate an effect on test scores.

Table 3: Treatment effects, target students in small groups 2017/18 and 2018/19

	(1) Dummy has y^9	(2) 9th grade score (y^9)	(3) Lowest proficiency (D^{L1})	(4) Low proficiency (D^{L2})
Effect estimates from specification with				
No controls	0.001	0.122**	-0.052*	-0.069*
	[-0.050, 0.050]	[0.016, 0.226]	[-0.104, 0.005]	[-0.144, 0.006]
Family controls	0.001	0.104**	-0.048*	-0.061**
	[-0.043, 0.044]	[0.008, 0.193]	[-0.098, 0.006]	[-0.124, -0.001]
Family $+y^8$ controls	-0.003	0.060**	-0.035*	-0.028
	[-0.047, 0.040]	[0.004, 0.118]	[-0.073, 0.003]	[-0.071, 0.015]
N	1142	1015	1015	1015
N clusters	48	48	48	48
$\bar{y}$	0.889	-0.720	0.141	0.603

Note: Each cell contains an estimate of θ from equation (1) for a given outcome (column) and set of controls (rows). Outcomes are (1) dummy for whether the student has a 9th-grade numeracy score, (2) (normalized) 9th-grade numeracy score, (3) dummy for 9th-grade numeracy score at lowest proficiency level and (4) dummy for 9th-grade numeracy score at either of the two lowest proficiency levels. The specifications in the first row control for student cohort and strata (group in randomization), the second row adds controls for family background, while the third row include 8th-grade numeracy test score (third-degree polynomials). The sample consists of target students *predicted* to get instruction in small groups in years 2017/2018 and 2018/2019 and corresponding students in control schools. Column (1) also includes students without a 9th-grade test score. Wild bootstrap 95 percent confidence sets in brackets. Statistical significance (from wild bootstrap test): ** 5 percent level and * 10 percent level.

In column (2), we present the effects on our main outcome variable, the 9th-grade test score. Low-performing students receiving small-group instruction perform 0.12 SD better than the corresponding students in the control schools (the top row, without controls). This effect is insufficient to overcome these students' disadvantage relative to the average student (corresponding to 0.7 SD, cf. the sample mean in the bottom row), but still a substantial improvement. A part of the difference between target students in treatment and control schools is attributable to their more advantageous family background. When family controls are added in the second row, the point estimate decreases to 0.10. In the third row, we add controls for prior achievement (8th-grade test scores) and obtain a difference of 0.06 SD in favor of the treated students. As this estimate is conditional on prior performance, and there is no impact on test-taking, this is our preferred specification, and we argue, a credible estimate of the intention-to-treat (ITT) effect for a target student predicted to receive small-group instruction.

Based on our results from Section 4.4, indicating a limited number of effective clusters, we base our inference on the wild bootstrap confidence sets shown in brackets in Table 3 and the associated p-values. As neither of the confidence sets in column (2) span zero, all three estimates, with different sets of controls, are significant at the 5 percent level. The p-value of the estimate with only family controls is 0.026 and for the estimate with control for 8th-grade numeracy is 0.042. As in Table 2 the randomization-based p-values are similar, at 0.019 and 0.039, thus providing further support to the validity of our inference.[21]

Assuming that the 2017/18 share of 89 percent of predicted small-group students who receive such instruction is representative for both years and that there is no effect on the remaining 11 percent who do not receive small-group instruction, the ITT effect estimate corresponds to an average treatment effect on the treated (ATT) of about 0.067 SD.

In columns (3) and (4), we study differences in the share of students performing at the lowest or either of the two lowest proficiency levels on the 9th-grade test. In line with the positive effect on test scores, we find a reduction of 3-4 percentage points in either measure of low-scoring students, i.e., reductions of about 25 and 5 percent relative to the base rates of 14 percent performing on the lowest level and 60 percent on either of the two lowest levels (cf. the bottom row of Table 3). However, these

[21] Randomization-based p-values are based on re-randomizing 10,000 times. Regular cluster-robust standard errors imply p-values <0.01 for both coefficients, thus over-rejecting, as in Section 4.4.

reductions are for the most part only significant at the 10 percent level, and the preferred specification yields an insignificant effect for the share at either of the two lowest proficiency levels.

We have investigated heterogeneous effects by a range of student and school characteristics.[22] We find significant effects (at the 5 or 10 percent level) for boys, children of parents with higher education and of Norwegian-born parents, and not for girls, children with lower-educated parents, and immigrant children. However, we cannot reject the possibility that the effects are the same. There are no clear differences by 8th-grade test score or cohort. We find a effects in schools with higher mean 8th-grade test scores and no impact in schools with lower mean test scores. This is the only case where the effects for subgroups of schools are significantly different. However, it does not point clearly to any mechanism, as schools with higher average test scores also have fewer target students and a higher share of target students receiving small-group instruction.

Large-group instruction and teacher training

Table 4 presents the effects on target students of being taught by trained teachers in large groups. The set-up is identical to Table 3. As in the case of small-group instruction, there are no major differences in test-taking across treatment and control schools (column (1)). Turning to our main outcome variable, 9th-grade test scores in column (2), the point estimate is negative and insignificant in all specifications. It is also close to zero, particularly in the preferred specification, where we control for the 8th-grade test score. Consistent with the result of no impact on test scores, we find no effects on the shares of low-performing students in columns (3) and (4).

[22] Results omitted for brevity.

Table 4: Treatment effects, target students in large groups 2017/18 and 2018/19

	(1) Dummy has y^9	(2) 9th grade score (y^9)	(3) Lowest proficiency (D^{L1})	(4) Low proficiency (D^{L2})
Effect estimates from specification with				
No controls	0.010	-0.036	0.012	0.022
	[-0.019, 0.053]	[-0.111, 0.068]	[-0.011, 0.033]	[-0.082, 0.111]
Family controls	0.014	-0.041	0.015	0.029
	[-0.015, 0.056]	[-0.114, 0.066]	[-0.014, 0.040]	[-0.077, 0.118]
Family $+y^8$ controls	0.016	-0.010	0.006	0.005
	[-0.017, 0.059]	[-0.083, 0.077]	[-0.023, 0.036]	[-0.100, 0.089]
N	835	760	760	760
N clusters	25	25	25	25
$\bar{y}$	0.910	-0.483	0.053	0.455

Note: Each cell gives an estimate of θ from equation (1) for a given outcome (column) and set of controls (rows). See note to Table 3 for details. The sample is target students predicted to get in instruction in large groups in years 2017/2018 and 2018/2019 and corresponding students in control schools. Wild bootstrap 95 percent confidence sets in brackets. Statistical significance (from wild bootstrap test): ** 5 percent level and * 10 percent level.

There are fewer students in the large-group sample than in the small-group sample. Furthermore, we only have 25 schools (11 treatment and 14 control), which reduces the power of the large-group analysis. A wild bootstrap test produces a confidence set of (-0.08, 0.08) in our preferred specification. Thus, a positive effect larger than 0.08 SD is highly unlikely, and the small point estimates (particularly when controlling for y^8) do not point to substantial effects that we are unable to detect due to low precision. While the confidence sets for the impacts on test scores in Tables 3 and 4 overlap, a wild bootstrap test rejects equality at the 10 percent level ($p = 0.065$).[23]

Investigating heterogeneous effects, no coefficient is individually significant, nor is any effect difference between groups of students. This is not surprising, given that the number of students randomized to large groups is smaller than the number randomized to small groups, the large-group students are distributed across fewer schools, and we find no indication of an average effect in Table 4.

[23]To compare the effects, we estimate the model in eq. (1), fully interacted with a dummy for belonging to the small-group sample, on data for all target students. With robust standard errors, this is equivalent to separate regressions.

Funding-only in the first year and full treatment in all years

Table A5 in Appendix C shows results for the first year of the intervention for the full treatment and funding-only, for target students in small and large groups and for non-target students.

Regarding full treatment schools, adjusting for the difference in 8th-grade score yields substantial but imprecise negative effect estimates, in particular for students in large groups. However, the treated students belong to only eight schools, and the wild bootstrap confidence sets are wide and provide insignificant estimates for both small and large group students.

Regarding schools only receiving funding for group instruction, we find a negative effect of 0.07 SD for small-group students and a negative effect of 0.08 SD for large-group students. However, neither of the effects are significant. While the results for the funding-only treatment may suggest that small-group instruction without a customized didactic method is not sufficient to improve the achievement of low-performing students, these results need to be interpreted with caution, as group size and student heterogeneity are not directly comparable across the first year and the subsequent two years (see Table A1 for an overview).

All treatment schools implemented the same treatments in the last two years. However, while teachers in the schools in wave 1 had previously received training, the teachers in wave 2 had not. We have investigated effects by wave, year, and treatment for all three intervention years. The estimated effect for each wave, year, and treatment is imprecisely estimated, and not individually significant. However, the estimated effects of the full treatment in small groups in schools in waves 1 and 2 are very similar in the second and third years. Thus, the treatments in the first year do not seem to have a lasting effect at the school level.[24]

5.2 Spillovers to non-target students

Table A6 in Appendix C reports results for non-target students. Regarding the main outcome variable, 9th-grade test scores, all estimates are close to zero, regardless of which controls are included. The wild bootstrap confidence sets rule out effects like the main effect in Table 3, and a wild bootstrap test rejects equality of effects at the 10 percent level ($p = 0.050$). There are no significant effects on

[24]Results are omitted for brevity. One reason may be that teachers receiving training in the first year were not teaching 8th-graders the following year. E.g., they continued teaching the students in 9th grade. Anecdotal evidence suggests that this was the case. Another possibility is that the first-year training had no lasting impact on the teaching. Recall, however, that there were implementation issues in the full treatment the first year and only eight schools.

the shares of non-target students who perform at the two lowest proficiency levels on the 9th-grade test.[25]

Overall, there is little indication of effects on non-target and target students in large groups. Recall, non-target students were often grouped with target students in large groups, i.e., large groups corresponded to regular classes with small-group students removed. Hence, class size and composition changed for all 8th graders in treated schools. However,, these changes are smaller for students in large groups than small groups. Previous studies have found no class-size effects in lower secondary education in Norway (Leuven et al., 2008; Leuven and Løkken, 2020). Moreover, even if large-group teachers participated in the teacher training program, the variation in the academic level of the adolescents in the large groups (spanning from proficiency levels 2-5, see Figure A1 in Appendix C) may have been too large for teachers to apply the didactic method of the intervention (Duflo et al., 2011). That is, it is hard for the teachers to respond to student variance and differentiate the instruction across student types. We discuss teacher fidelity to the didactic principles and tools in the next section.

6 Fidelity to the didactic method

For students to receive targeted instruction, teachers must apply the didactic method they learned during their training. Implementation fidelity of the didactic method implies adherence to the didactic principles and tools of the program.

During autumn 2017 and spring 2018, DPU collected data on fidelity through non-participative observation in randomly selected (treated) classrooms. DPU developed and used observation forms to assess the implementation fidelity of the six didactic principles and the four didactic tools.[26] DPU observed 47 intervention sessions, 35 in small groups and 12 in large groups. Each classroom

[25]Investigating heterogeneous effects for the non-target student, we find a negative effect for Norwegian-born children of immigrants, which is only significant at the 10 percent level. The estimated effect is significantly different from the estimated effects for children with Norwegian-born parents at the 10 percent level, but not from the effect for children of immigrant parents. However, as we test a large number of secondary hypotheses, this single significant coefficient may be spurious.

[26]These forms are available upon request. A brief example: The aim of principle 1 is to activate the memory of mathematical concepts and help students form mathematical connections. Adherence to this principle was coded with five observation nodes to qualify fidelity: (i) No link, (ii) teacher provides organizational link, (iii) students state organizational link, (iv) teacher states mathematical link, and (v) students state mathematical link. We find that about half of the small group sessions and less than 20 percent of the large-group sessions had teachers or students asserting mathematical links. A third of all large-group sessions did not state any links to either earlier or subsequent sessions. By contrast, very few small group sessions did not state any link.

observation followed one intervention session from start to end.

Overall, we find higher fidelity to the didactic principles and tools among small-group teachers than among large-group teachers. It is not surprising that small-group teachers show greater adherence to the didactic method of the intervention. The small-group students are more homogeneous, enabling teachers to target one academic level and concentrate their instructional effort on the low-performing target students (Duflo et al., 2011; Guryan et al., 2021). Teachers of small groups were also given additional teaching materials and more detailed instruction plans that incorporated the didactic principles and tools suited for low-performing students, making the implementation easier. Research shows that experiments are more effective when they are uncomplicated to follow. Together, ability grouping and teaching materials may have made the implementation easier for small-group teachers, thereby increasing their fidelity to the didactic method. Responses from teacher surveys (Kirkebøen et al., 2018) also indicate more enthusiasm and satisfaction among small-group teachers.[27]

We find that principle 2 (use low threshold-high ceiling tasks), alias differentiated instruction, which is suitable for the large and relatively heterogeneous groups of students, is little used. Teachers only applied this principle in one-third of the observed large-group sessions, suggesting that it is difficult and cumbersome to conduct differentiated instruction in the large and mixed-ability groups. The lack of differentiated instruction, and the low implementation fidelity of the other didactic principles and tools, likely explain the absence of impact on target (and non-target) students in large groups, as indicated in Table 4. Also, other scholars (e.g., Durlak et al., 2011) find that low program implementation of an evidence-based intervention results in poorer student outcomes than complete implementation.

The didactic principles and tools are not unique to the intervention. Comparing survey responses from schools, we find that the didactic method is somewhat familiar to teachers in control schools (Kirkebøen et al., 2018; see also Appendix B.2). However, when asked about the deployment of the different didactic methods, a clear difference emerged between treatment and control schools, with principles and tools adhered to more in treatment schools.

We restricted access to the teacher training program and the didactic material specific to the intervention to treatment schools. Thus, control school teachers were not directly exposed, provided they did not change school along the way. From matched employer-employee data, we identified

[27]For a thorough analysis of the classroom observations and teacher fidelity, see Lindenskov and Gunnes (2021).

2211 teachers working in control or treatment schools in October 2017.[28] In March 2019, only 18 out of 1115 teachers previously working in treatment schools were found to have moved to control schools. We cannot identify teachers who received training in the linked data, but the low number suggests that direct contamination through job changes is not a big issue. Note there were few job changes between sample schools, but not low mobility generally. Nineteen percent of the October 2017 teachers were not working at the same school in March 2019.[29]

7 Costs and benefits

In section 5, we found a significant intention-to-treat effect of the small-group intervention of 0.06 SD, corresponding to an average treatment effect on those treated of about 0.067 SD. In this section, we discuss how we can place a monetary value on this effect and how the value compares to the cost of the intervention. As we found no effect resulting from large-group instruction, we only focus on the small groups. That is, we focus on the costs-benefits as if the small-group instruction were to be implemented alone and not the total cost of our experiment with several treatments.

The cost of the small-group instruction was about USD 1200 per small-group student, while the cost of the entire intervention was about USD 1800 per small-group student, including the teacher training program. Implementing the small-group intervention will require funding for small-group instruction, teacher training, and administrative overheads.[30] We conclude that USD 1200 is a lower bound for the per-student cost, while USD 1800 is a reasonable upper bound. Thus, for the small-group treatment, we find an ITT effect of 0.033-0.050 SD per USD 1000 and an ATT effect of 0.037-0.056 SD per USD 1000. These effects are slightly lower than in Guryan et al. (2021).

To put a monetary value on the effect, we used Kirkebøen (2021), who studied the impact of school quality on long-term student outcomes. A 0.06 SD effect on numeracy early in lower-secondary education can be expected to increase end-of-compulsory-school grades by 0.04 SD, upper secondary

[28] We define teachers as employees with non-zero working hours and a teacher or principal occupation code.

[29] Both from treatment and control schools, about 3 percent move to other sample schools, 6-7 percent to other schools (schools outside Oslo or not lower secondary schools), and 10 percent do not work at schools anymore. The teachers identified from the data include 192 principals and other managers. One manager moves from a treatment to a control school and one from control to treatment. In total, 36 of 192 managers changed their workplace.

[30] The total costs of the intervention (not including the research) were around USD 1.7m. The intervention included extra administrative resources to enable UDE to communicate with and provide data for the researchers. For the small-group intervention only, the total administrative cost is lower. The per-student administrative cost will depend on the scale of the intervention, although some of the costs are one-off costs, e.g., the costs of teaching materials.

school completion rates by 0.6 percentage point, and earnings by 0.5 percent, or USD 265 per year. This is similar to or slightly lower than the valuation of test score effects in Guryan et al. (2021), based on Chetty et al. (2011).[31] With about 600 students receiving effective small-group instruction during the main intervention years, this corresponds to 3-4 more low-performing students completing upper secondary school. Although this is a small number, and any pay-off in the labor market will be several years into the future, the intervention may well be cost-effective. Using a discount rate of 4 percent (as the Norwegian Ministry of Finance recommends for public investments), the present value of the above earnings effect from ages 23-59 is about USD 3700, twice the total cost per small-group student and three times the cost of small-group instruction.[32] Thus, if there are sustained effects on employment and earnings similar to what we can expect from the shorter-term impact, the small-group instruction will be highly cost-effective.

8 Conclusion

Is it too late to implement measures in lower-secondary education for students falling behind in mathematics? This paper evaluates an intervention aiming to improve the achievement of low-performing 8th graders in mathematics. While it is too early to conclude about longer-term effects (e.g., completion of upper secondary school), the short-term impact on student achievement is promising.

A majority of the target students in our sample were randomized to small groups, where they received customized instruction by newly-trained teachers. The findings indicate that the intervention increased these students' test scores by about 6 percent of a standard deviation in the year following the intervention. The share of low-performing students was reduced by about 3 percentage points, corresponding to a 5-25 percent reduction for different low-performance measures. The intervention is cost-effective, with an estimated cost per small-group student of USD 1200-1800 and estimated

[31] Kirkebøen (2021) finds that 0.1 SD higher school value-added on 8th-grade test scores increases 10th-grade exam scores by 0.067 SD and the share completing upper secondary school by about 1 percentage point. 0.1 SD difference in exam value-added is associated with 1.7 percentage points higher completion rates and 1.5 percent higher earnings around age 30. Chetty et al. (2011) find that a one-percentile point increase in the 8th-grade test score is associated with USD 150 higher earnings at age 27. Guryan et al. (2021) find that percentile rank increases by one for about every 0.026 change in test scores, such that an effect of 0.06 SD corresponds to 2.25 percentile points or USD 340.

[32] Falch et al. (2009) estimate the social return to completing upper secondary school at USD 151k (adjusted for earnings growth since 2009), meaning that an effect on completion of 0.6 percentage points is valued at USD 900 per student. However, this disregards any impact of increased mathematics skills not operating through upper secondary school completion. Some international estimates of the value of completion are much higher. For the US, Levin et al. (2012) estimate the private return to high school to be USD 258k and the social return to be USD 756k.

benefits of USD 3700.

Some target students were assigned to large groups, mainly in their regular classes, minus the small group students. Although instructed by newly-trained teachers, we find no impact on these students. Classroom observations and teacher surveys suggest low implementation fidelity of the didactic method in large groups. There are several possible reasons for this. First, large groups require more time and effort spent on classroom management: students were heterogeneous and, therefore, more challenging to manage due to differentiated instruction being necessary. Second, unlike small-group teachers, teachers in large groups did not receive any concrete lesson plans or teaching material as they needed to adapt the instruction to several student types, not only low-performing students. The rationale was to let them adapt standard materials when appropriate. However, adaptation may have made compliance with the didactic method harder for large-group teachers. In the same vein, there is no indication from the first year that providing instruction without teacher training in small or large groups influences achievement.

Overall, our results suggest that neither small-group instruction nor targeted teacher training alone is enough to increase performance of low-performing adolescents. However, ability grouping combined with teacher training in didactic methods facilitates better-tailored instruction and improve low-performing students' math skills in lower secondary education.

References

[1] Andersen, S.C., Beuchert, L., Nielsen, H.S., and Thomsen, M.K. (2020). The Effect of Teacher's Aides in the Classroom: Evidence from a Randomized Trial. Journal of the European Economic Association 18, 1, 469-505.

[2] Athey, S. and Imbens, G.W. (2017). The econometrics of randomized experiments. Handbook of Economic Field Experiments. Vol. 1. North-Holland, 73-140.

[3] Aucejo, E.M., Coat, P., Fruehwirth, J.C., Kelly, S., and Mozenter, Z. (2018). Teacher effectiveness and classroom composition. Working paper.

[4] Bettinger, E., Lundvigsen, S., Rege, M., Solli, I.F., and Yeager, D. (2018). Increasing perseverance in math: Evidence from a field experiment in Norway. Journal of Economic Behavior and Organization 146, 1-15.

[5] Bietenbeck, J. (2014). Teaching practices and cognitive skills. Labour Economics 30, 143-153.

[6] Bligh, D. A. (2000). What's the use of lectures? San Fransisco: Jossey-Bass.

[7] Boaler, J. (2011). Changing students' lives through the de-tracking of urban mathematics classrooms. Journal of Urban Mathematics Education 4, 1, 7-15.

[8] Bonesrønning, H., Finseraas, H., Hardoy, I., Vaag Iversen, J.M., Nyhus, O.H., Opheim, V., Vea Salvanes, K., Jorde Sandsør, A.M. and Schøne, P. (2021). Small Group Instruction to Improve Student Performance in Mathematics in Early Grades: Results from a Randomized Field Experiment. CESifo Working Paper no. 9443.

[9] Borg, E. (2014). Et lag rundt læreren-kunnskapsoversikt. Rapport. Oslo: AFI.

[10] Bruhn, M. and McKenzie, D. (2009). In Pursuit of Balance. Randomization in Practice in Development Field Experiments. American Economic Journal: Applied Economics 1 (4), 200-232.

[11] Cameron, A.C and Miller. D. L. (2015). A Practitioner's Guide to Cluster-Robust Inference. Journal of Human Resources 50, 2, 317-372.

[12] Carneiro, P. and Heckman, J.J. (2003). Human Capital Policy. In James J. Heckman and Alan B. Krueger (eds.). Inequality in America: What role for human capital policy? Cambridge, MA. MIT Press. 77-240.

[13] Center for Excellence in Teaching and Learning (2021). The one-minute-paper. Available at Center for Excellence in Teaching and Learning: University of Rochester.

[14] Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., and Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. Quarterly Journal of Economics 126, 4, 1593-1660.

[15] Clotfelter, C. T., Ladd, H.F., and Vigdor, J. L. (2015). The aftermath of accelerating algebra: Evidence from district policy initiatives. Journal of Human Resources 50, 159-188.

[16] Cook, P.J., et al. (2014). The (surprising) efficacy of academic and behavioral intervention with disadvantaged youth: Results from a randomized experiment in Chicago. NBER Working Paper Nr. 19862.

[17] Connor, C. M., Morrison, F. J., Fishman, B., Crowe, E. C., Al Otaiba, S., and Schatschneider, C. (2013). A longitudinal cluster-randomized controlled study on the accumulating effects of individualized literacy instruction on students' reading from first through third grade. Psychological Science 24, 8, 1408-1419.

[18] Cortes, K.E., and Goodman, J.S. (2014). Ability-tracking, instructional time, and better pedagogy: The effect of double-dose algebra on student achievement. American Economic Review: Papers and Proceedings 104, 400-405.

[19] Cortes, K., Goodman, J.S., and Nomi, T. (2015). Intensive math instruction and educational attainment: Long-run impacts of double-dose algebra. Journal of Human Resources 50, 1, 108-158.

[20] Duflo, E., Dupas. P., and Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. American Economic Review 101, 5, 1739-74.

[21] Duncan et al. (2007). School readiness and later achievement. Developmental Psychology 43, 6, 1428-1446.

[22] Durlak, J.A., Weissberg, R.P., Dymnicki, A., Taylor, R.D., and Schellinger, K. (2011). The Impact of Enhancing Students' Social and Emotional Learning: A Meta-Analysis of School-Based Universal Interventions. Child Development 82, 1, 405-32.

[23] Esmonde, I. (2012). Mathematics learning in groups: Analysing equity within an activity structure. In: Herbel-Eisenmann B., Choppin J., Wagner D., and Pimm D. (eds). Springer.

[24] Falch, T., Johannessen, A. B., and Strøm, B. (2009). Kostnader av frafall i videregående opplæring. SØF-rapport 08.

[25] Faragher, R., Brady, J., Clarke, B., and Gervasoni, A. (2008). Children with Down syndrome learning mathematics: can they do it? Yes, they can! Australian Primary Mathematics Classroom 13, 4, 10-15.

[26] Forgazs, H. (2010). Streaming for mathematics in years 7-10 in Victoria: An issue of equity? Mathematics Education Research Journal 22, 1, 57-90.

[27] Freedman, D. A. (2008a). On regression adjustments to experimental data. Adv. in Appl. Math. 40 180-193.

[28] Freedman, D. A. (2008b). On regression adjustments in experiments with several treatments. Ann. Appl. Stat. 2 176–196.

[29] Fryer, R.G. (2017). The production of human capital in developed countries: Evidence from 196 randomized field experiments. Handbook of Field Experiments. Vol. 2. North-Holland. 95-322.

[30] Fryer, R.G and Howard-Noveck, M. (2020). High-dosage tutoring and reading achievement: Evidence from New York City. Journal of Labor Economics 38, 421-452.

[31] Guryan et al. (2021). Not too late: Improving academic outcomes among adolescents. NBER working paper series.

[32] Haaland, V.F., Rege, M., and Solheim, O. (2021). Complementarity in the Education Production Function: Teacher-Student Ratio and Teacher Professional Development, Working paper.

[33] Harder, J., Færch, J. V., Malm, S. G., Overgaard, S., Rasmussen, K. et al. (2020). Sammenfatning af følgeforskningen på matematikindsats 2017-TMTM, Tidlig matematikindsats til marginalgruppeelever. Trygfondens Børneforskningscenter, Aarhus Universitet, Københavns Professionshøjskole.

[34] Heckman, J.J. (2013). Giving kids a fair chance. (A strategy that works). The MIT Press.

[35] Jacob, B. (2017). When evidence is not enough. Findings from a randomized evaluation of Evidence-Based Literacy Instruction (EBLI). Labour Economics 45, 5-16.

[36] Jankvist, U.T. and Niss, M. (2015). A framework for designing a research-based "math counselor" teacher program. Educational Studies in Mathematics 90, 259-284.

[37] Kane, T.J., Taylor, E., Tyler, J., and Wooten, A. (2011). Identifying effective classroom practices using student achievement data. Journal of Human Resources 46, 3, 587-613.

[38] Kirkebøen, L. (2017). Targeted remedial mathematics teaching to improve upper secondary completion rates. AEA RCT Registry. https://doi.org/10.1257/rct.2308-1.0.

[39] Kirkebøen, L., Gunnes, T., Lindenskov, L., and Rønning, M. (2021). Didactic methods and small-group instruction for low-performing adolescents in mathematics: Results from a randomized controlled trial. Discussion papers 957, Statistics Norway.

[40] Kirkebøen, L. (2021). School value-added and long-term student outcomes. Discussion papers 970, Statistics Norway.

[41] Kirkebøen, L.J., Eilertsen, G., Rønning, M., Strømsvåg, S., Andresen, S., Reegård, K., Rogstad, J., Berge, J.E., and Lindenskov, L. (2018). Matematikkdidaktisk etterutdanning av lærere og målrettet strukturert matematikkundervisning ved overgang til 8. trinn og VG1. Foreløpig beskrivelse av utforming og gjennomføring av tiltak. Notater, 15. Statistics Norway.

[42] Lavy, V. (2016). What makes an efficient teacher? Quasi-experimental evidence. CESifo Economic Studies 1, 88-125.

[43] Leder, G. C., Pehkonen, E., and Törner, G. (Eds.). (2002). Beliefs: A hidden variable in mathematics education? Dordrecht: Kluwer Academic Publishers.

[44] Leuven, E. and Løkken, S. (2020). Long-term impacts of class size in compulsory school. Journal of Human Resources 55, 1, 309-348.

[45] Leuven, E., Oosterbeek, H., and Rønning, M. (2008). Quasi-experimental estimates of the effect of class size on achievement in Norway. Scandinavian Journal of Economics 110, 4, 663-693.

[46] Levin, H., et al. (2012). Cost-effectiveness analysis of interventions that improve high school completion. Center for benefit-cost studies of education. Teachers College, Columbia University.

[47] Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. The Annals of Applied Statistics 7, 1, 295-318.

[48] Lindenskov, L. and Tonnesen. P. B. (2020). A logical model for interventions for students in mathematics difficulties-improving professionalism and mathematical confidence. Nordic Studies in Mathematics Education 25 (3-4), 7-26.

[49] Lindenskov, L. and Gunnes, T. (2021). Didactic methods and teacher fidelity in large and small groups of students. Unpublished manuscript.

[50] Okazaki, M., Okamoto, K., and Morozumi, T. (2019). Characterizing the quality of mathematics lessons in japan from the narrative structure of the classroom. Hiroshima Journal of Mathematics Education 12, 49-70.

[51] Pellegrini, M., Lake, C., Neitzel, A., and Slavin, R. E. (2021). Effective programs in elementary mathematics: A meta-analysis. AERA Open 7,1, 1-29.

[52] Roodman, D. (2015). Boot-test: Stata module to provide fast execution of the wild bootstrap with null imposed, Statistical Software Components S458121, Boston College Department of Economics.

[53] Rønning, W., Hodgson, J., and Tomlinson, P. (2013). Å se og bli sett. Klasseromsobservasjoner av intensivopplæringen i Ny Giv. NF-rapport, 6.

[54] Scherer, P., Beswick, K., DeBlois, L., Healy, L., and Moser Opitz, E. (2017). Assistance of students with mathematical learning difficulties: how can research support practice? In G. Kaiser (Ed.). 249-259. Springer.

[55] Styles, B. and Torgerson, C. (2018). Randomised controlled trials (RCTs) in education research-methodological debates, questions, challenges. Educational Research 60, 255-264.

[56] Torgerson, C., Wiggins, A., Torgerson, D., Ainsworth, H., Hewitt, C. (2012). The effectiveness of an intensive individual tutoring programme (Numbers Count) delivered individually or to small groups of children: A randomised controlled trial. Effective Education 4, 1, 73-86.

[57] Valenta, A. (2015). Aspekter ved tallforståelse. Nasjonalt senter for matematikk i opplæringen.

[58] Van Klaveren, C. (2011). Lecturing style teaching and student performance. Economics of Education Review 30, 729-739.

APPENDIX

A The six chosen didactic principles and their scientific background

Principle 1: Create a link between learning sessions

The aim of principle 1 is to encourage teachers to help students experience connections between sessions to support their memory consolidation and tuning-in. As a consequence of the first international comparison of student achievement in mathematics (TIMSS 1995), Western mathematics educators studied high-achieving countries, such as Japan. Teachers in Japanese classrooms clarify to students how the content and working methods of one session relate to previous and future sessions (Okazaki et al., 2019). The hypothesis is that coherent elements can help the students form mathematical connections by supporting concentration, memory consolidation, perception of meaningfulness, and a deeper understanding in students.

Principle 2: Use low threshold-high ceiling tasks

Boaler (2011) finds that students benefit from non-tracking. The rationale is that teachers - independently of whether students perform at a low, medium, or high level - endorse teaching practices consisting of rich learning tasks and high expectations of students. Forgasz (2010), on the other hand, finds that tracking has benefits for high-performing students but disadvantages for low-performing students. While teachers allow high-achievers to engage in mathematical challenges, they offer simple math for low-performing students, restricting their learning opportunities. Promoting activities where all students have the prerequisites for getting started (low entry threshold) and continuing activities in more complex variants as far as the situation allows (high ceiling) is an essential principle of differentiated instruction. For small group teachers, Principle 2 is embedded in the teaching material. In large groups, the teachers must create new or adapt existing tasks into low-threshold-high-ceiling variants to enable all students to engage in real mathematical challenges from a safe starting point to as much as they can master. The teacher training program in years 2 and 3 presented ideas and examples of how to develop existing tasks.

Principle 3: Create motivation that leads to improved performance

At the beginning of this century, affective aspects were hidden variables in mathematics education research (Leder et al., 2002). Nowadays, affective aspects are increasingly in focus. Some scholars focus on affection as a cause of cognition, others on the reverse, but the relationship extends beyond causes and effects: Cognition is affective, and affection is cognition - they are interwoven. A feeling is a belief in what mathematics is, why it is essential to master mathematics, and is part of students' motivation. In the teacher training program, the interwovenness was labeled MO-FORMANCE from motivation-performance. Principle 3 states it is the responsibility of teahers to support the students' motivational and cognitive development in mathematics.

Principle 4: Initiate conversations with and among students about mathematical processes and concepts to boost mathematical understanding

In international comparisons like TIMSS and PISA, students' ability to communicate mathematical results is part of the skills measured. Communicative competence is a goal for mathematics instruction. Moreover, communication is a means of mathematics instruction, as students develop their mathematical thoughts and ideas by reading, listening, writing, and drawing mathematical and everyday words and symbols. Students' mathematical understanding and skills develop through individual activities and interaction with peers and teachers. Principle 4 states the teacher's decisive role in setting the scene for these interactions and initiating conversations with and among students about mathematical processes and concepts.

Principle 5: Set realistic but high expectations

The rationale for encouraging teachers to set realistic, but high expectations of students, is partly based on research on student ability tracking and partly on teacher perceptions of students with mathematics difficulties. Scherer et al. (2017) discuss the causes of mathematics difficulties, where one extreme is to regard difficulties as errors by the individual (neurological or psychological), and the other as a result of failures in the educational system (didactic) or other social features (sociological). Scherer et al. (2017) argue that teachers who are adherents of neurological and psychological theories see these students as deficient in skills, slow, and unable to learn. Those with more relational views

on math difficulties believe that all students have the potential to learn mathematics. Faragher et al. (2008) provide evidence that everyone can learn math by referring to students with Down syndrome. Another source of inspiration is the mathematics education literature on equity. This literature expands the opportunities-to-learn concept to include access to mathematical content and positional identities (Esmonde, 2011), which underline teachers' expectations of students' potential as decisive for achieving effects. Principle 5 focuses on not set too low expectations for low-achieving students as it might limit their learning opportunities.

Principle 6: Create a logbook to activate students' concentration and reflections and to support long-term memory

Writing a logbook is recommended to activate students' reflections. Students have classes in several subjects during a school day, rushing, for instance, from history to mathematics and then on to a foreign language class, which may challenge their memory skills. Reserving some minutes at the end of each session to summarize and document a few thoughts about the content, recorded in written notes or orally, seems to help students remember what they have learned. Bligh (2000) recommends advising students to document their questions and problems immediately, or else they will forget (p.145). Students' notes are valuable for teachers too. They allow teachers to get a sense of what students have learned and what confuses them. Instant feedback from students, for instance, a few minutes of writing at the end of a session, can help teachers adjust subsequent sessions to student needs (Center for Excellence, 2021).

B Evidence based on the autumn 2016 survey

After the first intervention period in 2016, we surveyed teachers in full treatment, funding-only, and control schools. Several teachers per school were invited to participate, and we got at least one answer from almost every school. Here we provide more details on potential implementation issues in the first intervention year (B.1) and on compliance, group sizes, and the organization of group instruction in treatment and control schools (B.2).

B.1 Potential implementation issues in the first intervention year

Only about 25 percent of the teachers in wave 1 schools (with the full treatment) responded that they had received sufficient information, and 35 percent the teacher training had enabled them to implement the intervention as intended. This partly reflected implementation challenges, e.g., that lesson plans and material for the small group teachers were not ready at the start of the intervention. There was also confusion regarding the intervention being part of a research project. In general, schools and teachers accepted the group assignment rules. However, some were unsure about implementing the didactic method and the extent to which they were allowed to use their professional judgment. UDE, previously unfamiliar with experiments, was not always able to provide clear answers. In part, this reflected a hands-off approach from the researchers and a desire from UDE for the intervention to be like a standard intervention.[33] After the initial months, lesson plans and other materials were ready. By the second year, 75 percent of the teachers answered that they had sufficient information and 70 percent enough training.

The schools in wave 2, which obtained funding but no teacher training, received the same explicit instructions on group sizes and which students to include in the small and large groups. However, the didactic method was for the schools and teachers to decide. Reports from these schools suggest nothing but satisfaction with the extra funding. We cannot rule out that there might have been issues associated with the fact that the funding-only treatment would be replaced by the full treatment the next year. The reports do not suggest that teachers in the funding-only saw the first year as a pilot.

[33] With experimental interventions, there is always the question of how far the effects can be extrapolated to non-experimental settings. To mimic a regular non-experimental intervention, the researchers initially had limited contact with the teachers. Because there was confusion concerning what the teachers were supposed to do, researcher visibility increased during the first year. It may have reduced the external validity of the experiment. However, researchers were limited to presenting the project and avoiding confusion that would not exist in a non-experimental setting.

Teachers may have been aware that other teachers received training. However, it is not uncommon that some teachers receive training and others do not, and it is not clear to what extent the teachers wanted the training ex-ante.

B.2 Small-group instruction in treatment and control schools

The survey included questions on the use of small-group instruction and the sizes of such groups. Most teachers in the full intervention and funding-only schools and about 40 percent in the control schools reported using small-group instruction. Figure A5 in Appendix C shows the numbers of students and groups in the funding-only and control schools.[34] We see that in both funding-only and control schools, there is extensive use of small-group instruction. However, in funding-only schools, the group sizes are more homogeneous and in line with the intervention (which limited group size to eight students in the first year). Disregarding groupings of more than 12 students, 70 percent of students in schools receiving funding only are in groups of 5-8 students versus 45 percent in control schools.

In total, 108 students get small-group instruction in the funding-only schools, and 157 students in groups of 12 students or less in control schools. It is hard to know how accurate and complete the data from the control schools are. However, as we have responses from all control schools (either group sizes or that they did not use small-group instruction), the number may be representative. Thus the intervention approximately doubled the number of students receiving small-group instruction during the intervention period (i.e., 299 students received small-group instruction in the treated schools in 2017/18).

While the funding-only schools had the same rule for assigning students to small groups as the schools receiving teacher training, the control schools had no such instructions. We might expect the students receiving small-group instruction in the control schools to overlap with our target group. Small-group instruction is often used in remedial teaching for low-performing students, so target students are likely over-represented among students receiving small-group instruction in control schools.

[34] We got survey responses from teachers in 15 of 16 funding-only schools and all control schools. 28 teachers from 14 funding-only schools and 24 teachers from 14 control schools reported using small-group instruction. The reports from teachers at funding-only schools were consistent, while there was more variation within the control schools. In Figure A5, we have used the answer that reports the largest number of groups.

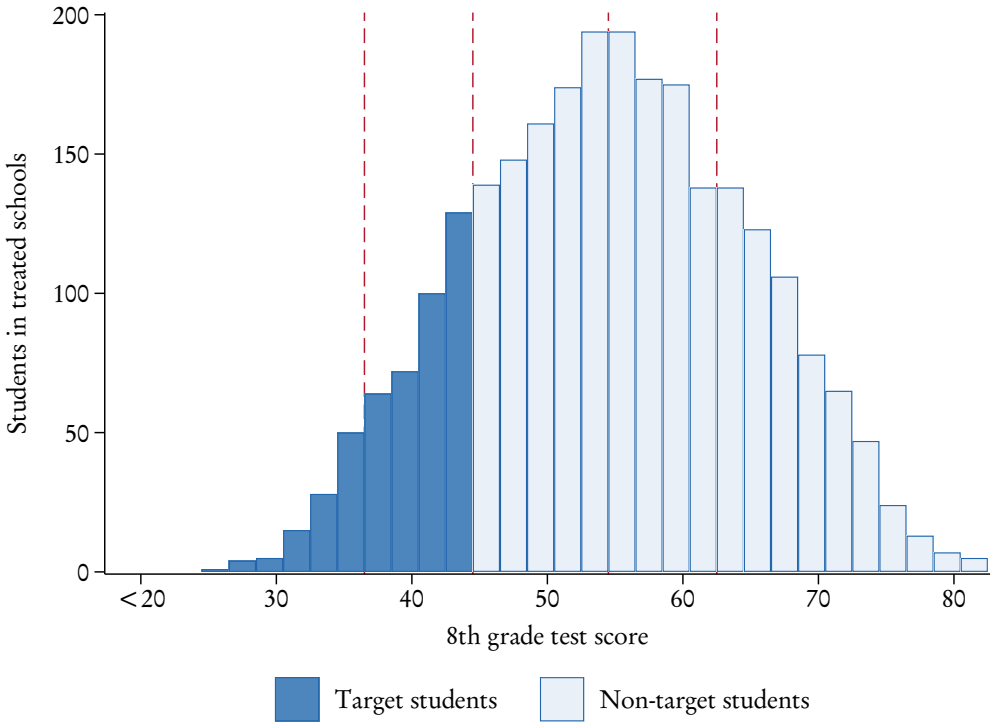
C Supplementary figures and tables

Table A1: Treatment characteristics

	1. year	2. year	3. year
Training program for math teachers			
Number of workshops	2	7	7
Hours of training	20	26	26
Instructors	Qualified national and international personnel		
Math instruction for 8th-graders			
Small group instruction			
Duration	6-8 weeks	6-8 weeks	6-8 weeks
Group size	max 8	max 6	max 6
Large group instruction			
Duration	6-8 weeks	6-8 weeks	6-8 weeks
Group size	Regular class minus small group students		
Group instruction only (no teacher training)			
Duration	6-8 weeks		
Group size	As for small and large groups		

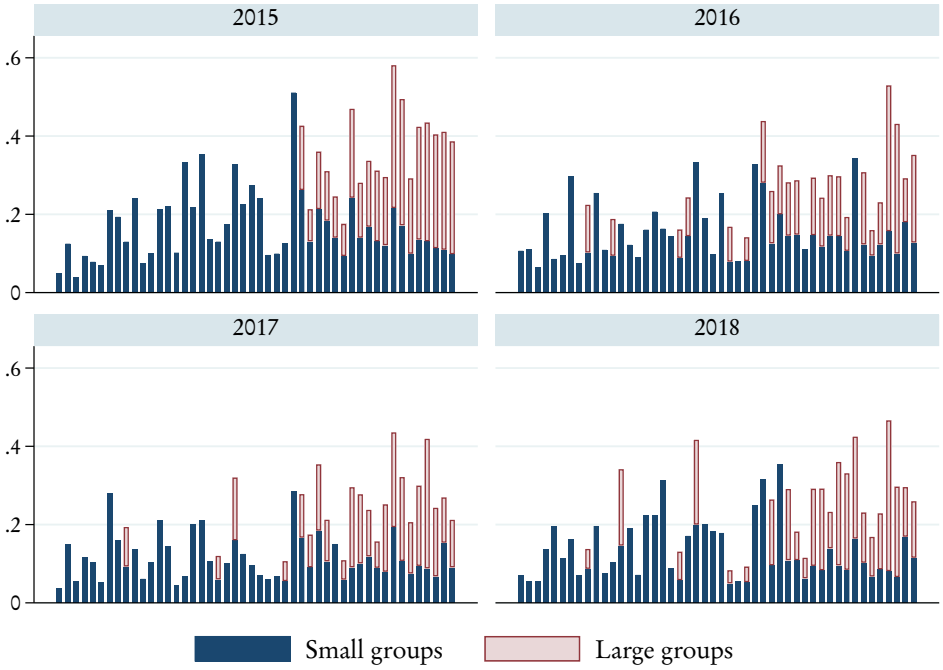
Note: 48 schools participated in the experiment. Math instruction for 8th-graders replaced regular math instruction. In the first year, only eight treatment schools received full treatment, while 16 out of 24 treatment schools received the funding-only treatment, i.e., funding for group instruction and no teacher training. Changes were made in the teacher training program and the size of the small groups between the first and the second year as indicated in the table.

Figure A1: The distribution of numeracy test scores of 8th graders, fall 2017



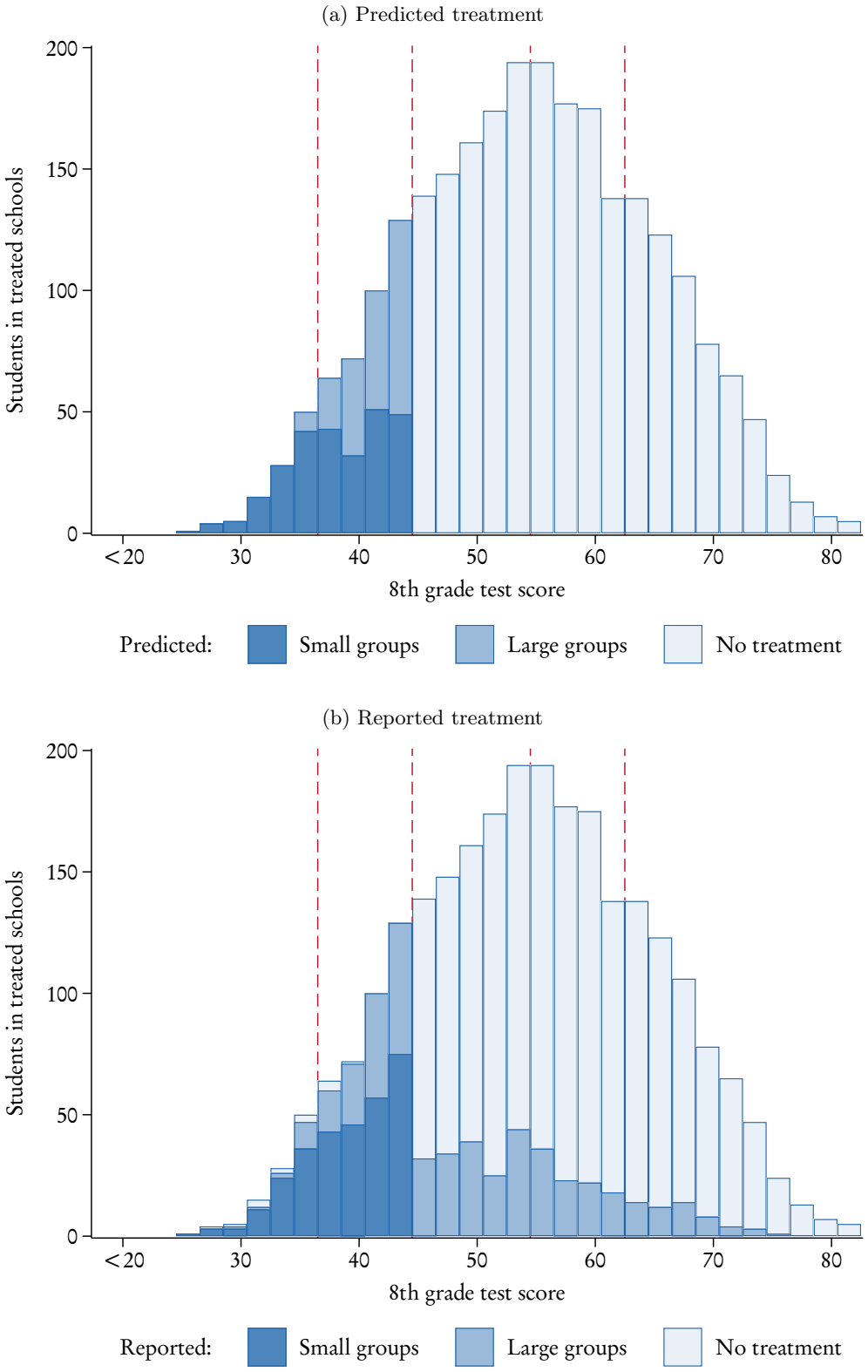
Note: Test scores have a national average of 50 and a standard deviation of 10. Vertical lines separate proficiency levels 1-5: Level 1 is test score $\leq$36, level 2 is test score $\in$ [37, 44] etc.

Figure A2: Share of target students per school and year



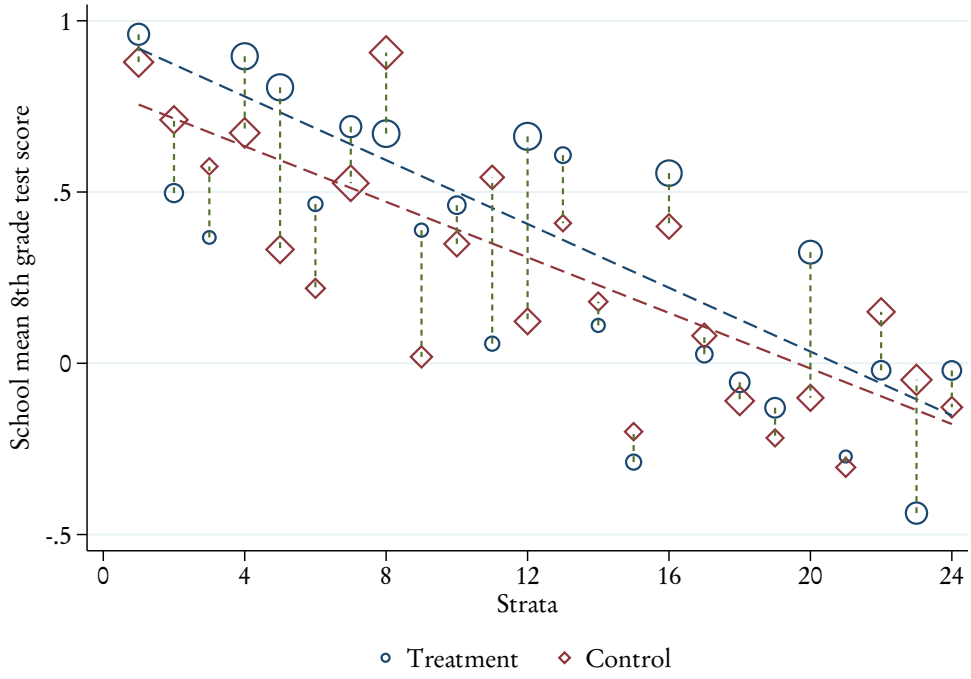
Note: Each bar represents the share of target students in one school and year. The bars distinguish between the share of target students predicted to get small and large group instruction if the school participates in the intervention. In 2015 (the year used as the basis for stratifying schools) and 2016 (the first intervention year), we use the 2016 maximum small-group size of eight students, while in 2017 and 2018, the reduced group size of six students.

Figure A3: Predicted and reported treatment by 8th-grade score, autumn 2017



Note: Predicted treatment is based on 8th-grade test scores and the assignment rule. Reported as by the schools.

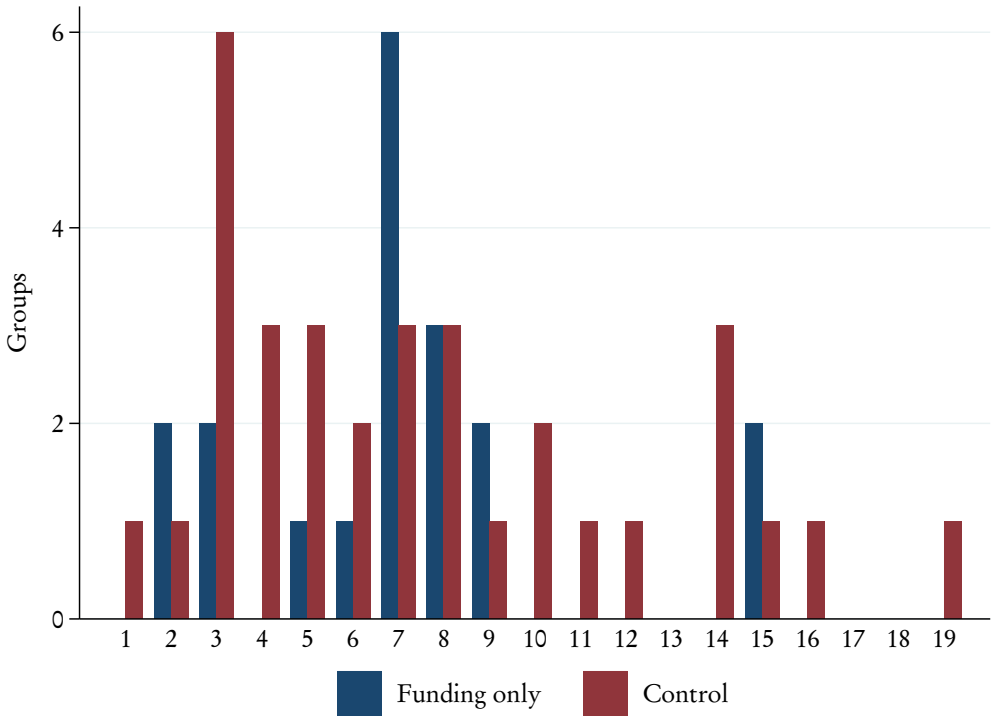
Figure A4: Average 8th-grade score by strata and treatment status



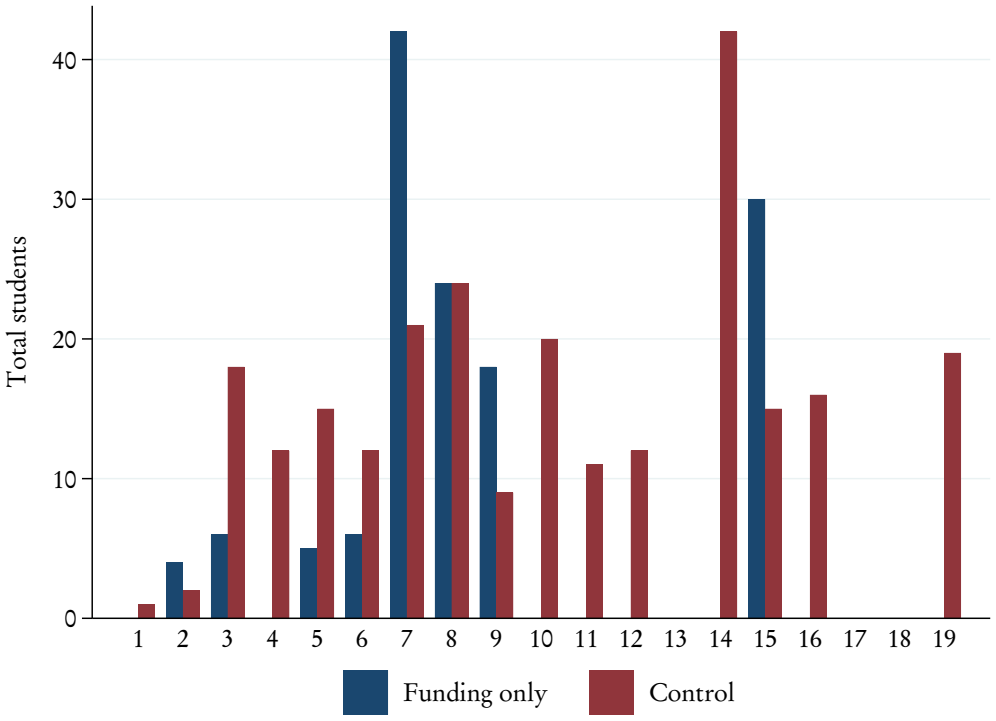
Note: The figure show school average 8th-grade math score in the main estimation sample (i.e., 2017/18 and 2018/19 students) by treatment status and strata. Larger markers indicate more students.

Figure A5: Number of groups and the total number of students in groups, autumn 2016

(a) Predicted treatment



(b) Reported treatment



Note: Based on teachers survey responses.

Table A2: Comparison of treatment and control schools

	Treated schools		Control schools		t-test (treated-control)		Student-weighted regression	
	Mean	SD	Mean	SD	Difference	SE	Coefficient	SE
Student characteristics (estimation sample, 2018/19 students)								
Number of students per year	120.1	(62.0)	110.0	(43.5)	10.0	(15.4)	24.2	(15.8)
Share female	0.51	(0.05)	0.49	(0.07)	0.01	(0.02)	-0.00	(0.02)
Share parent with higher education	0.60	(0.24)	0.56	(0.20)	0.04	(0.06)	0.07	(0.06)
Share foreign-born parents	0.59	(0.42)	0.72	(0.46)	-0.13	(0.13)	-0.16	(0.12)
Average 8th grade numeracy score (y^8)	0.20	(0.42)	0.13	(0.37)	0.07	(0.11)	0.11	(0.12)
School characteristics (the compulsory school register, 2017/18)								
Share combined schools	0.50	(0.51)	0.42	(0.50)	0.08	(0.15)	0.053	(0.14)
Number of lower secondary students	354.5	(152.2)	329.8	(129.9)	24.6	(40.9)	53.5	(40.6)
Number of primary school students	198.5	(213.8)	161.6	(210.0)	36.9	(61.2)	9.9	(56.1)
Average class size	23.3	(3.09)	23.5	(3.80)	-0.19	(1.00)	-0.18	(0.88)
Special needs teaching/regular teaching	0.25	(0.18)	0.21	(0.15)	-0.04	(0.05)	0.02	(0.04)
Norwegian for immigrants/regular teaching	0.10	(0.10)	0.14	(0.12)	0.04	(0.03)	-0.04	(0.03)
Share qualified teachers	0.97	(0.05)	0.97	(0.04)	-0.00	(0.01)	-0.00	(0.01)
Number of qualified math teachers	8.25	(3.26)	7.83	(2.65)	0.42	(0.86)	0.45	(0.75)
Teacher characteristics (employer-employee data, Oct. 2017)								
Average experience	11.1	(2.04)	11.0	(2.15)	-0.12	(0.61)	0.06	(0.58)
Share female teachers	0.70	(0.06)	0.67	(0.07)	-0.03	(0.02)	0.03	(0.02)*
Average age	39.8	(2.5)	40.0	(3.1)	0.2	(0.8)	-0.11	(0.78)
Sickness absence (hours per week, 2019)	1.63	(0.85)	1.62	(1.07)	-0.01	(0.28)	0.03	(0.27)
Number of schools	24		24		48		48	

Note: Data consists of all school averages for the indicated variables for all treatment and control schools. The first four columns show means and standard deviations by treatment status. The last four columns show the (school-level) difference and estimated standard error of the difference, the first two using equal weights for all schools and the last two weighed with the number of students in the sample (sum of weights is 5345). Sickness absence is not available in the Oct 2017 data. Thus, we use sickness absence for the October 2017 teachers from March 2019.

Table A3: Balancing of 8th grade scores by student group, 2017/18 and 2018/19

	(1) Small-group students	(2) Large-group students	(3) Non-target students
Estimates from specification with			
No controls	0.068	-0.027	0.040
	[-0.028, 0.165]	[-0.082, 0.029]	[-0.061, 0.130]
Family controls	0.047	-0.031	-0.005
	[-0.041, 0.132]	[-0.080, 0.022]	[-0.073, 0.063]
N	1142	835	7953
N clusters	48	25	48
$\bar{y}$	-1.241	-0.866	0.722

Note: Each cell gives an estimate of θ from equation (1) for a given outcome and student sample (column) and set of controls (rows). See note to Table 2 for details. Wild bootstrap 95 percent confidence sets in brackets. Statistical significance (from wild bootstrap test): ** 5 percent level and * 10 percent level.

Table A4: Share of randomizations producing a difference in 8th-grade numeracy greater than observed by the procedure for stratification

	8th grade test score			Number of target students	
Stratification	Share absolute weighted difference >.076	Mean absolute weighted difference	Mean absolute unweighted difference	Mean absolute weighted difference	Mean absolute unweighted difference
(0) None	0.560	0.104	0.090	4.002	3.501
(1) *Number and share target students*	0.326	0.061	0.051	2.667	2.465
(2) Two-year number and share target students	0.221	0.049	0.046	2.703	2.656
(3) One-year mean 8th grade score	0.081	0.035	0.037	2.842	3.014
(4) Two-year mean 8th grade score	0.110	0.038	0.038	2.764	3.176
(5) Number of target students	0.349	0.064	0.063	1.605	1.725
(6) Share of target students	0.187	0.046	0.041	2.633	2.775

Note: For each stratification, we have sorted the schools into 24 strata, randomized the schools to treatment and control 10,000 times and compared average 8th-grade numeracy and number of target students in treatment and control schools. The first column shows the share of student-weighted treatment-control differences greater than the difference in the data (0.076 SD). Student-weighted differences correspond to the main analyses, as these are at the student level. The remaining columns show: mean absolute student-weighted and unweighted differences in average 8th-grade test score and the number of target students. One-year stratification is based on 2015/16 students, and two-year stratification is based on 2014/15 and 2015/16. Comparison of treatment and control schools is done using the main estimation sample consisting of 2017/18 and 2018/19 students.

Table A5: Effects of treatments in the first year for different student groups, all 2016/17 students

	(1) Dummy has y^9	(2) 9th grade score (y^9)	(3) Lowest proficiency (D^{L1})	(4) Low proficiency (D^{L2})
Treatment: Teacher training and group instruction				
Small groups	0.041	-0.098	0.017	0.047
	[-0.061, 0.142]	[-0.263, 0.063]	[-0.083, 0.134]	[-0.050, 0.148]
Large groups	-0.023*	-0.183	0.077	0.160
	[-0.078, 0.001]	[-0.547, 0.373]	[-0.161, 0.214]	[-0.289, 0.477]
Non-target students	0.021*	-0.011	0.001	0.014
	[-0.002, 0.043]	[-0.129, 0.107]	[-0.004, 0.008]	[-0.003, 0.034]
	0.064	0.849	0.805	0.112
Treatment: Group instruction only (no teacher training)				
Small groups	0.049	-0.065	-0.038*	0.024
	[-0.046, 0.138]	[-0.217, 0.083]	[-0.080, 0.001]	[-0.094, 0.141]
Large groups	-0.030	-0.083	0.038*	-0.016
	[-0.124, 0.077]	[-0.173, 0.112]	[-0.000, 0.090]	[-0.138, 0.088]
Non-target students	-0.009	0.005	-0.000	0.011
	[-0.027, 0.010]	[-0.068, 0.068]	[-0.000, 0.000]	[-0.006, 0.028]
Family and y^8 controls	No	Yes	Yes	Yes
N	5261	4955	4955	4955
N clusters	48	48	48	48
$\bar{y}$	0.942	0.668	0.030	0.137

Note: Each cell gives an estimate of θ from equation (1) for a given outcome (column) and treatment and student group (rows). Sample is all students in 2015/2016, with separate specifications for each group of students, similar to tables 3, 4 and A6. See note to Table 3 for details about outcomes. All specifications control for strata, controls for family background and cubic in y^8 where indicated. Wild bootstrap 95 percent confidence sets in brackets. Statistical significance (from wild bootstrap test): ** 5 percent level and * 10 percent level.

Table A6: Treatment effects, non-target students 2017/18 and 2018/19

	(1) Dummy has y^9	(2) 9th grade score (y^9)	(3) Lowest proficiency (D^{L1})	(4) Low proficiency (D^{L2})
Effect estimates from specification with				
No controls	0.012	0.043	0.000	0.002
	[-0.001, 0.025]	[-0.060, 0.149]	[-0.001, 0.001]	[-0.008, 0.011]
Family controls	0.009	-0.008	0.001	0.007
	[-0.003, 0.022]	[-0.069, 0.058]	[-0.001, 0.002]	[-0.002, 0.015]
Family $+y^8$ controls	0.009	-0.001	0.001	0.006
	[-0.004, 0.022]	[-0.043, 0.037]	[-0.001, 0.002]	[-0.002, 0.014]
N	7953	7597	7597	7597
N clusters	48	48	48	48
$\bar{y}$	0.955	0.994	0.001	0.025

Note: Each cell gives an estimate of θ from equation (1) for a given outcome (column) and set of controls (rows). See note to Table 3 for details. The sample is non-target students in years 2017/2018 and 2018/2019. Wild bootstrap 95 percent confidence sets in brackets. Statistical significance (from wild bootstrap test): ** 5 percent level and * 10 percent level.