ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 24-931

Sustained Effects of Small-Group Instruction in Mathematics

Henning Finseraas
Norwegian University of Science and Technology

Ole Henning Nyhus
Norwegian University of Science and Technology

Kari Vea Salvanes
Institute for Social Research

Astrid Marie Jorde Sandsør
University of Oslo

Recent research suggests that using additional teachers to provide small-group instruction or tutoring substantially improves student learning. However, treatment effects on test scores can fade over time, and less is known about the lasting effects of such interventions. We leverage data from a Norwegian large-scale field experiment to examine the effects of small-group instruction in mathematics for students aged 7-9. This intervention shares many features with other high-impact tutoring programs, with some notable exceptions: instruction time was kept fixed, it had a lower dosage, and it targeted students of all ability levels. The latter allows us to assess fadeout across the ability distribution. Previous research on this intervention finds positive short-run effects. This paper shows that about 60% of the effect persists 3.5 years later. The effect size and degree of fadeout are surprisingly similar across the ability distribution. The study demonstrates that small-group instruction in mathematics successfully targets student performance and that effects can be sustained over time.

VERSION: March 2024

Suggested citation: Finseraas, Henning, Ole Henning Nyhus, Kari Vea Salvanes, and Astrid Marie Jorde Sandsør. (2024). Sustained Effects of Small-Group Instruction in Mathematics. (EdWorkingPaper: 24-931). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/d5dn-c793

Sustained Effects of Small-Group Instruction in Mathematics*

Henning Finseraas
Ole Henning Nyhus
Kari Vea Salvanes
Astrid Marie Jorde Sandsør†

Abstract

Recent research suggests that using additional teachers to provide small-group instruction or tutoring substantially improves student learning. However, treatment effects on test scores can fade over time, and less is known about the lasting effects of such interventions. We leverage data from a Norwegian large-scale field experiment to examine the effects of small-group instruction in mathematics for students aged 7-9. This intervention shares many features with other high-impact tutoring programs, with some notable exceptions: instruction time was kept fixed, it had a lower dosage, and it targeted students of all ability levels. The latter allows us to assess fadeout across the ability distribution. Previous research on this intervention finds positive short-run effects. This paper shows that about 60% of the effect persists 3.5 years later. The effect size and degree of fadeout are surprisingly similar across the ability distribution. The study demonstrates that small-group instruction in mathematics successfully targets student performance and that effects can be sustained over time.

Keywords: small-group instruction, tutoring, sustained effects, RCT, teacher density

JEL Codes: C93 (Field Experiments); H52 (Government Expenditures and Education); I21 (Analysis of Education)

* We thank Hans Bonesrønning, Inés Hardoy, Jon Marius Vaag Iversen, Vibeke Opheim, and Pål Schøne for collaboration on implementing the intervention. We also thank participants at the 49th annual AEFP conference in Baltimore 2024, the LEER 2023 conference at KU Leuven, the Workshop on Education Economics and Policy in Trondheim 2023 for comments on an earlier draft, and Gustav Agneman, Colin P. Green, and Susanna Loeb for many valuable suggestions. This research is part of the 1+1 Project, supported by the Norwegian Research Council (grant number 256217). The 1+1 Project was pre-registered at OSF prior to obtaining data (Bonesrønning et al., 2018), and the current study is a follow-up of Bonesrønning et al. (2022), who studied effects on national tests in grade 5. The preparation of this manuscript was partly supported by funding from the European Research Council Consolidator Grant ERC-CoG-2018 EQOP (grant number 818425) and by the Research Council of Norway through its Centres of Excellence scheme, CREATE - Center for Research on Equality in Education (project number 331640).
† Henning Finseraas (henning.finseraas@ntnu.no): Norwegian University of Science and Technology (NTNU); Ole Henning Nyhus (ole.nyhus@samforsk.no): NTNU Social Research; Kari Vea Salvanes (k.v.salvanes@samfunnsforskning.no): Institute for Social Research (ISF) & Nordic Institute for Studies in Innovation, Research and Education (NIFU), Astrid Marie Jorde Sandsør (a.m.j.sandsor@isp.uio.no): University of Oslo. Shared first authorship with authors listed in alphabetical order.

Introduction

Using additional teachers to reduce class size is a popular policy, argued to improve student performance as it increases the time teachers can focus on individual students. However, the evidence of such effects is mixed, and there is a lack of knowledge about under which circumstances class size matters (see Leuven and Oosterbeek (2018), Schanzenbach (2020), and Green & Iversen (2022) for recent reviews). In the resource-rich Norwegian context, research has consistently shown no effects of additional teachers, although a few studies document positive effects for some subgroups of students.[1] This raises the question of whether the effect of additional teachers on student outcomes varies depending on *how* additional teachers are used.

In a recent large-scale field experiment in Norway, covering 159 schools over four years, Bonesrønning et al. (2022) find that providing small-group instruction in mathematics for students aged 7-9 increases student performance. Small-group instruction is more flexible and potentially less costly than reducing class size, allowing schools to target subjects or students needing additional support. The result is consistent with findings from the literature where tutoring (either one-to-one or in small groups) has been shown to improve student learning substantially (Dietrichson et al., 2017; Nickow et al., 2020; Neitzel et al., 2022; Gersten et al., 2020) and has emerged as a promising strategy for addressing learning loss related to Covid-19.

The intervention used an additional teacher to provide small-group instruction in parallel to regular classroom instruction using a pull-out strategy. Students received two 4–6-week periods of small-group instruction per school year in groups of 4-6. Each group returned to regular instruction once their tutoring period was over, and small-group instructors were expected to cover the same topics as the regular class. Students of all ability levels were provided with small-group instruction, primarily in groups with similar ability levels.

While the intervention shares many features with other successful tutoring programs, it differs along three key dimensions: 1) the instruction time was unaltered, 2) it had a somewhat lower dosage than other programs,[2] and 3) students of all ability levels were targeted. As the curriculum was unchanged, the focus of the intervention was on the organizational aspect of how to use an additional teacher.

The initial results document that the intervention increased student test scores by approximately 6% of a standard deviation, as measured on national tests in grade 5, with consistent

[1] While Leuven et al. (2008), Falch et al. (2017), Leuven & Løkken (2018), and Borgen et al. (2022) show no effects of additional teachers, Iversen & Bonesrønning (2013), Haaland et al. (2024), and Kirkebøen et al. (2021) find positive effects for some subgroups of students.
[2] According to Nickow et al. (2020), most tutoring programs last for between 10 weeks and one school year.

effects across the ability distribution. The beneficial effects are based on test scores from five months after the program ended, which is also the case for most other tutoring interventions. However, the potential benefits of programs do not necessarily stem from their immediate effects but from the program helping students in the long run. Whether short-term effects persist over time and, if so, for whom is an essential part of determining their value. While many rigorously evaluated educational interventions find promising effects shortly after the intervention ends, the few studies investigating long-run effects on school outcomes often find that effects substantially diminish or completely fade out (Bailey et al., 2020; Hart et al., 2023).[3] However, there are some studies documenting re-emerging effects in adulthood (see e.g., Ludwig and Miller, 2007; Deming, 2009; Pages et al., 2020; Bailey et al., 2021; Dodge et al., 2015; Chetty et al., 2011), where improvements in socio-emotional skills are suggested as a potential channel.[4] Even with a vast literature documenting short-run effects of tutoring and small-group instruction, we have not yet seen a paper examining whether these effects persist in the longer run.

Bailey et al. (2017) argue that sustained effects of interventions can be achieved through developing relevant skills at the right time and creating a sustained environment for skill-building. This is also in line with the human capital production model developed by Cunha and Heckman (2007), where early human capital investment is complementary to later investments so that later investments leverage early investments. Moreover, the fadeout process may differ depending on the student's ability level or background. For instance, low-performers in the control group could receive extra support, thereby catching up to the intervention group. This process is particularly likely to happen in environments where lower-ability students are prioritized (Bailey et al., 2017).[5] Investigating the heterogeneity of effects in the short and longer run gives us an understanding of whether fadeout is particular to some contexts or groups.

In this paper, we extend on Bonesrønning et al. (2022) and ask whether the initial beneficial effects of the tutoring intervention are sustained over time. We study the effect of the intervention on national test scores in grade 8, conducted 3.5 years after the intervention ended. This allows us to investigate whether effects persist and whether the degree of persistence varies by baseline ability level in the longer run. Furthermore, we investigate whether the intervention generated

[3] Bailey et al. (2020) and Andersen et al. (2022) both conclude that fadeout is a substantive phenomenon, not merely a statistical artifact, although Andersen et al. (2022) argue that a more nuanced view of fadeout is warranted due to the concave nature of skill-development.

[4] Heckman et al. (2013) show that this appears to be the case in the Perry preschool project. However, a recent meta-analysis by Hart et al. (2023) shows that socio-emotional and cognitive skills demonstrate similar patterns of fadeout, suggesting that there may be other mediators for long-run adult outcomes.

[5] An early empirical example often used to illustrate the importance of sustained environments is Currie and Thomas (2000), who investigate the reasons behind the observed quicker fadeout of effects among Black Head Start participants compared to white participants. Their results suggest that this discrepancy is due to Black Head Start children being more likely to subsequently attend inferior schools.

beneficial spill-over effects on language skills by evaluating the observed impact on national tests in Norwegian and English.

Our study provides one of the first estimates of the long-run effects of a tutoring intervention and is one of the first studies to examine the longer-run effects of education programs more generally. Identifying the longer-term effects of school interventions can be difficult because teachers and schools respond to students' achievement levels, often providing more support for lower-performing students. However, with randomization taking place at the school level in a context where students typically attend the same neighborhood school over time, the same teacher will interact with either students in the treatment or control group. Such an environment is likely to be less prone to biases due to educators adjusting to ensure that students in the control group catch up. Combined with the finding that all children benefited from small-group instruction, our trial provides a unique opportunity to investigate whether this consistent and sustained environment can lead to persistent effects of an educational intervention.

We find measurable effects of the intervention 3.5 years after the intervention ended, with students of all ability levels still benefiting. While we observe fadeout, with about 60% of the effect persisting 3.5 years later, this pattern is similar for all students. This suggests that the initial intervention was able to target relevant skills in an environment that allows for this skill development to be sustained over time. Our results also provide evidence that targeting early mathematics skills creates beneficial spillovers to language skills. This result is in line with two potential mechanisms, either that their mathematics skills improved their further skill development directly or that the intervention itself enhanced other skills, such as socio-emotional skills, that are beneficial for later skill development across domains.

Institutional background and the 1+1 project

The Norwegian school system

Compulsory education consists of ten years of schooling, divided into lower primary school (grades 1-4), upper primary school (grades 5-7), and lower secondary school (grades 8-10). Students typically stay in the same school during primary school. However, when starting lower secondary school (grade 8), most students switch schools, as only about one-third of primary schools are combined schools that cover grades 1-10.[6] The school cut-off date is January 1, and children start school the year they turn 6. School is free, and nearly all students attend their local school, with

[6] Based on own calculations from information for the school year 2021/2022 in The Norwegian primary and lower secondary information system (GSI) – a school-level based register: https://gsi.udir.no/informasjon/apne/.

only about 4% attending private schools. Responsibility for providing compulsory education is at the local government level (municipalities). Norway has a national curriculum and common legal framework covering all schools. The Norwegian Education Act (1998, § 8-2) allows for small-group instruction, although results from teacher surveys conducted in connection with this research project suggest that small-group instruction was not in widespread use at the time of the intervention (see Bonesrønning et al., 2022).

The 1+1 project

The 1+1 Project was made possible through a Norwegian government grant of about 20 million Euros to hire 80 qualified teachers for four school years. The grant followed a political decision in 2015 to recruit more teachers to lower primary schools to reduce the student-teacher ratio. The 25% additional teachers were distributed in a way that could provide new knowledge about potential class size reductions in promoting student learning.[7]

Treatment schools were given an extra teacher in the school years 2016/17-2019/20 to carry out tutoring in groups of 4-6 students parallel to ordinary mathematics instruction for specific grades using a pull-out strategy. The extra teacher covered the same topics as in the regular class to ease the transition between small-group and regular-class instruction. Importantly, small-group instruction targeted students of all ability levels, not just the lower-performing students, who are the focus of much of the research literature. Furthermore, all students were supposed to participate in small-group instruction for at least two periods during each school year, with each period lasting 4 to 6 weeks. The project recommended that students be grouped into small groups by ability level, and nearly all schools followed this recommendation (Bonesrønning et al., 2022).[8]

The treatment shared many features with high-impact tutoring programs (Robinson & Loeb, 2021), which have been shown to result in substantial learning gains (see Nickow et al. (2020) and Dietrichson et al. (2017) for recent reviews). However, many of these programs have even higher dosages (the majority between 10 weeks and one school year in Nickow et al. (2020)), are targeted at low-ability students, have one-on-one tutoring, and include increased instruction time – replacing recreational activities, unfilled time, or potentially crowding out instruction time in other subjects. Our treatment had a relatively lower tutoring dosage (two sessions of 4-6 weeks per year), i.e. an intervention that is potentially less costly, and held instruction time in the subject

[7] Two projects were awarded equivalent resources to hire teachers, the 1+1 project and Two Teachers. Two Teachers used the extra teachers mostly within the classroom for literacy instruction along with teacher professional development. See Haaland et al. (2022) for details.

[8] This project's analyses were pre-registered at OSF (Bonesrønning et al., 2018).

fixed by integrating the intervention into a standard school day in a way that did not crowd out other activities.

The guidelines imply that all students should have received a minimum planned treatment of about 30 hours (1800 minutes) of small-group instruction per year.[9] Registration forms show that the 2008 cohort of students received a dosage of about 64% of planned treatment (Bonesrønning et al. 2022, online appendix C).[10] Specifically, students received 1103 minutes in grade 3 (school year 2016/17) and 1184 minutes in grade 4 (school year 2017/18). However, planned treatment and the reported treatment received are not directly comparable, as teachers were asked to deduct time spent on breaks and other interruptions from their reports.

As teachers conducting the small group instruction are important parts of the treatment, we collected survey information on how extra teachers were hired and their background characteristics. Only teachers formally qualified to teach mathematics to primary school students were hired, a pre-requisite according to national legislation. In a survey covering the primary schools' small-group instructors and other mathematics instructors, 31% of small-group instructors reported that they had previously worked at the same school; a majority were recruited externally. As compared to regular mathematics instructors in primary school, a larger fraction of small-group instructors was male (28% vs. 13%), they were younger (30 vs. 42), and they had less teaching experience (12 vs. 19 years). On average, small group instructors had more mathematics credits from higher education than other teachers, amounting to about 2/3 of a semester more mathematics.

Four cohorts of children born in 2008-2011 participated in the intervention, with different starting ages and treatment duration. The last two cohorts were affected by Covid-19 during substantial parts of the intervention, potentially confounding the results of the intervention. This paper follows the 2008 and 2009 cohorts, which we previously showed benefited from treatment, as evaluated on the national test in grade 5 (Bonesrønning et al., 2022). We now have access to their national test scores in grade 8, allowing us to establish whether the effects are sustained when evaluated about 3.5 years after treatment ended.

Table 1 shows an overview of when the 2008 and 2009 cohorts were treated and sat for national tests. The 2008 cohort was treated for two years, starting in grade 3, and the 2009 cohort was treated for three years, starting in grade 2. This variation in treatment duration allows us to investigate whether there are differential effects by treatment dosage.

[9] Calculations are based on legislation stating that students in grades 1-4 are to receive about 140 60-minute units of mathematics instruction per year, which amounts to about 3.7 hours per week, as a school year lasts 38 weeks. A minimum of 8 weeks of small-group instruction roughly amounts to 30 hours per year.

[10] Small group instructors received an electronic registration form where they provided detailed information on which students participated in small-group instruction during each mathematics lesson and the duration of each session. They were asked to report on actual time spent on instruction, excluding time spent moving between classrooms, breaks, etc.

Students sit for the national test in the fall, typically around September/October of grades 5 and 8. As no test results are available before grade 5 in national registers, we developed math tests for younger grades in collaboration with teachers and math educators. For the first two cohorts, the focus of this paper, tests to capture baseline abilities were carried out early in the school year (August) before starting the intervention. These students also sat for project-administered end-of-year tests to capture the immediate effects of the intervention. The 2009 cohort had two end-of-year tests in second and third grade, while the 2008 cohort sat for one end-of-year test in third grade. These post-tests were conducted at the end of the school year (May-June). We use these data to identify short-term treatment effects at a younger age than the national tests and to examine treatment heterogeneity by baseline ability, as captured by the baseline test scores. Parental consent was required to merge test data collected through the project with register data.

Given the evidence in the literature on spillover effects of teachers and programs to other outcomes, we extend our analyses to investigate potential spillovers to the other skills measured by the national tests in grades 5 and 8, Norwegian (reading), and English as a foreign language. This spillover analysis is particularly interesting given that test scores in mathematics are shown to be important predictors of later life outcomes (see e.g., Murnane et al., 1995). Furthermore, the child development literature consistently finds that early mathematics skills strongly predict later reading skills (e.g., Duncan et al., 2007), and research on a kindergarten mathematics intervention finds spillovers to language skills (Sarama et al., 2012).

Table 1. Treatment timing and test-taking

School year → Cohort ↓	**2016/17**	**2017/18**	**2018/19**	**2019/20**	**2020/21**	**2021/22**	**2022/23**
2008	Grade 3	Grade 4	Grade 5^{TEST}	-	-	Grade 8^{TEST}	-
2009	Grade 2	Grade 3	Grade 4	Grade 5^{TEST}	-	-	Grade 8^{TEST}

Note: The shaded area indicates treatment years, and TEST indicates the time of sitting for the national tests.

Randomization

We recruited ten large and densely populated municipalities to participate in the intervention. Within each municipality, we ranked schools based on their average score on the national test in grade 5 in the two preceding school years (2014/15 and 2015/16). We then grouped schools within each municipality into strata, where each stratum consisted of 4 to 6 schools. Within each stratum, 36 in total, we randomly allocated schools into either the treatment or control group, with an equal probability of assignment to either group. Our approach follows Imbens (2011) recommendations of having at least two treatment and control schools in each stratum. One school refused to

participate in the project after revealing their treatment status. Following the pre-analysis plan, we exclude schools in this stratum from the analyses.

All schools in the treatment group received one extra teacher, irrespective of the number of students in the target group. This approach implied that, without further restrictions, the largest schools would be unable to give all targeted students the planned minimum treatment dosage. Therefore, we randomized classes or groups of students into treatment at large schools (more than 48 students or two classes). Overall, about 73-74% of students at treatment schools participated in the intervention.

Data and research design

Data

The main data source is administrative data collected and organized by Statistics Norway. Our administrative data set includes test scores on national tests in grades 5 and 8 with background information about the students (birth year, gender, parental education, immigrant background, and treatment status).[11] In addition, the project collected data, most notably baseline ability tests administered at the beginning of grade 2 for the 2009 cohort and grade 3 for the 2008 cohort in the first year of treatment and post-tests administered at the end of grade 2 for the 2009 cohort and grade 3 for the 2008 and 2009 cohorts, see Table 1 for details.[12] Students take national tests in numeracy, Norwegian (reading), and English as a second language. Our primary outcome measure is the national test in numeracy, as small-group instruction in mathematics likely has the most direct impact on numeracy skills. However, as noted above, we also examine spillovers to reading and English skills.

Sample

Our study expands on Bonesrønning et al. (2022) by analyzing national tests in grade 8 and investigating treatment heterogeneity by baseline ability and treatment dosages. However, the sample we use differs from the one in Bonesrønning et al. (2022) in three important ways.

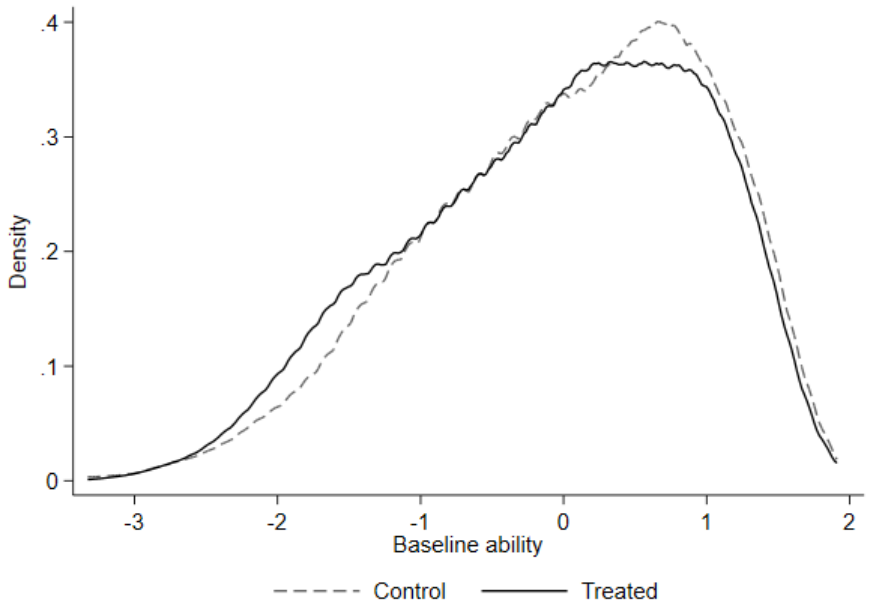
[11] National tests are digital, requiring access to a computer or tablet, and the main objective is to provide schools with information on core skills that they can use to improve their teaching quality. Exemptions from sitting the tests are to be practiced strictly, with exemptions given to students in special needs education or second language learners, and only if it is deemed that test scores will not provide schools with valuable information. Each year, about 5-6% of grade 5 students are exempt from sitting the test, and an additional 1-2% are coded as "did not participate", meaning they were absent on the test day. For the grade 8 test, about 4% are exempt, and 1-2% are coded as "did not participate". Even though students receive formal grades on these tests, they do not involve any stakes in terms of grade retention or promotion or contribute towards their grade point average at lower secondary school (students do not typically receive grades until lower secondary school).

[12] The project tests measured numeracy skills, similar to the national test but adjusted to better suit lower grade levels and developed in collaboration with teachers and math educators. The baseline tests were carried out early in the school year (August), and the post-tests were conducted at the end of the school year (May-June), see Table 1.

First, the national tests in grade 8 have missing scores for a subset of students in the 2009 cohort due to a nationwide teacher strike in the fall of 2022. The strike affected three out of the ten municipalities participating in the RCT, leading to a very high share of missing test scores among students in all relevant strata from these municipalities. Consequently, we have excluded all students in the 2009 cohort from these municipalities. This reduces the sample by about 15-17 percent (see Tables 3 and 4) compared to the sample used in Bonesrønning et al. (2022). Because the strike applied to all students within the municipalities, this attrition affects our statistical power but will not bias the causal estimates for the remaining schools in the sample.

Second, heterogeneity analyses require data linkage between administrative data and project-administered data, which is only possible for the subset of students whose parents provided written consent to merge the data files. This reduces the sample further by about 9%. Parental consent was collected after revealing treatment status, and the share with parental consent is, unfortunately, about 8 percentage points higher in the treatment group and higher among highly educated parents. This attrition can potentially bias our results.

Figure 1. Distribution of baseline ability in treatment and control group.



Note: The figure shows the distribution of test scores on the baseline ability test separately for students at treated and control schools. The x-axis shows standardized test scores.

Third, the baseline ability test, also administered after revealing the treatment status, was not carried out with similar thoroughness in control schools as in treatment schools. All tests show that randomization was successful in creating balanced treatment and control groups in the complete population of participants (Bonesrønning et al., 2022). However, Figure 1 shows that the share of students with low test scores at baseline is higher in the treatment schools. We

encouraged schools to allow students to retake the test if they were absent on the day of the test, and treatment schools apparently devoted more time to improving their participation rate (see Table A1), resulting in more low-performing students participating. Among the remaining sample of 11,260 students in grade 5 and 11,496 in grade 8, 504 (4.4%) and 529 (4.6%) students have missing baseline ability test scores in grades 5 and 8, respectively. In most specifications, we include students with missing baseline test scores while taking unbalanced missingness into account.

For these three reasons – missing and unbalanced consent, missing and unbalanced baseline ability test score, and missing national test score in grade 8 – our sample and analysis approach differs somewhat from the previous analysis of short-run effects on test scores in grade 5 (Bonesrønning et al., 2022).

Descriptive statistics and balance

Table 2 presents descriptive statistics on covariates and assesses the balance between treatment and control groups for students with non-missing national test scores in grade 5.[13] Panel A presents balance using the same sample as in Bonesrønning et al. (2022). Panel B presents balance when restricting the sample to students unaffected by the strike in the fall of 2022. Finally, Panel C presents balance when we restrict the sample to students not affected by the strike and who also provided parental consent to link test scores.

The covariates in Table 2 are those we use in the empirical analyses. We define two indicator variables for minority background: Born in Norway with foreign-born parents (second-generation immigrant) and foreign-born (first-generation immigrant). We divide parental education level (measured by the year the child turns six years of age, i.e. the year they start school) into four categories, based on the education level of the parent with the highest completed education level: *i*) Low: unknown or lower secondary (compulsory education), *ii*) USE: upper secondary education, *iii*) HE BSc: short higher education (Bachelor's degree), or *iv*) HE MSc+: long higher education (Master's degree or PhD).[14]

[13] The requirements to be exempted from the national tests in grade 8 are the same as for the national tests in grade 5, and the results are very similar for students with non-missing national test scores in grade 8.

[14] We deviate slightly from the pre-analysis plan regarding our choice of covariates, as we do not have the expected information on class size and teacher-student ratio broken down by cohort and school class. In addition, we have merged the groups "less than primary education or unknown education" and "primary education" into one group since less than four percent of the students had missing/unknown parental level of education.

Table 2. Descriptive statistics on covariates and balance tests

	Control Mean (SD)	Treatment Mean (SD)	Difference (1)-(2)
Panel A	Same sample as in Bonesrønning et al. (2022)		
Female	0.481 (0.006)	0.488 (0.007)	-0.007
Parental edu. Low	0.090 (0.008)	0.093 (0.009)	-0.003
Parental edu. USE	0.213 (0.012)	0.196 (0.013)	0.017
Parental edu. HE BSc	0.390 (0.009)	0.373 (0.009)	0.017
Parental edu. HE MSc+	0.308 (0.019)	0.339 (0.019)	-0.031*
1st gen. immigrant	0.063 (0.005)	0.064 (0.004)	-0.000
2nd gen. immigrant	0.100 (0.011)	0.101 (0.013)	-0.002
Cohort 2009	0.508 (0.005)	0.504 (0.006)	0.004
F-test joint significance			1.093
N [Clusters]	8128 [81]	8148 [78]	16276 [159]
Panel B	Sample not affected by strike		
Female	0.483 (0.007)	0.488 (0.008)	-0.005
Parental edu. Low	0.095 (0.009)	0.095 (0.010)	0.000
Parental edu. USE	0.223 (0.012)	0.205 (0.014)	0.018
Parental edu. HE BSc	0.392 (0.010)	0.374 (0.010)	0.018
Parental edu. HE MSc+	0.291 (0.020)	0.326 (0.020)	-0.036*
1st gen. immigrant	0.069 (0.005)	0.066 (0.005)	0.003
2nd gen. immigrant	0.104 (0.012)	0.106 (0.015)	-0.001
Cohort 2009	0.409 (0.021)	0.392 (0.022)	0.017
F-test joint significance			1.214
N [Clusters]	6772 [81]	6651 [78]	13423 [159]
Panel C	Sample not affected by strike and with parental consent		
Female	0.490 (0.007)	0.492 (0.008)	-0.002
Parental edu. Low	0.082 (0.008)	0.089 (0.010)	-0.008
Parental edu. USE	0.212 (0.013)	0.202 (0.014)	0.010
Parental edu. HE BSc	0.398 (0.010)	0.377 (0.010)	0.021*
Parental edu. HE MSc+	0.309 (0.021)	0.332 (0.020)	-0.023
1st gen. immigrant	0.065 (0.005)	0.060 (0.005)	0.004
2nd gen. immigrant	0.095 (0.011)	0.102 (0.014)	-0.007
Cohort 2009	0.417 (0.022)	0.398 (0.022)	0.019
Baseline ability	0.029 (0.032)	-0.081 (0.038)	0.110***
Missing baseline ability	0.067 (0.015)	0.034 (0.008)	0.033***
F-test joint significance			3.079***
N [Clusters]	5818 [81]	6263 [78]	12081 [159]

Note: The results in the last column come from an OLS regression comparing treatment and control schools while controlling for strata and cohort fixed effects. Standard errors clustered on school at time of randomization in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

For the sample used in Bonesrønning et al. (2022), Panel A, the share of students with parents with a long higher education is somewhat higher in the treatment group. Still, the joint F-test of all covariates is low and statistically insignificant. For the sample not affected by the strike, Panel B, we see a similar pattern with some imbalance for students with parents with longer

education. Again, the F-test of all covariates is low and not statistically significant. For the sample not affected by the strike and with parental consent, Panel C, we see strong imbalances for baseline ability and missing baseline ability, as expected. Students in the control schools have about .11 standard deviations higher test scores on the baseline test and about three percentage points higher missingness. The joint F test is also statistically significant (F=3.08, p<.01). However, if we exclude the baseline test score variables, the sample is balanced across the covariates (F=1.72, p=.11).

The imbalance in observed baseline ability in the sample in Panel C implies that an unadjusted treatment-control comparison when we condition on observing baseline ability may be a biased estimate of the treatment effect. In our analyses, we include covariates as controls, include specifications where all covariates interact with the treatment indicator using the Lin (2013) procedure, and estimate entropy balancing weights (Hainmueller, 2012) within each stratum to reweigh the control group so that it matches the covariate means in the treatment group within the sample.

Empirical specification

We estimate the intention-to-treat (ITT) effects using the following model:

$$y_{igs} = \beta TREATED_g + \alpha_s + X_i'\gamma + \epsilon_i$$

where *i* refers to individuals, *g* schools, and *s* randomization strata. *y* is the test score, and $TREATED_g$ is an indicator variable equal to 1 if the student was enrolled in a treatment school at baseline. Because randomization was performed within strata, we always include strata-fixed effects (α_s). The vector *X* represents the covariates. Standard errors are clustered at the school level.

We augment the above specification with an interaction term between $TREATED_g$ and the baseline test score to test whether the treatment effect differs for high and low-ability students. In these specifications, we include a full set of interactions between all covariates and $TREATED_g$, which will reduce bias in the treatment effect estimate from the imbalance between the groups. We follow Lin (2013) and include interactions between the treatment indicator and mean-centered covariates, which could improve statistical power. With this procedure, β continues to be the treatment effect for the average student even when all interactions are included, which eases comparison across specifications.

We treat all students at treatment schools as treated despite within-school randomization to treatment in larger schools. We take this approach as there may be potential spill-over effects from the treated classes to the other classes and out of a concern that treated schools could have changed the class compositions in response to their revealed treatment status. Our classification

ensures that β is a clean ITT estimate but likely represents a lower-bound estimate of the treatment effect. Given that 73-74% of students at treatment schools were treated, the ITT estimate can be scaled by 1.3 to provide an estimate of the local average treatment effect (LATE).

Main results

We first estimate the main treatment effects of the intervention. The outcome variables are test scores capturing numeracy skills from national tests in grades 5 and 8, i.e., about five months and 3.5 years after the treatment ended, respectively. Since, as explained in Section 3.2., the analyses for grade 8 use a different sample compared to that in Bonesrønning et al. (2022), we start by revisiting results for grade 5 to see how the sample restrictions affect our estimation results.

Baseline results for national tests in numeracy in grade 5

Table 3 presents the treatment effect for national tests in grade 5 using various specifications and sample restrictions. The first two columns replicate Bonesrønning et al. (2022), finding an effect of around 6% of a standard deviation on national tests in grade 5. The first column reports results without covariates, and the second column reports results including the covariates listed in Table 2.

Specifications in columns (3) and (4) are similar to those in columns (1) and (2), respectively, but exclude students affected by the teacher strike in the fall of 2022 (students in the 2009 cohort in three municipalities). We observe that the point estimate of the treatment effect on national test scores in grade 5 increases somewhat to about 9% of a standard deviation when excluding these students, but the estimates do not seem to differ statistically.[15] Recall that this sample is balanced across covariates; see Table 2.

Next, we present treatment effects when we further restrict the sample to students with parental consent, where we can merge baseline ability with national test scores. Columns (5) and (6) have the same specifications as columns (1) and (2), respectively. We observe that the size of the unconditional treatment effect in column (5) is 0.08, about 18% smaller than the estimate in column (3). Thus, the treatment-control imbalance in parental consent causes a downward bias in the estimate. The treatment effect decreases somewhat to 0.072 when we include covariates in

[15] The difference between the treatment effect estimates in columns (3) and (1) is .032, with a standard error of .046. The corresponding difference for columns 4 and 2 is .028 (SE= .040). The standard error is calculated through a simplified approach defined as: $se_{diff} = \sqrt{se_{\beta 1}{}^2 + se_{\beta 2}{}^2}$.

column (6) and is closer to that in column (4) (about 16% smaller), implying that the included covariates explain some of the imbalances caused by conditioning on parental consent.[16]

In column (7), we include the baseline test score, thereby reducing the sample size.[17] Columns (7)-(9) include control variables for baseline ability. In column (7), we include a z-score variable on baseline ability from the baseline test conducted at the beginning of grades 2 and 3 for the 2009 and 2008 cohorts, respectively. In this model, students with missing test scores are omitted but included in columns (8) and (9). In column (9), the model is augmented with a full set of treatment-covariate interactions to improve power (see above and Lin (2013)). In these specifications, the average treatment effect increases to 13-14% of a standard deviation but is very similar across models, i.e., how we control for baseline ability is not important for the estimates.

Table 3. Treatment effect on national numeracy test scores in grade 5

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Sample	All		No strike		No strike, consent				
Treatment effect	0.066**	0.058**	0.098***	0.086***	0.080**	0.072**	0.139***	0.133***	0.132***
	(0.031)	(0.026)	(0.034)	(0.030)	(0.035)	(0.031)	(0.036)	(0.035)	(0.035)
Baseline test, excl. missing (z-score)							0.588***		
							(0.013)		
Baseline test (z-score)								0.587***	0.596***
								(0.013)	(0.017)
Baseline test missing (D)								-0.211***	-0.160**
								(0.055)	(0.074)
Treatment interacted by:									
- Baseline test (z-score)									-0.016
									(0.026)
- Baseline test missing (D)									-0.149
									(0.113)
Covariates	No	Yes	No	Yes	No	Yes	Yes	Yes	Yes
Strata fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	14,891	14,891	12,350	12,350	11,260	11,260	10,756	11,260	11,260
R-squared	0.041	0.135	0.038	0.135	0.038	0.134	0.434	0.421	0.423

Note: Each column represents the results from a separate OLS regression, comparing students at treatment and control schools while controlling for strata- and cohort fixed effects. In models with covariates, we include controls for background characteristics. Robust standard errors, adjusted for clustering at the school level, are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Baseline ability is, as expected, strongly correlated with the national test scores, as a one standard deviation difference in baseline ability is associated with a difference in national test scores

[16] In Appendix Table A2, we replicate the model in column (6) with a sample that includes the municipalities affected by the strike. We see that this reduces the treatment effect, indicating that on average, students in the 2009 cohorts in striking municipalities seem to have relatively lower gains from the intervention. This suggests some heterogeneity in treatment effects across locations due to differences in implementation across schools or municipalities. Table A2 also shows point estimates for the control variables. They are as expected, where boys and non-immigrants perform better than girls and students with immigration backgrounds, respectively. We also observe a strong association between test scores and parental education, but no statistical difference in average test scores between the two cohorts.

[17] Columns (1) and (2) in Table A3 in the appendix report treatment effect when we run the same models as in columns (5) and (6) but restricted to students with information on their baseline ability. We observe similar treatment effects among this reduced sample of students as the full sample with parental consent. The effect is estimated to 0.072 and 0.070, respectively.

in grade 5 of almost 60% of a standard deviation. The missing baseline test dummy variable is negative which is consistent with there being a relatively high share of students in the lower part of the ability distribution who did not participate in the baseline test. The interaction term between treatment and missing baseline test in column (9) is not statistically significant, but the negative sign is consistent with the lower missing share in the treatment group.

The main takeaway from Table 3 is that when adjusting for the baseline test score, which is our preferred specification in the "No strike, consent" sample, we find larger treatment effects than in the two other samples. Even though we consider that adjusting for baseline ability is the best approach, given that we know that the selection into this sample depends on baseline ability, we cannot completely rule out that these differences reflect a bias that we are unable to fully adjust for. Alternatively, these differences may be a result of differences in treatment effects across samples.

Results for national tests in numeracy in grade 8

Table 4 presents estimated treatment effects on national tests in grade 8, 3.5 years after the treatment ended. The columns are comparable to columns (3)-(9) of Table 3. Columns (1) and (2) show treatment effects when we exclude students affected by the strike (students in the 2009 cohort in three municipalities). We find that the estimated treatment effect decreases somewhat compared to the estimated effects on the national test scores in grade 5. In column (1), where the effect is unconditional on covariates, the effect is 5.4% of a standard deviation and significant at the 10% level. When we include covariates in column (2), the effect size decreases somewhat and is no longer statistically significant.

The next two columns present the results when we restrict the sample to students with parental consent to link data. For this sample, the effect is estimated to be around 2-3% of a standard deviation; see columns (3) and (4). The smaller coefficients, as compared to columns (1) and (2), are in line with what we found for test scores in grade 5, namely that the treatment-control imbalance in parental consent causes a downward bias in the estimate.[18]

[18] Columns (3) and (4) in Table A3 in the appendix report treatment effect when we run the same models as in columns (3) and (4) only for the students with information on baseline ability. We observe quite similar treatment effects among this reduced sample of students compared to the full sample with parental consent. The effect is estimated to 0.015 and 0.013, respectively.

Table 4. Treatment effect on national numeracy test scores in grade 8

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Sample	No strike		No strike, consent				
Treatment effect	0.054*	0.039	0.027	0.019	0.083***	0.081***	0.081***
	(0.032)	(0.025)	(0.032)	(0.026)	(0.031)	(0.030)	(0.029)
Baseline test, excl. missing (z-score)					0.559***		
					(0.012)		
Baseline test (z-score)						0.557***	0.565***
						(0.012)	(0.015)
Baseline test missing (D)						-0.244***	-0.241***
						(0.046)	(0.054)
Treatment interacted by:							
- Baseline test (z-score)							-0.015
							(0.024)
- Baseline test missing (D)							-0.017
							(0.102)
Covariates	No	Yes	No	Yes	No	Yes	Yes
Strata fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	12,649	12,649	11,496	11,496	10,967	11,496	11,496
R-squared	0.040	0.154	0.038	0.151	0.432	0.420	0.421

Note: Each column represents the results from a separate OLS regression, comparing students at treatment and control schools while controlling for strata- and cohort fixed effects. In some specifications, we also include controls for background characteristics. Robust standard errors, adjusted for clustering at the school level, are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

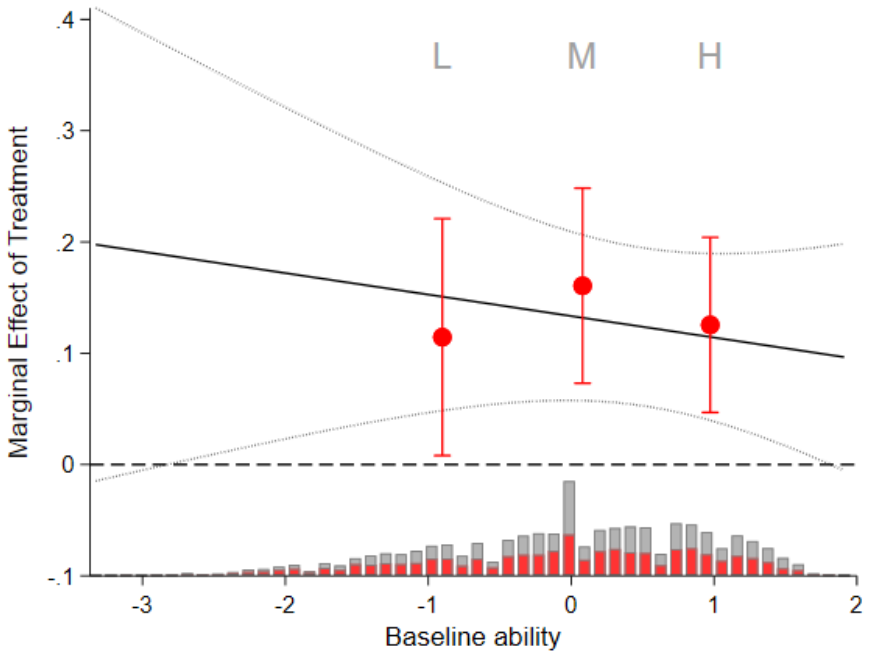
When we control for baseline test scores, which we do in columns (5) to (7), using the same specifications as columns (7) to (9) in Table 3, the estimates increase to around 0.08 standard deviations when taking baseline test scores into account. We find similar magnitudes when utilizing entropy reweighting (Hainmueller, 2012), as shown in column (2) in Table A4. The models with baseline test controls suggest a sustained effect of small-group instruction in mathematics. Some of the impact fades out as the treatment effect decreases from around 0.14 standard deviations five months after treatment ended to around 0.083 standard deviations 3.5 years after treatment ended. The decline from grade 5 to 8 is about 40-50% across specifications.

The fadeout from grade 5 to grade 8 is a continuation of a fadeout process from grades 2 and 3, measured using project-administered tests, to national tests in grade 5; see Table A5. Our estimates suggested an end-of-the-year treatment effect of about 0.20 standard deviations, which declines to 0.14 standard deviations in grade 5 and about 0.083 standard deviations in grade 8. This continued but substantially slowed fadeout is in keeping with prior studies and with the concept that some of what is learned in a year of school are general knowledge and skills that continue to be relevant for future learning, and some are grade-specific and not captured by future assessments (Masters et al., 2017).

Heterogeneity by baseline ability

Our main results document significant and sustained average effects of small-group instruction in mathematics. Earlier studies of similar treatment mainly focus on low-performing students and, thus, have a limited range of starting achievement for which to assess heterogeneity. Our sample includes the full range of prior achievements, and analyzing whether the impact differs along the ability distribution is highly relevant for understanding how small-group instruction works as a universal intervention. In addition, our two cohorts received different dosages, where the 2009 cohort was treated for one more year than students in the 2008 cohort and started treatment at an earlier age (age 7 versus 8). This variation further allows us to examine if fadeout along the ability distribution depends on treatment dosage.

Figure 2. Heterogeneity across baseline ability on grade 5 numeracy test scores



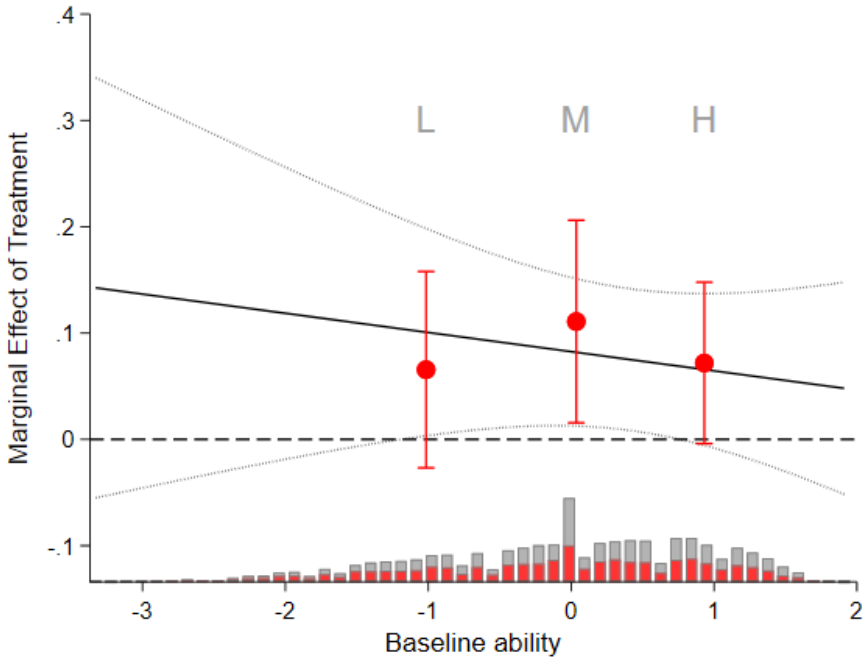
Note: The figure shows the estimated marginal effects using both a conventional linear interaction model and a binning estimator. The total height of the stacked bars refers to the distribution of the moderator (individual baseline test score) in the pooled sample, and the red and grey shaded bars refer to the distributions in the treatment and control groups, respectively.

Figure 2 shows treatment effects on national test scores in grade 5 for students with different baseline test scores, as measured in our project-administered baseline test in mathematics. The regression line indicates a negative slope in treatment effects, but this association is not statistically significant, as seen also from the coefficient and standard error of the treatment-baseline test score interaction in column (9) of Table 3. Similarly, the effect size is similar if we group the students into three ability groups (low, medium, and high). The L (low), M (medium),

and H (high) point estimates and bars in red are treatment effect estimates from a regression where the baseline test score is divided into three equal-sized bins; see Hainmueller et al. (2019) for details. The point estimates indicate the largest gains for students in the middle of the ability distribution, but the differences between the groups are small.[19]

Figure 3 shows heterogeneity in national test scores in grade 8 by baseline test scores. The pattern is similar to what we find in Figure 2 when analyzing national test scores in grade 5. The negative association with baseline ability is strikingly similar across the tests and not statistically significant (see interaction term and standard error in column (7) of Table 4). The pattern for the group-specific estimates is also strikingly similar; while larger for the middle group, the differences are small and not statistically significant. The pattern is similar also when examining the short-term test scores from grades 2 and 3 (see Bonesrønning et al., 2022).

Figure 3. Heterogeneity across baseline ability on grade 8 numeracy test scores



Note: The figure shows the estimated marginal effects using both a conventional linear interaction model and a binning estimator. The total height of the stacked bars refers to the distribution of the moderator (individual baseline test score) in the pooled sample, and the red and grey shaded bars refer to the distributions in the treatment and control groups, respectively.

Finally, since the two cohorts received different treatment dosages (2 vs. 3 years), Table 5 presents cohort-specific estimates for test scores in grades 5 and 8. We conduct this analysis on the sample of schools unaffected by the teacher strike in 2022 and where parents consented to

[19] The high share of students with baseline test scores equal to 0 in the figure is misleading because students with missing baseline test scores are set to 0. Since a dummy for missing baseline test and its interaction with treatment is included in the analysis, the students with missing baseline test scores do not contribute to the point estimate for any groups.

merge baseline test scores with register data because we need the baseline scores to examine heterogeneity.[20] All regressions include controls for background characteristics and baseline test scores in mathematics, and all controls are interacted with the treatment status.

Table 5. Heterogeneity across cohort and treatment dosage

	(1)	(2)	(3)	(4)
National test	Grade 5, numeracy		Grade 8, numeracy	
Cohort	2008	2009	2008	2009
Treatment effect	0.149***	0.166***	0.098**	0.104**
	(0.049)	(0.048)	(0.044)	(0.045)
Baseline test (z-score)	0.647***	0.521***	0.601***	0.514***
	(0.022)	(0.026)	(0.019)	(0.022)
Baseline test, missing (D)	-0.098	-0.028	-0.186**	-0.345***
	(0.103)	(0.097)	(0.080)	(0.091)
Treatment interacted by:				
- Baseline test (z-score)	-0.005	0.024	0.008	-0.014
	(0.029)	(0.036)	(0.027)	(0.032)
- Baseline test, missing (D)	-0.487**	-0.159	-0.489**	0.114
	(0.205)	(0.145)	(0.189)	(0.138)
Covariates	Yes	Yes	Yes	Yes
Covariates interacted with treatment	Yes	Yes	Yes	Yes
Strata fixed effects	Yes	Yes	Yes	Yes
Observations	4,571	4,321	4,500	4,663
R-squared	0.502	0.351	0.475	0.378

Note: Each column represents the results from a separate OLS regression, comparing students at treatment and control schools while controlling for strata- and cohort-fixed effects. Included schools are identical across the two cohorts since we exclude students in the 2008 cohort at schools affected by the teacher strike in 2022. As a result, the number of observations and estimates are not directly comparable to the corresponding models in Table 3 and Table 4. We also exclude students without parental consent. All specifications include controls for background characteristics, baseline test scores in numeracy, and all covariates are interacted with the treatment indicator. Robust standard errors, adjusted for clustering at the school level, are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Standard errors clustered on school at time of randomization in parentheses.

Columns (1) and (2) in Table 6 indicate that the treatment effects are about the same magnitude on national test scores in grade 5 regardless of treatment duration (2 vs. 3 years). The effect is estimated to be 0.149 standard deviations and 0.166 standard deviations, respectively. We find similar results in columns (3) and (4) when analyzing effects on national test scores in grade 8. In all models, the interactions between treatment and baseline ability are small, meaning there is no heterogeneity by baseline ability for either cohort. All in all, upholding small-group instruction for a third year seems inefficient for the outcomes we measure.

[20] Note that the samples differ from the corresponding models in Tables 3 and 4 since we also exclude the 2008 cohort students at schools affected by the teacher strike in 2022. Thus, the sample of schools is identical across the two cohorts in the analyses in Table 5.

Spillover effects on other cognitive skills

The intervention may not only have affected math learning; it might also have generated spillover effects to other cognitive skills. In the same period as the national test in numeracy is conducted, students also sit for national tests in Norwegian (reading) and English as a foreign language. Panel A in Table 6 shows results for national tests in grade 5, whereas Panel B presents the results for national tests in grade 8. Since we know the main sample has some imbalances (i.e., the sample of students with parental consent), baseline controls are necessary in these analyses. Unfortunately, we do not have information about baseline ability in Norwegian (reading) and English as a foreign language, as students were not tested on these skills before the intervention. As a proxy, we include baseline ability in mathematics as a control variable.

Table 6. National tests in Norwegian (reading) and English as a second language

	(1)	(2)	(3)	(4)
National test	Norwegian (reading)		English	
Sample	No strike	No strike, consent	No strike	No strike, consent
Panel A: Grade 5 test:				
Treatment effect	0.048*	0.067**	0.002	0.029
	(0.0269)	(0.0311)	(0.0282)	(0.0298)
Observations	12,212	11,132	12,393	11,292
R-squared	0.119	0.273	0.050	0.153
Panel B: Grade 8 test:				
Treatment effect	0.035	0.072**	0.022	0.059**
	(0.0237)	(0.0280)	(0.0234)	(0.0277)
Observations	12,666	11,517	12,604	11,464
R-squared	0.160	0.332	0.082	0.235
Covariates incl. baseline ability in mathematics	No	Yes	No	Yes
Covariates interacted with treatment	No	Yes	No	Yes
Strata fixed effects	Yes	Yes	Yes	Yes

Note: OLS regression with robust standard errors adjusted for clustering on school in parentheses. Observations for students in the 2009 cohort attending schools in strata affected by the teacher strike in 2022 are excluded from all samples. *** p<0.01, ** p<0.05, * p<0.1***.

Columns (1) and (3) show the results for the "No strike" sample, with controls for background characteristics but without controls for baseline ability. Columns (2) and (4) show the same estimations for the "No strike, consent" sample, where we control for the unbalance in parental consent by including baseline ability in mathematics as a control as well as interacting all covariates with the treatment indicator. The results provide evidence that the intervention created beneficial spillover effects, most strongly so for Norwegian (reading) test scores, which are sustained over time with measurable effects detected for national tests in grade 8 for the "No strike, consent"

sample. The result for Norwegian (reading) test score is in line with the child development literature on the predictive power of mathematics on reading skills (see, e.g., ten Braak et al., 2022). Although not significant, the coefficients in the "No strike" sample are 3.5 and 2.2 percent of a standard deviation in Norwegian and English, respectively. As discussed in connection to Table 3, any differences in the treatment effects between the two samples can be due to differential treatment effects. However, we cannot rule out that these differences are driven by a bias in the "Consent" sample that we cannot fully adjust for. The results from the "No strike" sample therefore provide the most conservative estimates.

Discussion and conclusion

In this paper, we contribute to the literature on student-teacher ratio on student outcomes, and to the growing literature documenting beneficial effects of tutoring interventions. We investigate the longer-run effects of an intervention that used additional teachers to provide small-group instruction in mathematics to all students using a pull-out strategy, and where instruction time was held constant (Bonesrønning et al., 2022).

An important feature of this intervention is that it did not introduce any new curriculum, as teachers were instructed to cover the same topics in both the regular class and in the small groups. This was done to ease the transition between small-group and regular-class instruction. As a result, our findings suggest that the increase in skills comes from how the intervention was organized, allowing students of all ability levels to receive tailored instruction in small groups. Another aspect that distinguishes this intervention from other tutoring interventions is that instruction time was held fixed, with a lower dosage (on average 8 weeks per year) than most other tutoring interventions, which typically last from ten weeks and one school year (Nickow et al., 2020), and as such contributes to the understanding of the impact of lower-dosage tutoring.

A growing number of interventions have shown beneficial effects of high-impact tutoring programs on outcomes measured shortly after the interventions ended (Robinson & Loeb, 2021; Nickow et al., 2020; Dietrichson et al., 2017). However, much less is known about the lasting effects of tutoring interventions, which is also true for education interventions more generally. The few studies that provide estimates of longer-run outcomes during students' compulsory schooling often find that effects substantially diminish or completely fade out (Bailey et al., 2020; Hart et al., 2023; Andersen et al., 2022).

Identifying the longer-term effects of school interventions can be challenging because teachers and schools respond to students' achievement levels, often providing more support for lower-performing students. If an intervention benefited some students, teachers and schools may

then work to help the other students. The school-level randomization in this intervention helps reduce these responses since all students in the school either received or did not receive the treatment. As a result, the long-run estimates are likely to be more accurate than they would have been for student-level randomization within schools. Moreover, Norway provides a particularly stable context for estimating fadeout because Norwegian primary school students usually remain with the same school over time, so few, if any, students move between treatment and control schools. Our results suggest that in this environment, it was possible to sustain the effects of the intervention over time.

The results in this paper indicate that the initial beneficial effects of the tutoring intervention persisted – about 60% of the effect persisted 3.5 years after the program ended, with a similar degree of persistence across the ability distribution. Furthermore, we find that the intervention created spillovers to reading skills, with suggestive evidence that these beneficial effects were sustained over time. This result is in line with the child development literature, where it is well-established that early mathematics skills have strong predictive power on reading skills (see e.g. ten Braak et al., 2022). It could also be due to the intervention itself improving school engagement, well-being, or skill development in other domains, e.g. socio-emotional skills that were beneficial for later language development.

Finally, our findings demonstrate that lowering the student-teacher ratio can increase student outcomes in a resource-rich context, where previous research has shown no or small effects on student outcomes (Leuven et al., 2008; Iversen & Bonesrønning, 2013; Falch et al., 2017; Leuven & Løkken, 2018; Haaland et al., 2022; Borgen et al., 2022). This encouraging finding suggests that it matters *how* teachers are being used and that targeting mathematics can generate beneficial spillovers to other cognitive skills.

References

Andersen, S. C., Bodilsen, S. T., Houmark, M. A., & Nielsen, H. S. (2022). Fade-Out of Educational Interventions: Statistical and Substantive Sources. CESifo Working Paper No. 10094.

Bailey, D. H., Duncan, G. J., Odgers, C., & Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of Research on Educational Effectiveness*, 10, 7–39.

Bailey, D. H., Duncan, G. J., Cunha, F., Foorman, B. R., & Yeager, D. S. (2020). Persistence and fade-out of educational-intervention effects: Mechanisms and potential solutions. *Psychological Science in the Public Interest,* 21(2), 55-97.

Bailey, M. J., Sun, S., & Timpe, B. (2021). Prep School for poor kids: The long-run impacts of Head Start on Human capital and economic self-sufficiency. *American Economic Review*, 111(12), 3963-4001.

Bonesrønning, H., Finseraas H., Hardoy I., Iversen J. M. V., Nyhus O. H., Opheim V., Salvanes, K. V., Sandsør, A. M. J., & Schøne, P. (2018). The Effect of Small Group Instruction in Mathematics for Pupils in Lower Elementary School. OSF pre-registration. doi:10.17605/OSF.IO/YWQVC

Bonesrønning, H., Finseraas, H., Hardoy, I., Vaag Iversen, J. M., Nyhus, O. H., Opheim, V., Salvanes, K. V., Sandsør, A. M. J. & Schone, P. (2022). Small-group instruction to improve student performance in mathematics in early grades: Results from a randomized field experiment. *Journal of Public Economics 216*.

Borgen, N.T., Kirkebøen, L.J., Kotsadam, A., & Raaum, O. (2022). Do funds for more teachers improve student performance? *CESifo Working Paper, No. 9756*, 10.2139/ssrn.4120148

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *The Quarterly journal of economics,* 126(4), 1593-1660.

Cunha, F., & Heckman, J. J. (2007). The technology of skill formation. *American Economic Review*, 97(2), 31–47.

Currie, J., & Thomas, D. (2000). School quality and the longer-term effects of Head Start. *Journal of Human Resources*, 35, 755–774.

Deming, D. (2009). Early childhood intervention and life-cycle skill development: Evidence from Head Start. *American Economic Journal: Applied Economics, 1*(3), 111-134.

Dietrichson, J., Bøg, M., Filges, T., & Klint Jørgensen, A.-M. (2017). Academic Interventions for Elementary and Middle School Students with Low Socioeconomic Status: A Systematic Review and Meta-Analysis. *Review of Educational Research, 87*(2), 243–282.

Dodge, K. A., Bierman, K. L., Coie, J. D., Greenberg, M. T., Lochman, J. E., McMahon, R. J., & Pinderhughes, E. E. (2015). Impact of early intervention on psychopathology, crime, and well-being at age 25. *American Journal of Psychiatry*, *172*(1), 59–70. doi:10.1176/appi.ajp.2014.13060786

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. *American Economic Review, 101*(5), 1739-1774, doi:10.1257/aer.101.5.1739

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K., & Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43*(6), 1428-1446.

Falch, T., Sandsør, A.M.J., & Strøm, B. (2017). Do smaller classes always improve students' long-run outcomes? *Oxford Bulletin of Economics and Statistics, 79*(5), 654-688, 10.1111/obes.12161

Gersten, R., Haymond, K., Newman-Gonchar, R., Dimino, J., & Jayanthi, M. (2020). Meta-analysis of the impact of reading interventions for students in the primary grades. *Journal of Research on Educational Effectiveness, 13*(2), 401-427.

Green, C., & Iversen, J. M. V. (2022). The effect of class size in schools on student outcomes. In S. Mendolia, M. O'Brien, A. R. Paloyo, & O. Yerokhin (Eds.), *Critical Perspectives on Economics of Education*. Routledge. https://doi.org/10.4324/9781003100232

Guryan, J., Ludwig, J., Bhatt, M. P., Cook, P. J., Davis, J. M., Dodge, K., Farkas, G., Fryer Jr, R. G., Mayer, S., & Pollack, H. (2021). Not Too Late: Improving Academic Outcomes Among Adolescents. *NBER Working Paper No. 28531*. 10.3386/w28531

Haaland, V. F., Rege, M., & Solheim, O. J. (2024). Do students learn more with an additional teacher in the classroom? Evidence from a field experiment. *The Economic Journal*, 134(657), 418-435.

Hainmueller, J. (2012). Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Political Analysis, 20*(1), 25-46. doi:10.1093/pan/mpr025

Hainmueller, J., Mummolo, J., & Xu, Y. (2019). How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice. *Political Analysis, 27*(2), 163-192, doi:10.1017/pan.2018.46

Hart, E. R., Bailey, D. H., Luo, S., Sengupta, P., and Watts, T. W. (2023). Do Intervention Impacts on Social-Emotional Skills Persist at Higher Rates than Impacts on Cognitive Skills? A Meta-Analysis of Educational RCTs with Follow-Up. (EdWorkingPaper: 23-782). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/7j8s-dy98

Heckman, J., Pinto, R., & Savelyev, P. (2013). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review, 103*(6), 2052-2086.

Imbens, G. (2011). Experimental Design for Unit and Cluster Randomized Trials. *International Initiative for Impact Evaluation Paper.*

Iversen, J.M.V., & Bonesrønning, H. (2013), Disadvantaged Students in the Early Grades: Will Smaller Classes Help Them? *Education Economics, 21*(4), 305-324. doi: 10.1080/09645292.2011.623380

Kirkebøen, L. J., Gunnes, T., Lindenskov, L. & Rønning, M. (2021). Didactic methods and small-group instruction for low-performing adolescents in mathematics. Results from a randomized controlled trial. *Discussion Papers 957, Statistics Norway.*

Leuven, E., & Løkken, S.A. (2020). Long-term impacts of class size in compulsory school. *Journal of Human Resources, 55*(1), 309-348. doi: 10.3368/jhr.55.2.0217.8574R2

Leuven, E., & Oosterbeek, H. (2018). Class size and student outcomes in Europe. *EENEE Analytical Report No. 33.* doi: 10.2766/28340

Leuven, E., Oosterbeek, H., & Rønning, M. (2008). Quasi-Experimental Estimates of the Effect of Class Size on Achievement in Norway. *Scandinavian Journal of Economics, 110*(4) (2008), 663-693, doi: 10.1111/j.1467-9442.2008.00556.x

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

Ludwig, J., & Miller, D. L. (2007). Does Head Start improve chidren's life chances? Evidence from a regression discontinuity design. *The Quarterly journal of economics*, 122(1), 159-208.

Master, B., Loeb, S., & Wyckoff, J. (2017). More Than Content: The Persistent Cross-Subject Effects of English Language Arts Teachers' Instruction. *Educational Evaluation and Policy Analysis, 39*(3), 429-447. doi: 10.3102/0162373717691611

Murnane, R. J., Willett, J. B., & Levy, F. (1995). The Growing Importance of Cognitive Skills in Wage Determination. *The Review of Economics and Statistics, 77*(2) 251-266.

Neitzel, A. J., Lake, C., Pellegrini, M., & Slavin, R. E. (2022). A synthesis of quantitative research on programs for struggling readers in elementary schools. *Reading Research Quarterly, 57*(1), 149-179.

Nickow, A., Oreopoulos, P., & Quan, V. (2020). The Impressive Effects of Tutoring of PreK-12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence. *NBER Working Paper No. 27476*. 10.3386/w27476

Pages, R., Lukes, D. J., Bailey, D. H., & Duncan, G. J. (2020). Elusive longer-run impacts of head start: Replications within and across cohorts. *Educational Evaluation and Policy Analysis, 42*(4), 471-492.

Robinson, C. D., & S. Loeb. (2021). High-impact tutoring: State of the research and priorities for future learning. (EdWorkingPaper: 21-384). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/qf76-rj21

Sarama, J., Lange, A. A., Clements, D. H., & Wolfe, C. B. (2012). The impacts of an early mathematics curriculum on oral language and literacy. Early Childhood Research Quarterly, 27(3), 489-502.

Schanzenbach, D. W. (2020). The economics of class size. In The Economics of Education (pp. 321-331). Academic Press.

Smith, T. M., Cobb, P., Farran, D. C., Cordray, D. S., & Munter, C. (2013). Evaluating Math Recovery: Assessing the Causal Impact of a Diagnostic Tutoring Program on Student Achievement. *American Educational Research Journal, 50*(2), 397–428. https://doi.org/10.3102/0002831212469045

ten Braak, D., Lenes, R., Purpura, D. J., Schmitt, S. A., & Størksen, I. (2022). Why do early mathematics skills predict later mathematics and reading achievement? The role of executive function. *Journal of Experimental Child Psychology*, *214*, 105306.

Appendix

Table A1. Missing test scores and treatment

Missing test	(1) Grade 5	(2) Grade 5	(3) Grade 5	(4) Grade 8	(5) Baseline test
Treatment effect	-0.00584	-0.00391	0.000	-0.001	-0.033***
	(0.00657)	(0.00570)	(0.006)	(0.004)	(0.012)
Female		-0.00529	-0.001	-0.011**	-0.001
		(0.00480)	(0.005)	(0.004)	(0.004)
Parental edu. HS		-0.0470***	-0.052***	-0.022*	-0.025**
		(0.0132)	(0.013)	(0.012)	(0.011)
Parental edu. HE BSc		-0.0927***	-0.089***	-0.044***	-0.045***
		(0.0120)	(0.012)	(0.012)	(0.011)
Parental edu. HE MSc+		-0.118***	-0.115***	-0.048***	-0.046***
		(0.0130)	(0.013)	(0.012)	(0.011)
1st gen. immigrant		0.0599***	0.048***	0.063***	0.064***
		(0.0123)	(0.013)	(0.013)	(0.011)
2nd gen. immigrant		0.0462***	0.041***	0.009	0.013*
		(0.00995)	(0.010)	(0.009)	(0.008)
Cohort 2009		0.0881***	0.088***	0.003	0.004
		(0.00687)	(0.007)	(0.005)	(0.008)
Constant	0.0828***	0.122***	0.107***	0.085***	0.097***
	(0.00533)	(0.0135)	(0.013)	(0.012)	(0.016)
Sample	All	All	No strike, Consent	No strike, Consent	No strike, Consent
Observations	13,423	13,423	12,081	12,081	12,081
R-squared	0.017	0.065	0.063	0.016	0.100

Note: Each column represents the results from a separate OLS regression, comparing students at treatment and control schools while controlling for strata- and cohort fixed effects. Columns (3) to (5) exclude students in the 2009 cohort attending schools in strata affected by the teacher strike in 2022 as well as excluding students without the parental consent to merge baseline test scores with register data. Columns (2) to (5) include controls for background characteristics. Robust standard errors, adjusted for clustering at the school level, are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A2. Treatment effect on grade 5 test scores, with and without students affected by the strike (students in the 2009 cohort in three municipalities) for students with parental consent

	(1)	(2)
Sample	Consent	No strike, consent
Treatment effect	0.045	0.072**
	(0.027)	(0.031)
Female	-0.241***	-0.252***
	(0.016)	(0.018)
Parental edu. HS	0.061	0.087**
	(0.039)	(0.040)
Parental edu. HE BSc	0.424***	0.462***
	(0.038)	(0.038)
Parental edu. HE MSc+	0.729***	0.768***
	(0.037)	(0.040)
1st gen. immigrant	-0.177***	-0.177***
	(0.038)	(0.043)
2nd gen. immigrant	-0.228***	-0.222***
	(0.037)	(0.040)
Cohort 2009	-0.026	0.021
	(0.024)	(0.030)
Constant	-0.294***	-0.345***
	(0.044)	(0.044)
Observations	13,684	11,260
R-squared	0.129	0.134

Note: Each column represents the results from a separate OLS regression, comparing students at treatment and control schools while controlling for strata- and cohort fixed effects. Columns (3) to (5) exclude students in the 2009 cohort attending schools in strata affected by the teacher strike in 2022 as well as excluding students without the parental consent to merge baseline test scores with register data. Columns (2) to (5) include controls for background characteristics. Robust standard errors, adjusted for clustering at the school level, are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A3. Treatment effects on numeracy test scores. The sample is students with both baseline test score and parental consent, who were not affected by the strike.

National test	(1) Grade 5	(2) Grade 5	(3) Grade 8	(4) Grade 8
Treatment effect	0.072**	0.070**	0.015	0.013
	(0.035)	(0.031)	(0.032)	(0.026)
Covariates	No	Yes	No	Yes
Strata fixed effects	Yes	Yes	Yes	Yes
Observations	10,756	10,756	10,967	10,967
R-squared	0.039	0.132	0.039	0.148

Note: Each column represents the results from a separate OLS regression, comparing students in treatment and control schools while controlling for strata- and cohort fixed effects. Specification (2) and (4) include controls for background characteristics. Baseline ability is not included as a covariate. Robust standard errors, adjusted for clustering at the school level, are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A4. Treatment effects when applying entropy weights.

National test	(1) Grade 5	(2) Grade 8
Treatment effect	0.132***	0.082**
	(0.038)	(0.032)
Observations	10,226	10,482
R-squared	0.438	0.436

Note: OLS regression with robust standard errors adjusted for clustering on school in parentheses. Observations in the control group are weighted so that the control group is on average similar to the treatment group on background characteristics. *** p<0.01, ** p<0.05, * p<0.1.

Table A5. Treatment effect across time using all available tests.

Test	(1) Grade 2	(2) Grade 3	(3) Grade 5	(4) Grade 8
Treatment effect	0.194***	0.209***	0.133***	0.081***
	(0.060)	(0.041)	(0.035)	(0.030)
Baseline test (z-score)	0.577***	0.593***	0.587***	0.557***
	(0.020)	(0.013)	(0.013)	(0.012)
Baseline test missing (D)	-0.338***	-0.204***	-0.211***	-0.244***
	(0.118)	(0.075)	(0.055)	(0.046)
Cohorts	2009	Both	Both	Both
Covariates	Yes	Yes	Yes	Yes
Covariates interacted with treatment	No	No	No	No
Strata fixed effects	Yes	Yes	Yes	Yes
Observations	4,751	11,279	11,260	11,496
R-squared	0.412	0.408	0.421	0.420

Note: Each column represents the results from a separate OLS regression, comparing students in treatment and control schools. All regressions include strata and cohort fixed effects and controls for background characteristics and baseline test score. Robust standard errors, adjusted for clustering at the school level, are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1.