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

Working Paper

Didactic methods and small-group instruction for low-performing adolescents in mathematics: Results from a randomized controlled trial

Discussion Papers, No. 957

Provided in Cooperation with:
Research Department, Statistics Norway, Oslo

Suggested Citation: Kirkebøen, Lars J.; Gunnes, Trude; Lindenskov, Lena; Rønning, Marte (2021) : Didactic methods and small-group instruction for low-performing adolescents in mathematics: Results from a randomized controlled trial, Discussion Papers, No. 957, Statistics Norway, Research Department, Oslo

This Version is available at:
https://hdl.handle.net/10419/249147

Statistisk sentralbyrå
Statistics Norway

Didactic methods and small-group instruction for low-performing adolescents in mathematics: Results from a randomized controlled trial

TALL SOM FORTELLER
DISCUSSION PAPERS
957

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

Discussion Papers No. 957, June 2021
Statistics Norway, Research Department

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

Didactic methods and small-group instruction for low-performing adolescents in mathematics: Results from a randomized controlled trial

Abstract:
Can high-dosage tutoring help low-performing adolescents? We implement a randomized experiment to test a twofold intervention: A teacher training program customized for instructing 8th graders who perform poorly in mathematics and two 4-6 week periods of targeted math instruction for low-performing 8th graders, a majority in small homogeneous groups and the rest in larger and more heterogeneous groups. We randomized 24 schools to treatment and 24 schools to control. For students receiving small-group instruction, we find that test scores increase by .06 SD. Moreover, the share of low-performing students decreases by up to 25 percent. We find no impact on treated students in large groups. Classroom observations and surveys to teachers indicate higher fidelity to the didactic methods among teachers managing small groups.

Keywords: Low-performing students, Ability grouping, High-dosage tutoring, Classroom management, Didactic methods, Mathematics, RCT, Stratified randomization, Cost-benefit of interventions

JEL classification: I21, I24, I28

Acknowledgements: 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. The usual disclaimer applies.

Address: Statistics Norway, Research Department. E-mail: kir@ssb.no

Discussion Papers comprise research papers intended for international journals or books. A preprint of a Discussion Paper may be longer and more elaborate than a standard journal article, as it may include intermediate calculations and background material etc.

Abstracts with downloadable Discussion Papers
in PDF are available on the Internet:
http://www.ssb.no/en/forskning/discussion-papers
http://ideas.repec.org/s/ssb/dispap.html

ISSN 1892-753X (electronic)

Sammendrag

Denne artikkelen studerer et forsøk med tilrettelagt matematikkundervisning på for elever med svake resultater fra nasjonale prøver. Tiltaket består av to deler: Kursing av lærere i didaktikk tilpasset elever med lav kompetanse i matematikk og tilrettelagt undervisning for elever i en klart definert målgruppe i to perioder på fire til seks uker på 8. trinn. Et flertall av målgruppeelevene fkk tilpasset opplæring i små grupper, bestående av elever med svake resultater fra nasjonale prøver, de øvrige elevene i større grupper. Tiltaket ble gjennomført som et randomisert forsøk, der 24 av 48 ungdomsskoler i Oslo ble tilfeldig valgt ut til å delta. Dette gjør at vi kan studere effekter av tiltaket ved sammenligne resultater for forskjellige grupper av elever i tiltaks- og kontrollskoler.

Vi finner at elever som fikk tilrettelagt undervisning av kursede lærere på 8. trinn får et resultat på nasjonal prøve i regning på 9. trinn som er omtrent 6 prosent av et standardavvik (tilsvarende 0,6 skalapoeng) høyere enn sammenligningsgruppen, og i mindre grad presterer på de laveste mestringsnivåene. En økonomisk verdsetting av denne effekten, basert på andre studier av sammenhengen mellom skoleresultater og arbeidsmarkedsutfall, tyder på at gevinsten er klart større enn kostnaden av tiltaket.

Vi finner ingen effekter på resultatene til elever som fikk opplæring av kursede lærere i store grupper. Vi finner heller ingen effekter på elever som, i det første forsøksåret, fikk opplæring i små grupper uten at lærerne fikk kursing. Verken kursing av lærere eller små grupper ser ut til å være tilstrekkelig til å gi mer læring hver for seg.

Klasseromsobservasjoner og læreres svar i spørreundersøkelser viser at undervisningen i de små gruppene med kursede lærere i stor grad bruker didaktikken de ble kurset i, mens dette i mindre grad er tilfellet i de store gruppene. I tillegg til å være mindre har de små gruppene mindre variasjon i elevenes faglige nivå, og lærerne som underviste små grupper fikk en litt annen kursing enn lærerne som underviste store grupper. Alt dette kan ha bidratt til forskjeller i bruk av didaktikken fra tiltaket og forskjeller i effekter for elevene.

1 Introduction

Youths from families with low socioeconomic status (SES) are over-represented among those who perform poorly in school and have lower prospects for labor market careers. Reducing achievement gaps among socioeconomic groups and increasing educational attainment among low SES students is high on the political agenda, and research points to the importance of math skills to complete high school (e.g., Duncan et al., 2007). Although previous research to a large degree concludes that early investments are more beneficial than later investments (Carneiro and Heckman, 2003; Heckman, 2013), recent findings indicate high returns from programs directed towards adolescents with low numeracy skills (Cook et al., 2014; Clotfelter et al., 2015; Cortes et al., 2015; Fryer and Howard-Noveck, 2020; Guryan et al., 2021).[1]

This paper contributes to this burgeoning literature arguing that it is not too late to implement interventions for adolescents falling behind. We design and test an intervention targeting 8th-graders with low numeracy skills. The intervention combined customized training for qualified math teachers with targeted instruction in two periods (each lasting 4-6 weeks) for low-performing students, mostly in small groups of six or fewer students. The intensive math course replaced regular math classes during the intervention period, and the small group instruction largely corresponds to what Fryer (2017) defines as high-dosage tutoring.[2] Due to organizational constraints, some target students got instruction in larger groups taught by newly trained teachers, mostly in their regular classes. In the first year, some randomly selected schools only got funding for small and large-group instruction, and no teacher training. While the ultimate objective of the intervention is to increase the proportion of students completing high school, in this paper, we study shorter-term effects on numeracy skills in the

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

[2]Fryer (2017) describes 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), the total extent of instruction (three hours per week for 9-12 weeks, i.e., 27-36 hours) may be somewhat less than Fryer (2017) classifies as high-dosage.

year after the treatment.

Our intervention combines small- and large-group instruction for low-performing students with teacher training. The training program built on well-known didactic methods but focused on how these specific targeted didactic principles and tools can be combined, re-composed, and used to boost the achievement of low-performing students (Torgerson et al., 2012; Harder et al., 2020; Pellegrini et al., 2021). Many of these methods have proven to be successful in lower grades. The idea is to apply some of the didactic methods used in lower grade levels to boost achievement among low-performing students in higher grades.

The intervention took place in 2016/17, 2017/18, and 2018/19. We randomly selected 24 out of 48 lower secondary schools in Oslo (the capital of Norway) to participate, one from each of 24 matched pairs (following the recommendations of Bruhn and McKenzie, 2009). Schools were matched on the number and share of low-performing students, and we show that stratifying schools significantly reduced the ex-ante probability of imbalances. Still, we demonstrate that our sample of 48 schools is sufficiently heterogeneous to produce imbalanced groups with a high probability, even with pairwise matching, and we do find imbalances in pre-determined characteristics across treatment and control schools. However, since we have good controls for pre-existing differences, we can still provide credible effect estimates, despite the imbalances (Lin, 2013).

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. The share of low-performing students is reduced by about 3 percentage points, corresponding to a reduction of 5-25 percent for different measures of low performance. Using other studies to value our results, we conclude that the small-group intervention is cost-effective, with an estimated cost per small-group student of USD 1200-1800 and estimated benefits of USD 3700. Our incomplete data on small-group assignment suggest that 89 percent of students predicted to get instruction in small groups actually do get it, implying a treatment effect of 0.067 SD on the treated. We find no impact

on target students who receive instruction from newly-trained teachers in large groups. There is also no indication from the first year that providing instruction in small or large groups without teacher training influences achievement.

Our paper contributes to the literature on experimental teaching interventions in schools, and has several similarities with Guryan et al. (2021). We find similar effect per dollar and cost-benefit ratios for adolescents as Guryan et al. (2021).[3] However, despite important similarities, our intervention and context differ from Guryan et al. (2021) in several ways. First, the teachers teach small groups of students, requiring fewer teachers than more individualized one-on-one tutoring. Second, the targeted instruction in our case replaces regular math instruction for two limited periods. While perhaps contributed to a lower effect, these differences reduce the cost of the intervention. Guryan et al. (2021) rely on relatively low-cost tutors. In other contexts, such tutors may not be the available.[4] We demonstrate that we can achieve effects per dollar similar to Guryan et al. (2021) with regular teachers and little disruption to schedules (as the targeted instruction does not replace other subjects).

Extensive supplementary data allow us to further investigate and expand upon the findings from the effect analyses, and contribute to the burgeoning literature on teacher fidelity to implementing new didactic principles and tools (e.g., Durlak et al. 2011). Classroom observations and surveys to teachers show high teacher fidelity to the didactic methods in the small groups, but lower in the larger groups. Teacher satisfaction is also higher in the small group. The paper demonstrates how extra funding can help implement effective teaching

[3]Guryan et al. (2021) carried out an RCT 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 find that personalization of the instruction increased math test scores by 0.16 percent of a standard deviation. They do not implement any particular didactic methods. 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.

[4]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.

strategies, which has often proved difficult (e.g., Forgasz, 2010; Rønning et al., 2013; Jacob, 2017).[5] Small-group instruction with homogeneous students simplifies the teaching task in several ways. Teachers need to spend less time on classroom management and are left with more time to concentrate their effort on teaching to one academic level (e.g., Connor et al., 2013). Thus, our paper also relates to the literature on ability tracking (e.g., Duflo et al., 2011).[6]

Finally, we contribute to the literature on the practical design and implementation of moderate-scale 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). 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 investigate how our population of 48 schools can give imbalanced treatment and control groups and to what extent this can be mitigated ex-ante by stratifying on different variables. In particular, our findings highlight the tension between a desire to balance several characteristics and to better balance one (c.f. Bruhn and McKenzie, 2009). In our case, a small increase in the expected balance of school size comes at the cost of substantially reduced expected balance in baseline outcomes. 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 methods, organization, 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

[5]The literature on teaching practices (e.g., Kane et al.,2011; Bietenbeck, 2014; Lavy, 2016; and Aucejo, 2018) focuses on mapping teaching practices to student types. It is less concerned with implementation issues.

[6]The evidence on ability tracking is mixed (Cortes and Goodman, 2014). Mainly, the effect depends on to what extent the teaching matches the level of the ability group.That is, ability tracking affects students in both the top and bottom halves of the achievement distribution if the benefits of better-targeted pedagogy (i.e., personalization) outweigh the negative impact of being exposed to lower-skilled peers (Duflo et al., 2011; Guryan et al., 2021).

effect estimates, and section 6 discusses channels of impact, that is, the implementation quality of the didactic methods. Section 7 provides a cost-benefit analysis and section 8 concludes.

2 Institutional setting

Compulsory education in Norway consists of seven years of primary education and three years of lower secondary education. Children start primary school the year they turn six. Schools at the primary and the secondary level are almost all public and have a local catchment area.[7] Early/late starting and grade retention are rare, such that nearly everybody starts middle school the year they turn fourteen. Ability tracking is controversial in Norway, and persistent ability tracking is not allowed. There are standardized national tests in numeracy, literacy, and English in 5th, 8th and 9th grade. 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 heterogeneous in terms of parents' education and ethnic background. There are substantial differences between schools, reflecting residential segregation. Within municipalities, school funding is compensatory, such that schools with students of less advantageous backgrounds get increased funding.

High school is not compulsory, but students are entitled to three years of upper secondary education. Almost all students start high school directly after lower secondary education. However, about 25 percent do not complete within five years. For many students, passing mathematics is a binding constraint for completing upper secondary education. Thus, better numeracy skills will enable more students to graduate from high school. Furthermore, an improved understanding of mathematics may create a greater sense of mastery, which low-performers may be lacking. Low completion rates are a policy concern and the backdrop for

[7]Parents can apply for 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.

the intervention.

3 The intervention

The intervention ran during the school years 2016/17, 2017/18, and 2018/19 and consisted of teacher training and targeted instruction of students in 8th grade with low proficiency in mathematics. In the remainder of the paper, we denote these students as target students. In the first part of the intervention, qualified teachers attended the training program that provided them with didactic principles and tools adapted for students who perform poorly in mathematics. Then, in the second part, target students from 24 treatment schools received two periods (5-6 weeks during October-November and 4-6 weeks around April) of instruction by the newly trained teachers. The targeted math instruction replaced regular instruction in mathematics, typically three hours per week, during the intervention period.[8] A majority of the target students were in small groups consisting of six or fewer students. The remaining minority stayed mainly in their regular classes (large groups). The small-group treatment fits Fryer's (2017) definition of high-dosage tutoring (see footnote 2). The first year served as a pilot. We get back to how the pilot year differed from the last two years in section 3.3.

3.1 The didactic methods and organization of the teacher training

According to Valenta (2015), five components are crucial for understanding numerical reasoning: Conceptual understanding, calculation, application strategies, rational thinking, and commitment. Previous tests and analyses by the local school authorities in Oslo (UDE) show that the target students have poor comprehension of these five components, suffer from misconceptions, and have little learning effect of ordinary teaching. Without (basic) knowledge

[8]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 8th grade.

and skills from primary education, the target students lack the prerequisites for mastering mathematics at the lower secondary level, and their challenges propagate (Borg et al., 2014). Identified shortcomings and misconceptions have influenced the mathematics content covered and didactic methods used in the intervention.

UDE was responsible for the content and organization of the teacher training program. The Danish School of Education (DPU) provided professional guidance. DPU has extensive experience with research on students with low math skills. They have conducted several interventions to improve students' numeracy skills (Jankvist and Niss, 2015; Lindenskov and Tonnesen, 2020; Harder et al., 2020).

The didactic methods are based on internationally acknowledged teaching practices and supplemented with experience based on other Norwegian teacher training programs. The didactic methods consist of principles and tools. DPU and UDE incorporated six principles into the teacher training program and the instruction of students. (i) Create a link between learning sessions to activate student memory of mathematical concepts and help 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 so everybody can reach their potential. (iii) Foster motivation leading to improved performance, 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 prepared and 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. 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 methods in the professional community, representatives from the school administration also attended the training.

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 would then receive lectures and participate in workshops during autumn and spring. The focus was on the theoretical and practical aspects of implementing the new didactic methods for low-achieving math students.

The teacher training program separated small and large group teachers. The six didactic principles and four didactic tools were the same for small and large group teachers. However, teachers selected to teach small groups got additional instruction materials, including concrete lesson plans and exercises. Teachers teaching large groups did not receive any. The rationale was to let the large group teachers themselves adapt standard materials when appropriate.

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

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

3.2 Organization and funding of small- and large-group instruction

The 24 treatment schools received funding for small-group instruction for the three years the intervention lasted. The exact amount of funding depended on the number of students belonging to the target group in 8th grade in 2015/16 (i.e., the year before the first year of the intervention). The remaining 24 control schools only received information (at the management level) about the experiment.

Providing small-group instruction for all target students would require many small groups in some schools, putting demands on available classrooms and qualified teachers. In coordination with UDE, we decided that there would be a maximum of (three) small groups per school. 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 financial support to create two small groups for the 12 lowest-performing students and a smaller amount of funding to facilitate instruction in line with the didactic methods of the intervention in larger groups for the remaining target students. Based on information from UDE, large groups coincide with regular classes minus the low-performing students receiving small-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 didactic methods, class size, and class composition during the treatment periods.

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 students to small- or large-group instruction followed explicit assignment rules. Intervention instruction would start early/mid-October, and UDE followed up with the schools during the treatment years. Before each intervention year, UDE informed the treatment schools about the intervention and what it meant in terms of extra funding, teacher training, student selection, implementation of small-group instruction, and reporting.

3.3 The pilot year

Due to limited time for preparing the teacher training program, the first year (the school year 2016/2017) served as a phase-in and a pilot. Only eight of the 24 treatment schools received training for teachers and implemented the full treatment the first year. The remaining 16 treatment schools only received funding for group instruction. They got identical directions concerning which students to assign to small and large groups, teacher qualifications, and the extent and timing of group instruction. However, teachers from these schools did not receive training the first year. In the remainder of the paper, we denote this treatment as funding-only. The size of the small groups was eight students the first year, meaning that schools with up to 24 target group students would have three small groups in the pilot year.[10]

A survey following the first intervention period in 2016 showed that fidelity to the didactic methods among teachers was very low. It was mainly due to a shortage of information and course material (see more in Appendix B.1). Based on experiences from the first year, there were changes also to the teacher training program. In the first year, the sessions were, to a large extent, theoretically oriented and focused primarily on presenting the didactic principles and tools, followed by teacher reflections. To raise fidelity, that is, induce a high-quality implementation of the didactic methods, the training sessions in the two following years included additional workshops. The latest workshop of the training program included practicing and observations in classrooms.

4 Data and empirical strategy

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

[10]According to Fryer's (2017) definition (see footnote 2), the small group instruction in the pilot year is not defined as high-dosage tutoring.

4.1 Data and target students

The data are mainly from national registers or registers from the municipality of Oslo. Additionally, we use self-collected data from teacher surveys and classroom observations to shed light on mechanisms. The national employer-employee register allows us to track teachers across employers. From the National Education Database (NUDB), we have detailed information on students' previous results from standardized national tests in 5th grade (NP5) and 8th grade (NP8). NUDB also provides 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-need education and results on national tests in 9th grade (NP9). UDE also collected data on group assignments in treatment schools.

Our complete student sample includes all students in 8th grade in Oslo in the school years 2016/17, 2017/18, and 2018/19, about 5500 students per year. We focus on 2017/18 and 2018/19 for our main analysis of the intervention and separately study the treatments in the pilot year as described in section 3.3.[11] We exclude students receiving special needs education, as they already receive customized instruction and were not eligible for targeted instruction in the intervention. Furthermore, we exclude students with no data from the 8th-grade numeracy test, as we are not able to detect whether these students belong to the target group or not. In total, we exclude about 10 percent of the gross sample.[12]

We define target students as those who score at the two lowest proficiency levels (out of five) on the standardized national test in 8th grade, NP8. Figure A1 in Appendix C shows the distribution of test scores on NP8 for fall 2017 (the other years have a very similar distribution). The target group constitutes about 20 percent of the students, i.e., about 1100 students per

[11]The intervention follow the pre-registration published in July 2017 (Kirkebøen, 2017) with one exception: Initially, the treatment was planned to be identical in the three intervention years. Given the changes made in the size of the small groups and the teacher training from the first to the second year, we believe it is more reasonable to analyze the pilot separately as we do in this paper.

[12]4.4 percent lack NP8 while 8.1 percent receive special-need education, with some overlap between these two groups.

year. To ease interpretation of the estimated intervention effects, we will normalize the test scores with the national mean and standard deviation.

Table 1 presents descriptive statistics for our main estimation sample, where we also separate between target and non-target students. 49 percent of the students are female, 36 percent have parents without higher education, and 31 percent have two foreign-born parents. As expected, there is an over-representation of boys and students of lower educated and foreign-born parents among the target students.

Compared to the national average test score, students in Oslo score about 37 percent of a standard deviation better, both on the 8th grade and the 5th-grade numeracy tests. Target students, selected on their 8th-grade performance, score almost 1.1 standard deviation below the national average in grade five and 0.8 standard deviation below in grade eighth. We will use the numeracy test score in 9th grade, which is directly comparable to the 8th-grade score, to measure treatment effects. The average progress from 8th to 9th grade corresponds to about 32 percent of a standard deviation. However, the average improvement of the students belonging to the target group is only about 17 percent. While 20 percent of all students in the sample perform at proficiency level one or two in 8th grade, only 12 percent do so in 9th grade. Among the target students, 10 percent perform at the lowest proficiency level in 9th grade and another 44 percent at the second lowest. Few non-target students perform at the two lowest levels.

4.2 Randomization and implementation of the different treatments

We conducted a randomized controlled trial (RCT) at the school level to evaluate the intervention.[13] The randomization took place in May 2016. Principals of all lower secondary schools in Oslo were informed about the project in February 2016. Shortly after randomization, schools knew whether they were in the treatment or control group, and the treatment schools started

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

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 5 numeracy (y^5)	0.36	-1.08	0.72
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	9929	1977	7952

Note: The sample consists of students sitting 8th-grade numeracy test in 2017 or 2018 in Oslo who do not receive special needs education.

to make plans for teacher training and small and large group instruction.

Schools in Oslo are heterogeneous, with the number of target students in 2015/16 (the year before the intervention and the most recent available test results at the time of randomization) ranging from six to 64. To increase the likelihood of the treatment and control groups 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 way of stratifying schools prior

[14]Matching was done by constructing a distance measure based on standardized numbers and shares of target students. The number of target students crucially impacts the implementation of the intervention (number of small groups and number of target students in large groups), while the share of low-performing students measures the average performance level at the school. To ensure 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 as the share of target students when matching schools. Randomization was done by writing a script that randomized schools. After testing, a random seed was set, and it ran once.

to the randomization follows the recommendations of Bruhn and McKenzie (2009).[15] For the pilot intervention in 2016/17, we randomly selected eight of the 24 treatment schools to full treatment in the following way: After sorting the strata, we pooled them into groups of three (eight groups in total) and selected one of the three treatment schools from each group to 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.

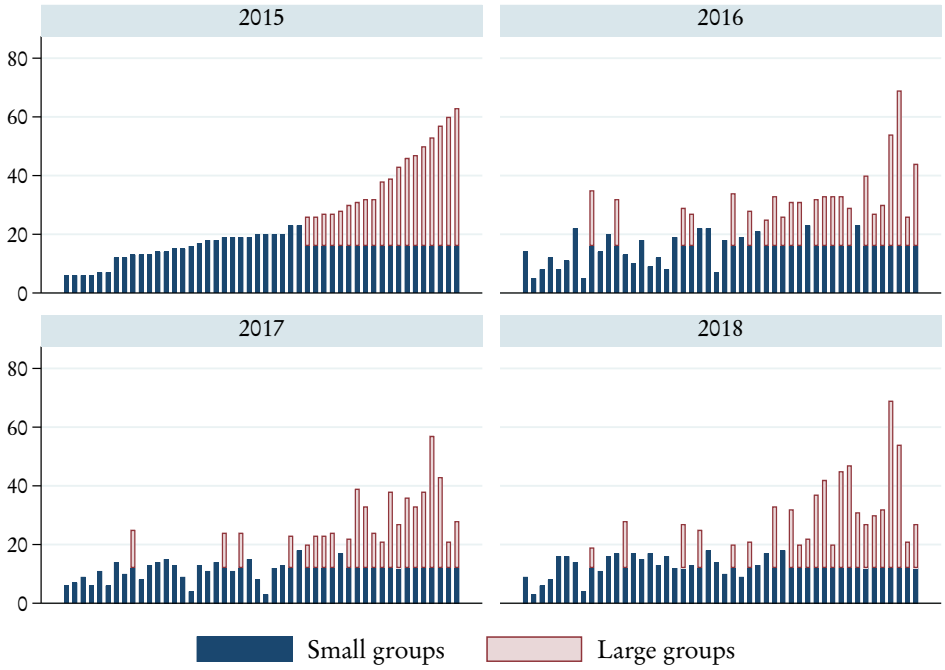
According to the assignment rules and administrative data, 560 target students in treatment schools got instruction in small groups and 400 in larger groups in the school years 2017/18 and 2018/19. In the pilot year, 2016/17, about 130 target students in the eight full-treatment schools got small-group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

The upper left panel of Figure 1 shows the number of target students assigned to small or large groups in the school year 2015/16, i.e., the year used as the basis for stratifying schools for randomization in our sample. The remaining panels show how we distributed target students in small and larger groups in 2016/17 - 2018/19. The number of target students varies, partly due to differences in school size (ranges from 37 to 203 target students) and partly due to differences in 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 equivalent to Figure 1 apart from that it reports the share of target students instead of the number.

For the 2017/18 students, we have data from the municipality in Oslo on actual assignments to small and larger groups. Of 466 target students in treatment schools, 299 got 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

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

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 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 group students in 2015.

numbers. For the lowest-performing students, there is a vast overlap between observed and predicted treatment. 89 percent of the lowest-performing students, who should get small-group instruction according to the assignment rule, do get small-group instruction. However, about 1/3 of the target students predicted to get large-group instruction are reported to get small-group instruction[16], and in some schools, a substantial number of non-target students are reported to get large-group instruction.[17]

[16]About half of these students come from three schools which report 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 getting small-group instruction.

[17]In total, 329 non-target students are reported to get large-group instruction. All these students, apart from 16, belong to seven schools that report that all their students, including non-target students, get small- or large-group instruction. Likely, this is due to mixing large-

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

4.3 Empirical strategy

As we assigned students to small- and large-group treatments based on observed test scores, we can identify the corresponding groups of students in control schools, i.e., the counterfactual outcome. Hence, we can identify the effects for the following 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.

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

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

In the main effect analyses, y_{ist} is the 9th-grade test score of student i at school s in year t. T_s equals 1 if school s is a treated school, 0 otherwise. 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 5th and 8th-grade test scores). We allow for the residuals ε_{ist} to be correlated over time within schools and adjust standard errors for school-level clustering. The number of schools (48) is in line with the

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 when presenting the results. Athey and Imbens (2017) caution against studying RCTs with regression models and recommend using re-sampling methods. While the randomization is done by strata based on data for previous student cohorts, in line with the recommendation of Athey and Imbens (2017), we also have pre-treatment data for the actual participants. Adjusting for individual baseline outcomes has a large impact on precision, our ability to handle (random) imbalances, and heterogeneous effects.

rule-of-thumb, the minimum number of clusters for cluster-robust estimation, to be reliable. However, with heterogeneous cluster sizes, the effective number of clusters is smaller (Cameron and Miller, 2015). Also, in some analyses, we have fewer clusters. Therefore, we also have applied wild bootstrap tests to the estimates and will comment on these tests when presenting the results.[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). Regarding pre-determined student- and school characteristics, we can use the same model framework to investigate whether the treatment and control schools are similar, as expected from the randomization. If alike, we interpret θ as a causal effect of the intervention for post-intervention outcomes. If the treatment and control groups are not alike, we will still get an unbiased effect estimate if we, through γ, δ, and X, manage 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.

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 schools. However, as we only have a limited number of schools, we may still get imbalances by chance.

Table A1 in Appendix C compares treatment and control schools. There is little evidence of systematic differences between treatment and control schools. The only difference which is significant (only at the 10 percent level) is the share of female teachers when weighting with the number of students. There are, however, insignificant differences in student composition. Students in treated schools are: more likely to have parents with tertiary education, less likely

[19]For the main estimates, we have used Stata's cluster option. For the wild bootstrap tests, we use a boot-test with the standard 999 replications (Roodman, 2015). As the wild bootstrap is sampling-based, p-values and confidence sets will vary between replications. We have fixed the random seed to make the results presented reproducible.

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

	(1) Dummy main sample	(2) Background index ($\hat{y^9}$)	(3) 8th grade score (y^8)	(4) Dummy target group	(5) Small-group instruction	(6) Large-group instruction
Effect estimates from specification with						
No controls	0.001	0.099**	0.076*	-0.022*	-0.011	-0.011
	(0.011)	(0.035)	(0.045)	(0.012)	(0.008)	(0.011)
Family controls			-0.005	0.004	0.006	-0.002
			(0.026)	(0.007)	(0.008)	(0.010)
N	11106	9930	9930	9930	9930	9930
$\bar{y}$	0.894	0.596	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 not special needs education), (2) 9th-grade numeracy score predicted from observed family background, (3) 8th-grade numeracy score, (4) dummy for being in the target group (i.e., low 8th-grade numeracy score), (5) dummy for getting small-group instruction if treated and (6) 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. Cluster (school) robust standard errors in parentheses. Statistical significance: ** 5 percent level and * 10 percent level.

to have immigrant parents, and have higher average 8th-grade numeracy scores.

In Table 2, we investigate the similarity of the treatment and control schools. 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 rows which control variables we include.

We start by looking at the first-row specifications, controlling 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 who have non-missing test scores from grade eight and are not receiving special needs education, and thus are in the main sample. We find no such difference. In both treatment and control schools, we

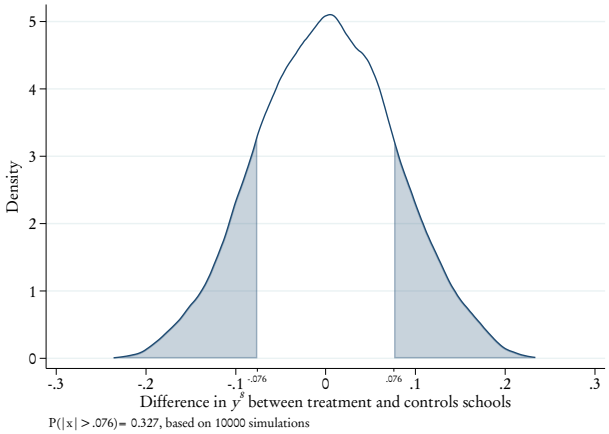
include just under 90 percent of students in the analyses (cf. outcome means in the bottom row of Table 2).

In the following columns, we investigate differences in student characteristics within the estimation sample and find significant differences. Column (2) shows the difference in an index of student and parental background, constructed as the predicted score on the numeracy test in grade 9. This index is about 10 percent of a (test score) standard deviation higher in treatment schools than in the control schools. The difference in measured score on the grade eight numeracy test in column (3) is slightly smaller and amounts to 7.6 percent of a standard deviation. As a result of the better prior performance of the students in the treatment schools, fewer students belong to the target group in treatment schools than in control schools. This difference amounts to 2.2 percentage points (column (4)) and can be compared to the sample average of 20 percent target students. Finally, columns (5) and (6) decompose the target students into those that would get small-group and large-group instruction if treated. For both treatments, 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 (column (3)). In the effect analyses, we will study several samples, corresponding to different treatments.

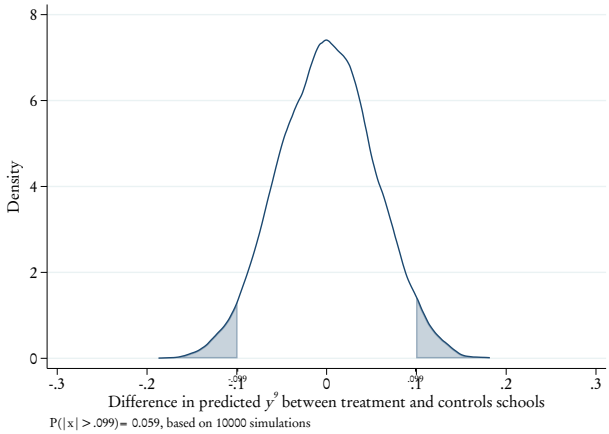
In Table A2 in Appendix C, we show differences in family background and 8th-grade numeracy score for the small-group, large-group, and non-target samples. In particular, in the small-group sample, we find substantial treatment-control differences in both family background and 8th-grade numeracy, with family background unable to explain the difference in numeracy. We will address this imbalance by adding different sets of pre-determined controls when analyzing the effects, primarily controls for 8th-grade numeracy.

Figure 2: Treatment-control differences across many randomizations

(a) 8th grade numeracy



(b) Predicted 9th grade numeracy



Note: Figures show the distributions of the treatment-control differences from 10,000 randomizations. The shaded areas indicate the share of randomizations with an absolute difference larger than the observed differences in the experiment, 0.076 and 0.099.

4.5 Investigating stratified randomization

With the random assignment of schools, we may be surprised to see significant (and substantial) 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 differently-sized schools and a heterogeneous student population (cf. Table A1) make differences like those we observe somewhat likely. In Figure 2, we present the distribution of differences in 8th-grade numeracy scores and predicted 9th-grade scores between treatment and control schools across 10,000 randomizations. Sub-figure (a) shows that we find absolute differences in 8th-grade numeracy as large or larger than those we observe in Table 2 in 33 percent of the randomizations, as indicated by the shaded areas. The difference in predicted 9th-grade numeracy is as big or bigger than what we observe in 5.9 percent of the randomizations (cf. sub-figure (b)).

We stratified 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 higher strata, as predicted. However, the relationship is not monotone. Many schools have lower average scores than other schools in higher strata. Also, while many strata have minor within-strata differences, in several strata, the differences are substantial. It is not entirely unexpected. Figure 1 is sorted by the number of target group students in 2015/16, such that a school retains its position in subsequent years. We see that the number of target students (and the share of target students in Figure A2) does not increase monotonously with rank in later years, while the number of target students correlates over the years, the ranking of schools does change.

Given the imperfect sorting of schools into strata, it is reasonable to ask if we could have done better regarding stratified and randomized schools. In Table A3 in Appendix C, we compare the performance of alternative stratification schemes. In addition to the stratification used for the randomization (numbered 1 in Table A3), we have investigated randomization without stratification (0), a two-year version of the implemented scheme (2), stratification based on one- (3) and two-year mean 8th-grade score (4), the number of target students (5), and the share of target students (6). The two-year schemes (i.e., 2 and 4) use data from 2014/15 and 2015/16, i.e., the most recent years available when randomizing. These different schemes produced similar but not identical stratifications of the schools. Looking at the correlation matrix for the different schemes, most correlations between schemes are close to or greater than .9, and the scheme used correlates more than .86 with all alternatives.

For each scheme, we stratify schools and randomize to treatment and control within strata 10,000 times. For each randomization, we find the treatment-control difference in the 8th-grade test score, controlling for strata dummies. The first column in Table A3 shows the share of randomizations that give an absolute student-weighted difference between treatment and control schools greater than the observed difference. We see that the stratification we used produced a difference in 33 percent of the randomizations. Without stratification, we get differences in 56 percent of the randomizations. However, most other stratification schemes

perform better than the one we implement. The only exception is the scheme where we use the number of target students, which produces differences in 35 percent of the randomizations. For the remaining, the share ranges from 8 to 22 percent.

The same pattern is visible for the mean absolute difference in the next column. With our chosen stratification, this is 6.1 percent of a standard deviation. With no stratification, the mean absolute difference is 10.4 percent, and for the other schemes, it ranges from 3.5 percent (when stratifying by average score) to 6.4 percent (when stratifying by the number of target students). While student heterogeneity between schools is the main reason for these differences, differences in school sizes also contribute. We see this by comparing the student-weighted differences in the second column with the unweighted school differences in the third. For the stratification used, the mean absolute unweighted difference is 5.1 percent of a standard deviation, almost 20 percent smaller than the weighted difference.

Table A3 demonstrates that stratifying partly by the number of target students produces larger average differences in test scores than if schools were stratified exclusively by average test scores. However, the stratification was also based on the number of target students because the intervention depends crucially on the number of target students and to ensure sufficiently many target students receiving instruction in large groups both in treatment and control schools. The fourth and fifth columns of Table A3 show the student-weighted and unweighted absolute mean difference in the number of target students between the treatment and control schools. With an unweighted mean absolute difference of 2.5 students, the stratification used is the second best-performing, beaten only by stratification by the number of target students. Randomization without stratification stands out with poor performance, as for average test scores. However, stratification by average test scores gives mean absolute differences in the number of target students of about 3, only moderately higher than the difference for the stratification used.

5 Results

In this section, we present our effect estimates. We first investigate the effects on target students receiving high-dosage tutoring, large-group instruction, and spill-overs to non-target students. We then study the treatments in the pilot year.

5.1 Effects of the main intervention

High-dosage tutoring

In Table 3, we report the results on student achievement of receiving instruction from trained teachers in small groups. Each cell represents a separate regression. We study different outcome variables (indicated by the columns) and include various control variables (indicated by the rows). As shown in section 4, we have found evidence of random differences between treatment and control schools. We will thus need to take pre-existing differences into account when estimating treatment effects.

We start by establishing (in column (1)) that the difference in test-taking across treatment and control schools is essentially zero, irrespective of controls. It 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.

In column (2), we present the effects on our main outcome variable, the 9th-grade test score (for the 89 percent of the students that took the 9th-grade test). Low-performing students receiving small-group instruction perform 0.12 SD better than the similar students in the control schools (cf. the top row, without controls). A large part of this difference is attributable to their more advantageous background. When conditioning on family controls in row two, the point estimate decreases to 0.10. In row three, we further add controls for prior achievement (8th-grade test scores) and obtain a statistically significant 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, we argue that this is a credible estimate of the intention-to-

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**
	(0.017)	(0.036)	(0.019)	(0.025)
Family controls	0.001	0.104**	-0.048**	-0.061**
	(0.015)	(0.032)	(0.018)	(0.021)
Family $+y^8$ controls	-0.003	0.060**	-0.035**	-0.028*
	(0.015)	(0.021)	(0.014)	(0.016)
Family $+y^5$ controls	0.004	0.104**	-0.048**	-0.060**
	(0.015)	(0.030)	(0.017)	(0.022)
N	1142	1015	1015	1015
N clusters	48	48	48	48
$\bar{y}$	0.889	-0.720	0.141	0.603

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 whether the student has a 9th-grade numeracy score, (2) 9th-grade numeracy score, (3) dummy for 9th-grade numeracy score at lowest proficiency level and (4) dummy for 9th-grade numeracy score at two lowest proficiency level. 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 and the fourth rows include (third-degree polynomials) 8th-grade or 5th-grade numeracy test score. The sample is target students predicted to get instruction in small groups in years 2017/2018 and 2018/2019 and corresponding students in control schools, and except for column (1) have a 9th-grade test score. Cluster (school) robust standard errors in parentheses. Statistical significance: ** 5 percent level and * 10 percent level.

treat effect of a target student predicted to get small-group instruction by trained teachers.[20] Assuming that the 2017/18 share of 89 percent of predicted small-group students getting such instruction is representative for both years and that there is no effect on the remaining 11 percent that do not receive small-group instruction, this corresponds to a treatment effect on the treated of about 0.067 SD for the students receiving small-group instruction by trained teachers. In columns (3) and (4), we study differences in the share of students performing at the lowest and either of the two lowest proficiency levels on the 9th-grade test. In line

[20]This effect is also significant in a wild bootstrap test ($p = .041$).

with the positive effect on test scores, we find a reduction of 3-4 percentage points in either measure of low-scoring students, with base levels of about 14 and 60 percent, corresponding to about 25 and 5 percent.[21]

Although test score in grade eight is the best proxy for prior performance and thus gives the lowest residual variance and the most precise estimates, it is potentially endogenous to the treatment. The test in 8th grade is conducted about 1.5 months into the school year, which is after the teacher training has started. In the last row, we substitute 8th-grade test scores with 5th-grade test scores, which are indeed pre-determined. We find a significant difference of 0.10 SD in favor of the treated students, very similar to the results where we only control for family background. While less vulnerable to endogeneity for the treatment, this specification takes less account of pre-existing random differences between the treatment and control schools. We will thus focus on the results conditional on 8th-grade scores as our main effect estimates.

Large-group instruction

Table 4 presents the effects on target students receiving instruction by trained teachers in larger groups. The set-up is identical to Table 3. As for small-group instruction, there are no large 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 and close to zero, particularly in the preferred specification where we control for the 8th-grade test score. Consistent with no impact on test scores, we find no effects on the share of low-performing students in columns (3) and (4). The confidence interval for effect on test scores is (-.07, 05).

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).

[21] A wild bootstrap test of the former effect is significant at the 10 percent level ($p = .059$) but not for the latter.

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
	(0.012)	(0.031)	(0.008)	(0.035)
Family controls	0.014	-0.041	0.015	0.029
	(0.012)	(0.032)	(0.010)	(0.036)
Family $+y^8$ controls	0.016	-0.010	0.006	0.005
	(0.013)	(0.029)	(0.011)	(0.034)
Family $+y^5$ controls	0.015	-0.035	0.014	0.025
	(0.011)	(0.027)	(0.010)	(0.035)
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. Cluster robust standard errors in parentheses. Statistical significance: ** 5 percent level and * 10 percent level.

Thus, a positive effect larger than .05-.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. Even though the confidence intervals for the impacts on test scores in tables 3 and 4 overlap, a formal t-test reject equality of effects ($p = .015$), while a wild bootstrap test only rejects equality at the 10 percent level ($p = .066$).[22]

Spillovers to non-target students

Table A4 in Appendix C reports results for non-target students, similar to those for target students. Concerning the main outcome variable, 9th-grade test scores, all estimates, regard-

[22]We estimate the model fully interacted with students belonging to the small-group sample on data for all target students to compare the effects. With robust standard errors, this is equivalent to separate regressions.

less of controls included, are close to zero. Although less precise, wild bootstrap estimates are qualitatively similar and rule out effects like the main effect in Table 3.[23] There is also no significant effect on the share of non-target students who perform at the lowest proficiency level on the 9th-grade test, but there is an increase in the share of non-target students on either of the two lowest levels. The latter should, however, be interpreted in light of test-taking: marginal test-takers are often low-performing. If a large share of the extra test-takers among the non-target students in the treatment schools (as is indicated in column (1) in Table A4) perform at the lowest two levels, this is sufficient to explain the difference in the share of low-performing non-target students.[24]

Overall, there is little indication of effects neither on non-targeted students or target students in large groups. Recall that non-target students often were mixed with target students randomized to large groups , suggesting that there were some changes regarding teacher training, class composition, and class size for these students, still the changes were probably not very large.[25] Previous studies have found no class size effects in middle schools in Norway (Leuven et al., 2008; Leuven and Løkken, 2020). Moreover, even if teachers instructing large groups participated in the teacher training program, the variation in the academic level of these adolescents (spanning from proficiency level 2-5, see Table A1 in appendix C) may have been too high for endorsing the newly learned didactic methods (Duflo et al., 2011). We discuss channels of impact, that is, teacher fidelity to the didactic principles and tools, in section 6.

[23]Wild bootstrap produces a confidence interval of (-0.038, 0.037). A t-test comparing the main effects on test scores of the small-group students and the non-target students gives a p-value of .005, while a wild bootstrap test gives a p-value of .044.

[24]A wild bootstrap test gives no significant effect (p-value = .180).

[25]Many schools reported that all non-target students received large-group instruction (cf. section 4.2). It may indicate that large-group instruction did not deviate much from ordinary classroom instruction.

Heterogeneous effects

Tables A5 and A6 in Appendix C, report effect estimates by student and school characteristics. Despite problems with balancing, there are indications of effect differences by student characteristics for the small group treatment. We find significant effects on boys and students with parents with higher education, while effects on girls and students with lower parental education are close to zero and insignificant. However, we cannot reject that the effects are the same. There are no clear differences by 8th-grade test score, immigration status, or cohort.

Although, we find indications of heterogeneous effects among large group students, we are reluctant to emphasize or interpret them. The number of students randomized to large groups is smaller than in small groups and distributed across fewer schools. As we find no indication of an average effect in Table 4, any significant estimate for a subgroup is likely spurious. The estimates for non-target students are all close to zero.

We find an effect in schools with higher average 8th-grade test scores and no impact in schools with lower average test scores. It is the only case where the effects for different school subgroups are significantly different. However, it does not point clearly to any mechanism. Schools with higher average test scores have fewer target students and a higher share of target students receiving small-group instruction than schools with lower average scores.

5.2 Treatments in the pilot year

Table A7 in Appendix C shows results based on the first year of the intervention for each combination of student group (target students randomized to small and large groups) and treatment (full treatment or funding only).

We see that the imbalance in pre-intervention characteristics notably regards the small-group students in schools implementing the full intervention. Adjusting for the difference in 8th-grade score yields substantial but imprecise negative effect estimates - for both groups of target students - and in particular for students in large groups. The estimates are significant at the 10 percent level. However, as the treated students belong to only eight schools, the cluster-

robust estimates may under-reject. Wild bootstrap confidence sets are wider. In particular for students in large groups, and insignificant both for small- and large-group students. Imbalance in pre-determined characteristics for the small-group students makes these estimates hard to interpret.[26] Similarly, the low number of students and the conflicting differences in background and 8th-grade scores make the estimates for large-group students also hard to interpret.

Schools only receiving funding are more similar to their control schools before the intervention, and thus the estimates for these schools are easier to interpret. We find an insignificant negative effect of 0.07 SD for small-group students and a negative effect of 0.08 SD, significant at the 10 percent level, for other target students. For each of the main effect estimates in column (4) of Table A7, a t-test rejects equality of effects with the main estimate from Table 3. Wild bootstrap tests only reject equality of the full treatment and the main intervention for small groups, and only at the 10 percent level. Taken at face value, the funding-only treatment suggests that small group instruction for low-performing students (i.e., ability grouping) without customized didactic methods is not sufficient to improve student achievement. However, recall that the group size in the first year is beyond what Fryer (2017) defines as high-dosage tutoring. And, even though schools got detailed instruction on how to spend the extra resources, including how to group students, we cannot guarantee the lack of discretionary adjustments.

6 Teacher fidelity to the didactic methods

For students to benefit, it is necessary that the teachers apply the targeted didactic methods they learned during their training. Fidelity, i.e., a high-quality implementation, means endorsing the didactic principle and tolls intended by the program.

During autumn 2017 and spring 2018, DPU collected data on fidelity through non-participative observations in randomly selected (treated) classrooms. In total, DPU observed 47 interven-

[26] 8th-grade test scores are strongly related to 9th-grade scores. If there is a difference in the 8th-grade score, a bias when controlling for the 8th-grade score can give a substantial relative bias in the estimated effect.

tion sessions, 35 in small groups and 12 in large groups. Each classroom observation followed one intervention session from start to end. DPU developed and used observation forms to assess the fidelity of implementing the six didactic principles and the four didactic tools.[27]

Overall, we find higher fidelity to the didactic principles and tools among small-group teachers than large-group teachers. We find that principle 2 (use low threshold-high ceiling tasks), primordial for differentiated instruction, hence suited 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 conducting differentiated instruction in a classroom of heterogeneous students is hard to achieve. It may explain the lack of impact on target students in large groups, as indicated in Table 4. Durlak et al. (2011) find that a lower-quality program implementation induces poorer student outcomes than a higher-quality or more complete implementation.

Given the existing findings in the literature, it is not surprising that small-group teachers show greater fidelity to the didactic methods of the intervention. The small groups consist of more homogeneous students, enabling teachers to concentrate their instruction on where students are academically and adjust to one academic level (Duflo et al., 2011; Guryan et al., 2021). That is, the teacher can personalize the instruction relative to classroom teaching in large groups. Teachers teaching small groups also got additional teaching materials and more detailed instruction plans, incorporating the didactic principles and tools. Research shows that programs are more effective when they are easy to follow. Together, ability grouping and teaching materials may have facilitated the instruction for small group teachers and thereby

[27]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 linking, (ii) the teacher states organizational link, (iii) students state organizational link, (iv) the 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 large-group sessions had teachers or students asserting mathematical links. A third of all large-group sessions did not explicate any links to neither former nor following sessions, which was only the case in very few small group sessions.

increased their fidelity to the didactic methods. Responses from teacher surveys (Kirkebøen et al., 2018) also point to more enthusiasm and satisfaction among small-group teachers.[28]

The didactic principles and tools are not unique to the intervention. Comparing survey responses from schools, we detect that the didactic methods are not unfamiliar to teachers in control schools (Kirkebøen et al., 2018; see also Appendix B.2). However, when asked about using the different didactic methods, there is a clear difference between treatment and control schools with principles and tolls endorsed 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 identify 2211 teachers working in control or treatment schools in October 2017.[29] When inquiring about where they worked in March 2019, we find that only 18 out of 1115 teachers working in treatment schools moved to control schools. We cannot identify teachers receiving training in the linked data, but the low number suggests that direct contamination through job changes is not a big issue. Note, there are few job changes between sample schools, but not low mobility in general. 19 percent of the October 2017 teachers are not working at the same school in March 2019.[30]

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

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

[30]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 we identify from the register 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 workplace.

7 Costs and benefits

In section 5, we found a significant intention-to-treat (ITT) effect of the small-group intervention of .06 SD, corresponding to an average treatment effect on the treated (ATT) of about .067 SD. In this section, we discuss how we can value this effect and how this value compares to the cost of the intervention. As we found no effect of large-group instruction, we only focus on the small groups.

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. To inform policy, we care about the cost of continuing or introducing the small-group intervention, which will require funding for small-group instruction and some administrative overhead.[31] 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 1000 USD and an ATT effect of 0.037-0.056 SD per 1000 USD. These effects are slightly lower than what Guryan et al. (2021) find in their study.

We use Kirkebøen (2021) - who studied the effect of school quality on long-term student outcomes - to value this effect. A 0.06 SD effect on numeracy early in lower secondary can be expected to increase end-of-compulsory school grades by 0.04 SD, high school completion rates by 0.6 percentage points, and earnings by 0.5 percent, or USD 265 per year. This is similar or slightly lower than the valuation of test score effects in Guryan et al. (2021), based on Chetty et al. (2011).[32] With about 600 students receiving effective small-group instruction during the

[31]The total costs of the intervention (not including the research) were around USD 1.7M. The intervention included extra administrative resources for UDE to communicate with and provide data to the researchers. Furthermore, some administrative costs pertain to the large groups. For small-group intervention only, the total administrative will be lower than in the current intervention. The per-student administrative cost will, however, depend on the scale of the intervention. Some of the costs are likely to be one-time costs, e.g., costs of teaching materials.

[32]Kirkebøen (2021) finds that 0.1 SD higher school value-added on 8th-grade test scores increases 10th-grade exam scores by about 0.067 SD and the share completing high school by about 1 percentage point. 0.1 SD difference in exam value-added is associated with 1.7

main intervention years, this corresponds to 3-4 more students completing high school. Even if 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 costs per student, and three times the cost of small-group instruction.[33] 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 to help adolescents falling behind? This paper adds to the burgeoning literature on this essential topic by designing and evaluating a high-dosage tutoring intervention, aiming to improve the performance of low-performing 8th graders in mathematics. While it is too early to conclude about longer-term effects (e.g., completion of upper secondary education, the main objective), the short-term impact is promising.

A majority of the target students in our sample were randomized to small groups consisting of a maximum of six students where they received customized instruction by newly-trained teachers. The findings indicate that these students increased their average test scores by about 6 percent of a standard deviation in the year following the intervention. The share of low-

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.

[33]Falch et al. (2009) estimate the social return to completing high 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 high 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 USD 258k and the social return to USD 756k.

performing students is reduced by about 3 percentage points, corresponding to a reduction of 5-25 percent for different measures of low performance. The intervention is cost-effective with an estimated cost per small-group student of USD 1200-1800 and estimated benefits of USD 3700. It is a similar effect per dollar and similar cost-benefit ratios as Guryan et al. (2021), but with a different intervention and context.

Some target students were randomized to large groups, mainly their regular classes, instructed by newly-trained teachers. We find no impact on these students. Although teachers in large groups applied the didactic principles and tools promoted by the intervention more than teachers in the control schools, classroom observations and teacher surveys suggest low teacher fidelity to the didactic methods in these groups. Low-quality implementation of the didactic methods in large groups may be due to different reasons. First of all, larger groups may require more time and effort spent on classroom management, leaving less opportunity for conscious changes of the teacher's practices. Second, the larger groups were more heterogeneous than the small groups and, therefore more challenging to manage due to differentiated instruction being necessary. Finally, unlike small-group teachers, teachers in large groups did not get any concrete lesson plans and teaching materials to facilitate the use of the promoted principles and tools.

The results from the funding-only treatment point indicate zero impact, suggesting that small-group instruction and ability grouping is not sufficient to increase performance among low-performing adolescents if no better pedagogy is involved. However, these results need to be interpreted with caution as group size is not directly comparable across the pilot year and the remaining two years (25 percent larger in the pilot year).

Nevertheless, even though we cannot separate the effect of the different elements in our intervention, we have sufficient evidence to conclude that ability grouping and a detailed instruction plan in small groups seem to be salient channels of impact. Together with our uplifting cost-benefit estimates, we, therefore, advocate implementing the small group intervention in settings where the purpose is to ameliorate the math skills among low-performing

adolescents.

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

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

[10] 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.

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

[12] 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.

[13] 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.

[14] 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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

[22] Falch, T., Johannessen, A. B., and Strøm, B. (2009). Kostnader av frafall i videregående opplæring. SØF-rapport 08/09.

[23] 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.

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

[25] 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.

[26] 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.

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

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

[29] 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.

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

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

[32] 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.

[33] 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.

[34] 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.

[35] Kirkebøen, L. (2021). School value-added and long-term student outcomes. Unpublished manuscript.

[36] 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.

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

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

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

[40] 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.

[41] 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.

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

[43] 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.

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

[45] 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.

[46] 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.

[47] 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.

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

[49] 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.

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

[51] 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.

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

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, like Japan. Japanese classrooms were characterized by teachers clarifying to students how the content and working methods of one session relate to previous sessions and possibly also to 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 on whether students perform at a low, medium, or high level - endorse teaching practices consisting of rich assignments, formative assessment, and high expectations of students. Forgasz (2010), on the other hand, finds benefits of tracking for high-performing students but disadvantages for low-performing students. Although teachers allow high-achievers to engage in mathematical challenges, they offer simple math for low-performing students, restricting their learning opportunities. Endorsing activities where all students have sufficient prerequisites to get started (low entry threshold) and continue the activities in more complex variants as far as the situation allows (high ceiling) is an essential principle of differentiated instruction. For the small group teachers, Principle 2 is embedded in the formulated tasks in the specific detailed teaching material. In the large groups or regular classes with target students included, the teachers should themselves create new or adapt existing ones into low

threshold-high ceiling tasks to let all students engage in real mathematical challenges from a safe starting point to as much as they can master. The teacher training sessions in years 2 and 3 presented ideas and examples of how to adapt to existing tasks.

Principle 3: Initiate 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 take a growing focus in mathematics education research. Some scholars focus on affection as a cause for cognition, while others focus on the other way round. The relationship also goes beyond causes and effects: Cognition is affective, and affection is cognition. These aspects are interwoven. A feeling is a belief of what mathematics is and why mathematics is essential to master, which is part of students' motivation. At the teacher training sessions, the interwovenness was presented with the abbreviation MO-FORMANCE from MOtivation-perFORMANCE. Principle 3 states that it is part of the teachers' responsibility to support the students' motivational and cognitive development in mathematics.

Principle 4: Initiate conversations with and among students on mathematical processes and concepts to support mathematical understanding

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

Principle 5: Set realistic, but high expectations

The rationale for encouraging teachers to set realistic, but high expectations for students, is partly based on research on student ability tracking and partly based on teacher perceptions of students with mathematics difficulties. Scherer et al. (2017) discuss causes of mathematics difficulties, where one extreme is to consider difficulties as errors by the individual (neurological or psychological), and the other extreme is that they arise due to failures in the educational system (didactic) or other social features (sociological). Scherer et al. (2017) argue that teachers relating to neurological and psychological theories see these students as being in deficit, cumbersome, and unable to learn. Those with more relational views on math difficulties think 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 inspiration comes from the mathematics education literature on equity. This literature enlarges the opportunities-to-learn concept to include access to mathematical content and discourse practices, as well as positional identities (Esmonde, 2011), which underline teachers' expectations of students' potentials as decisive for effect. Principle 5 focuses on the importance of avoiding setting too low expectations for low achieving students - as it might delimit their learning opportunities.

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

Writing a journal (i.e., logbook) is recommended to activate students' reflections. Students follow several different subjects during a school day, rushing from, for instance, history class to mathematics class, and further on to a foreign language class, which may challenge students' 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 remembering what they have learned. As underlined by Bligh (2000), teachers need to guide their students. Bligh recommends advising students to document their experienced questions

and problems immediately, or else they will forget (p.145). Students' notes are valuable for their teachers, too. These notes allow teachers to get a sense of what students have learned and what confuses students. Instant feedback from students, for instance, few minutes prompted writing at the end of a session, can help teachers adjust the following sessions to students' needs (Center for Excellence, 2021).

B Evidence based on the autumn 2016-survey

In Appendix B, we provide more details on the pilot year (B.1), and on compliance, group sizes, and organization of small-group instruction in treatment and control schools (B.2).

B.1 The pilot year

In the survey following the first intervention period in 2016, only about 25 percent of the pilot school teachers answer that they had received sufficient information, and 35 percent that the teacher training 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 what it meant that the intervention was part of a research project. Schools and teachers, in general, accepted the group assignment rules. However, some were unsure about implementing the didactic methods and to what extent they were allowed to use their professional judgment. UDE, previously unfamiliar with implementing experiments, was not always able to provide clear answers. Partly, it reflected a hands-off approach from the researchers and a desire from UDE for the intervention to be like a standard intervention.[34] After the initial months, lesson plans and other materials were ready. By the second year, 75 percent of the teachers answered having sufficient information and 70 percent enough training.

The schools receiving funding but no teacher training got explicit instructions on group sizes and which students to include in the small and large groups, but the didactic methods were for the schools and teachers to decide. Reports from these schools suggest nothing but

[34]With experimental interventions, there is always a question of to what extent we can extrapolate the effects 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 present the project and avoid confusion that would not be present in a non-experimental setting.

satisfaction with the extra funding. Still, we cannot rule out that there might have been issues associated with the fact that this was indeed a pilot year. The size of the small groups in the pilot year was beyond Fryer's (2017) definition of high-dosage tutoring (see footnote 2).

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

After the 2016 intervention period, we surveyed teachers in full intervention, funding-only, and control schools. Several teachers in each school type could participate, and we got at least one answer from almost every school. Among the questions asked were 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 report using small-group instruction. In Figure A5 in Appendix C, we display students and the number of groups in the funding-only and control schools.[35] 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 groups of more than 12 students, 70 percent of students receiving only funding 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, mostly being made up of groups of 9-12 students. It is hard to know to what extent the data from the control schools are accurate and complete. However, as we have responses from all control schools (either group sizes or an answer that they do not use small-group instruction), the number may be representative. Thus the intervention approximately doubled the number of students receiving small-group

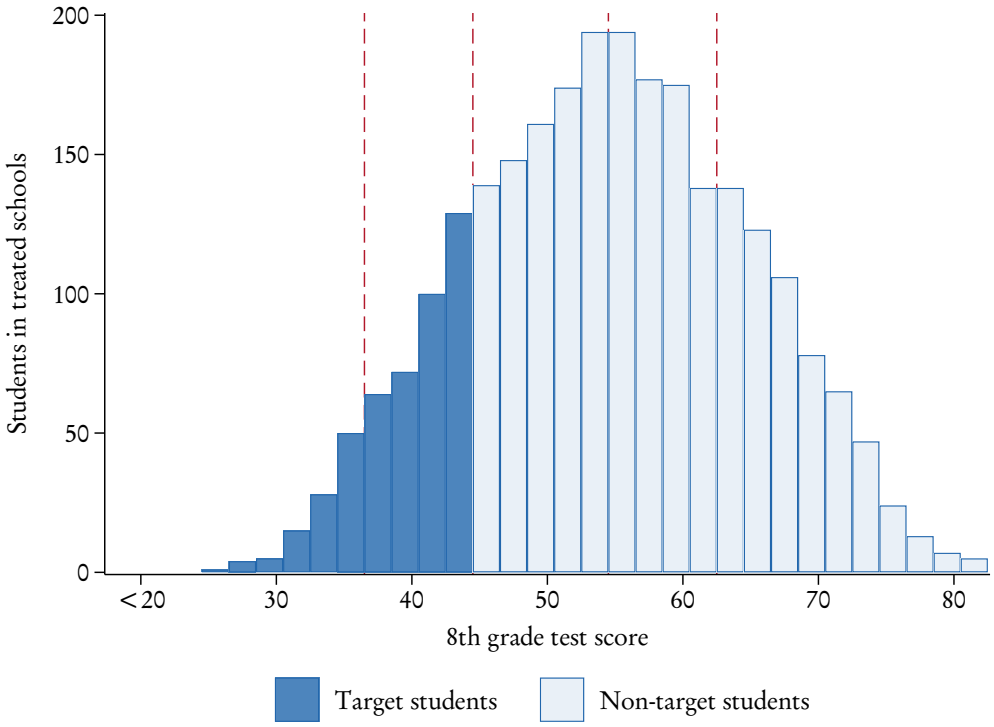
[35]We got survey responses from teachers in 15 out of 16 funding-only schools and all 24 control schools. 28 teachers from 14 funding-only schools and 24 teachers from 14 control schools reported using small-group instruction. The reports from teachers at funding-only schools were typically approximately 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.

instruction (cf. the 299 students that got small-group instruction in the treated schools in 2017/18) during the intervention period.

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

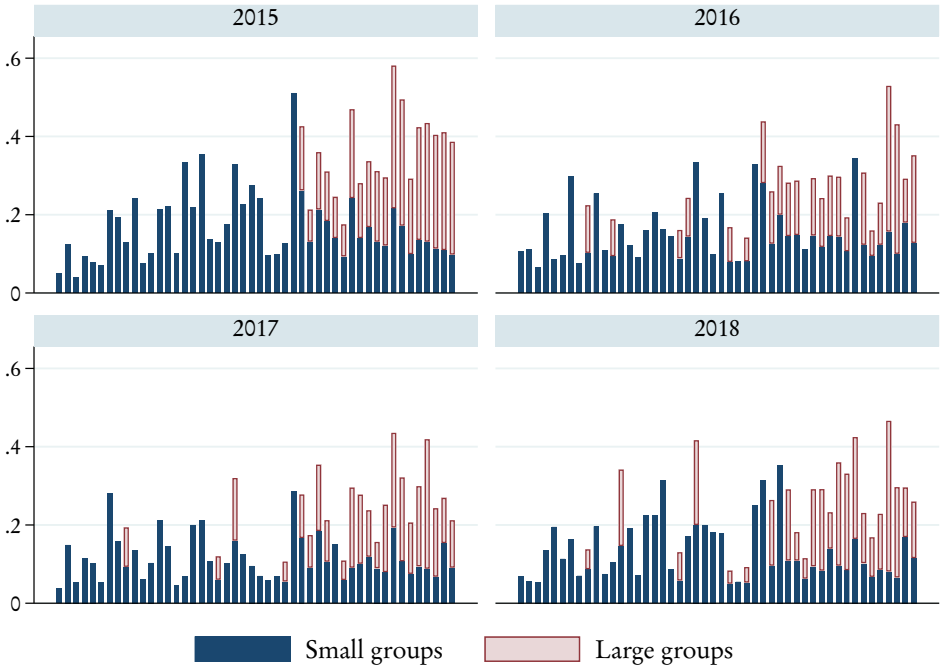
C Supplementary figures and tables

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



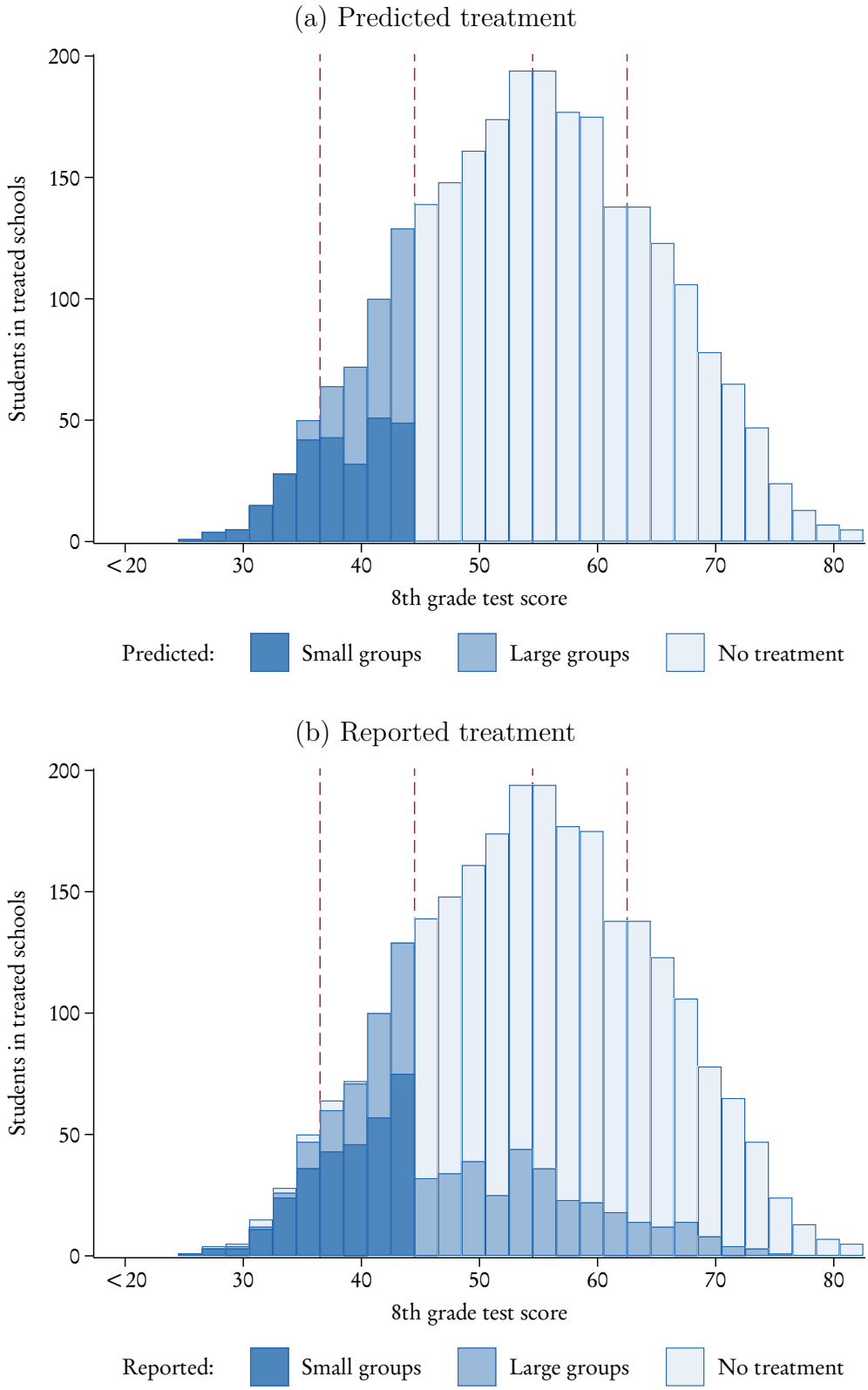
Note: Test scores have a national average of 50 and a standard deviation of 10. Vertical lines separate proficiency levels 1-5: Level 1 is test score $\leq$36, level 2 is test score $\in$ [37, 44] etc.

Figure A2: Share of target students per school and year



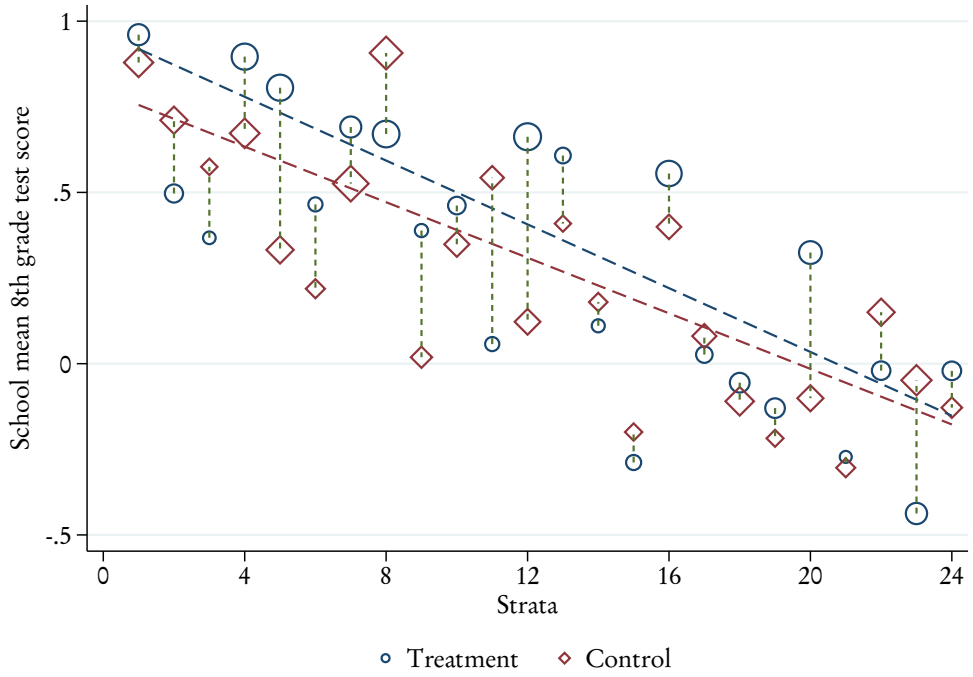
Note: Each bar represents the share of target students in one school and year. The bars distinguish between the share of target students predicted to get small and large group instruction if the school participates in the intervention. In 2015 (the year used as the basis for stratifying schools) and 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 is as reported by the schools.

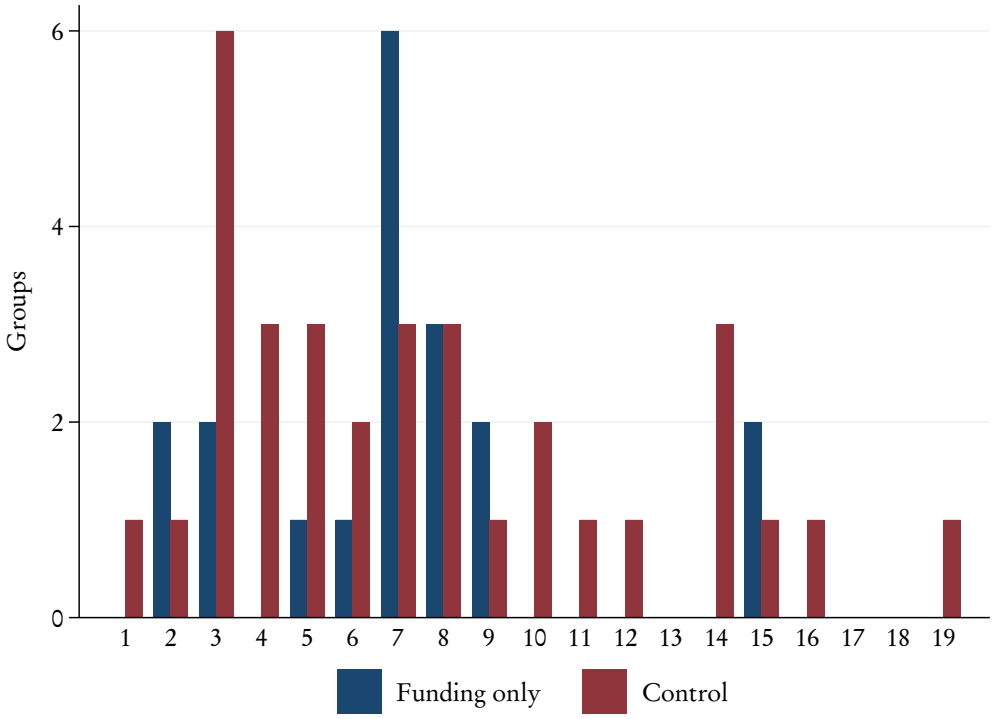
Figure A4: Average 8th-grade score by strata and treatment status



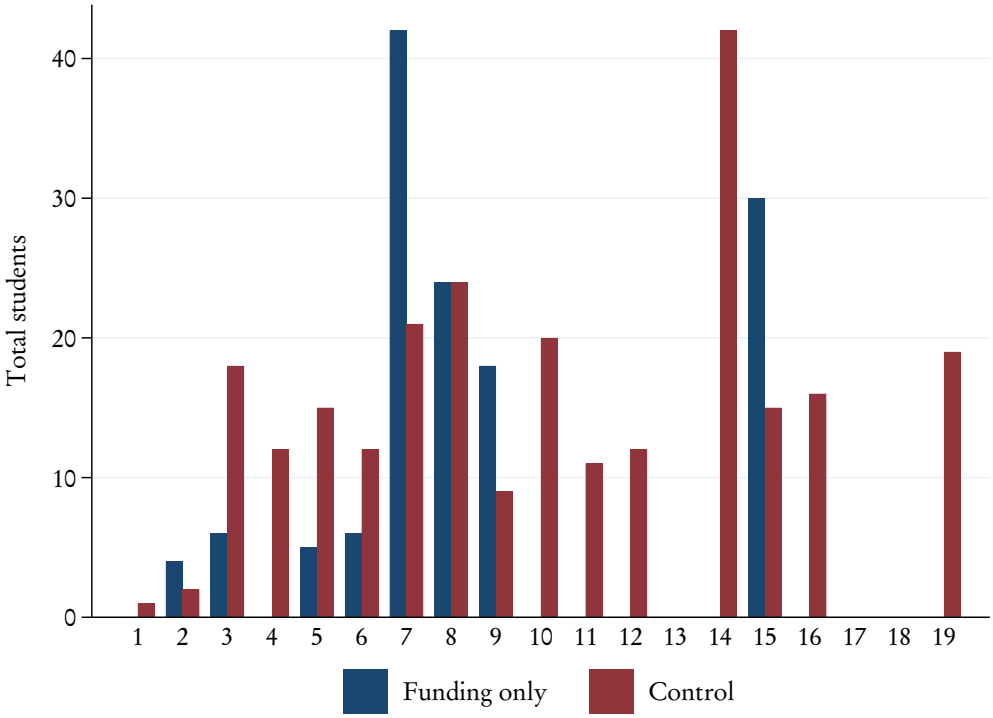
Note: The figure show school average 8th-grade math score in the main estimation sample (i.e., 2017/18 and 2018/19 students) by treatment status and strata. Larger markers indicate more students.

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 A1: 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 data (from 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 data (from the compulsory school register, 2017/18)								
Share combined primary school	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)
Spec. needs teach./reg. teach.	0.25	(0.18)	0.21	(0.15)	-0.04	(0.05)	0.02	(0.04)
Norwegian for immigrants teach./reg. teach.	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 (from matched 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 (hrs./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 A2: Balancing by student group, 2017/18 and 2018/19

	(1)	(2)	(3)	(4)	(5)	(6)
	Small-group students		Large-group students		Non-target students	
	Family index ($\hat{y^9}$)	8th grade score y^8	Family index ($\hat{y^9}$)	8th grade score y^8	Family index ($\hat{y^9}$)	8th grade score y^8
Effect estimates from specification with						
No controls	0.080*	0.075**	0.026	-0.027	0.092**	0.037
	(0.040)	(0.034)	(0.024)	(0.019)	(0.036)	(0.033)
Family controls		0.058*		-0.032*		-0.005
		(0.031)		(0.018)		(0.023)
N	1015	1015	760	760	7597	7597
N clusters	48	48	25	25	48	48
$\bar{y}$	0.220	-1.217	0.122	-0.860	0.710	0.729

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 3 for details. Cluster robust standard errors in parentheses. Statistical significance: ** 5 percent level and * 10 percent level.

Table A3: 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 A4: 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
	(0.005)	(0.035)	(0.000)	(0.003)
Family controls	0.009**	-0.008	0.001	0.007**
	(0.005)	(0.022)	(0.001)	(0.003)
Family $+y^8$ controls	0.009**	-0.001	0.001	0.006*
	(0.005)	(0.013)	(0.001)	(0.003)
Family $+y^5$ controls	0.010**	0.000	0.001	0.005*
	(0.004)	(0.018)	(0.001)	(0.003)
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. Cluster robust standard errors in parentheses. Statistical significance: ** 5 percent level and * 10 percent level.

Table A5: Heterogeneous treatment effects, student characteristics

	(1) Family index ($\hat{y^9}$)	(2) 8th grade score y^8	(3) Dummy has y^9	(4) 9th grade score (y^9)	(5) Lowest proficiency (D^{L1})	(6) Low proficiency (D^{L2})
Student's sex						
Male	0.046	-0.020	0.166**	0.119**	-0.062**	-0.073**
	(0.040)	(0.023)	(0.063)	(0.046)	(0.028)	(0.034)
Female	0.095**	0.016	0.091**	0.019	-0.017	0.002
	(0.039)	(0.021)	(0.039)	(0.035)	(0.020)	(0.026)
8th grade proficiency						
Level 2	0.064**	0.009	0.120**	0.056*	-0.008	-0.033
	(0.024)	(0.019)	(0.038)	(0.033)	(0.022)	(0.024)
Level 1	0.004	-0.021	0.054	0.066	-0.080**	-0.021
	(0.033)	(0.034)	(0.068)	(0.054)	(0.038)	(0.033)
Parental education						
No higher education	0.045	0.024	0.062	0.031	-0.028	-0.004
	(0.042)	(0.021)	(0.046)	(0.031)	(0.021)	(0.026)
Higher education	0.105**	-0.049*	0.193**	0.108**	-0.047	-0.070*
	(0.042)	(0.029)	(0.053)	(0.051)	(0.032)	(0.039)
Immigrant background						
Native	0.117**	-0.002	0.193**	0.094**	-0.060**	-0.026
	(0.029)	(0.028)	(0.051)	(0.044)	(0.025)	(0.036)
Immigrant	0.034	0.043	0.095	0.082	0.018	-0.044
	(0.074)	(0.056)	(0.121)	(0.108)	(0.069)	(0.066)
Second gen.	0.019	-0.008	0.035	0.015	-0.025	-0.026
	(0.050)	(0.023)	(0.050)	(0.042)	(0.028)	(0.035)
Cohort (year of 8th grade test)						
2017	0.081*	0.025	0.136**	0.061	-0.043*	-0.043
	(0.045)	(0.025)	(0.062)	(0.046)	(0.024)	(0.036)
2018	0.067	-0.021	0.108*	0.058	-0.028	-0.015
	(0.052)	(0.028)	(0.061)	(0.047)	(0.025)	(0.035)
Controls for background and y^8				Yes	Yes	Yes

Note: Each cell gives an estimate of θ from equation (1) for a given outcome (column) and sub-sample (row). See note to Table 3 for details. Heterogeneous effects are estimated from fully interacted models (corresponding to separate regressions for each sub-sample). The sample is target students predicted to get in instruction in small groups in years 2017/2018 and 2018/2019. Cluster robust standard errors in parentheses. Statistical significance: ** 5 percent level and * 10 percent level.

Table A6: Heterogeneous treatment effects, school characteristics

	(1) Family index ($\hat{y^9}$)	(2) 8th grade score y^8	(3) Dummy has y^9	(4) 9th grade score (y^9)	(5) Lowest proficiency (D^{L1})	(6) Low proficiency (D^{L2})
Wave school among start in 2016 or start in 2017						
Start 2016 (pilot year)	0.187**	-0.019	0.237**	0.083**	-0.061**	-0.066**
	(0.050)	(0.034)	(0.065)	(0.030)	(0.022)	(0.013)
Start 2017	0.022	0.011	0.067*	0.049*	-0.024	-0.011
	(0.040)	(0.020)	(0.040)	(0.028)	(0.017)	(0.022)
Average 8th grade test score						
Lower	-0.064	-0.017	-0.072	-0.018	0.011	-0.019
	(0.058)	(0.020)	(0.053)	(0.035)	(0.019)	(0.037)
Higher	0.152**	0.019	0.239**	0.113**	-0.067**	-0.033
	(0.034)	(0.023)	(0.043)	(0.038)	(0.020)	(0.029)
Controls for background and y^8				Yes	Yes	Yes

Note: Each cell gives an estimate of θ from equation (1) for a given outcome (column) and sub-sample (row). See note to Table 3 for details. Heterogeneous effects are estimated from fully interacted models (corresponding to separate regressions for each sub-sample). The sample is target students predicted to get instruction in small groups in years 2017/2018 and 2018/2019. Cluster robust standard errors in parentheses. Statistical significance: ** 5 percent level and * 10 percent level.

Table A7: Effects of treatments in the pilot year for different student groups, 2016/17 students

	(1) Family index ($\hat{y^9}$)	(2) 8th grade score y^8	(3) Dummy has y^9	(4) 9th grade score (y^9)	(5) Lowest proficiency (D^{L1})	(6) Low proficiency (D^{L2})
Treatment: Teacher training and funding						
Predicted small groups	0.221**	0.244**	0.041	-0.098*	0.017	0.047
	(0.063)	(0.063)	(0.034)	(0.057)	(0.037)	(0.036)
Predicted large groups	0.104	0.039	-0.023**	-0.183*	0.077*	0.160
	(0.090)	(0.025)	(0.007)	(0.109)	(0.041)	(0.110)
Non-target students	0.185**	0.110	0.021**	-0.011	0.001	0.014**
	(0.033)	(0.067)	(0.007)	(0.038)	(0.002)	(0.006)
Treatment: Funding-only						
Predicted small groups	0.022	-0.038	0.049	-0.065	-0.038**	0.024
	(0.038)	(0.043)	(0.032)	(0.052)	(0.016)	(0.039)
Predicted large groups	0.048	-0.021	-0.030	-0.083*	0.038**	-0.016
	(0.031)	(0.042)	(0.025)	(0.044)	(0.015)	(0.040)
Non-target students	0.020	-0.034	-0.009	0.005	-0.000	0.011*
	(0.053)	(0.041)	(0.006)	(0.023)	(0.000)	(0.006)
Family and y^8 controls	No	No	No	Yes	Yes	Yes
N	4955	4955	5261	4955	4955	4955
N clusters	48	48	48	48	48	48
$\bar{y}$	0.652	0.324	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 (rows). Sample is all students in 2015/2016, with separate specifications for each group of students, similar to Tables 3, 4 and A4. 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. Cluster robust standard errors in parentheses. Statistical significance: ** 5 percent level and * 10 percent level.