CESifo
AREA
CONFERENCES 2022

Economics of Education
Munich, 2 – 3 September 2022

Small Group Instruction to Improve Student Performance in Mathematics in Early Grades: Results from a Randomized Field Experiment

Hans Bonesrønning, Henning Finseraas, Ines Hardoy, Jon Marius Vaag Iversen, Ole Henning Nyhus, Vibeke Opheim, Kari Vea Salvanes, Astrid Marie Jorde Sandsør, and Pål Schøne

CESifo

Small Group Instruction to Improve Student Performance in Mathematics in Early Grades: Results from a Randomized Field Experiment[*]

Hans Bonesrønning, Henning Finseraas, Ines Hardoy, Jon Marius Vaag Iversen, Ole Henning Nyhus, Vibeke Opheim, Kari Vea Salvanes, Astrid Marie Jorde Sandsør, and Pål Schøne[†]

Abstract

We use an RCT to investigate whether small group instruction improves student performance in mathematics in the early grades. The RCT is large-scale, covering 159 Norwegian schools over four years. The students - 7-9 years old - are pulled out from their regular mathematics classes into small, homogenous groups for mathematics instruction for 3 to 4 hours per week, for two periods of 4-6 weeks per school year. Unlike many other recent tutoring experiments, all students are pulled out, not only struggling students. In our intention-to-treat analysis, we find that students in treatment schools increased their performance by .06 standard deviations in national tests 0.5 years after the intervention, with no differential effect by pre-ability level or gender. Our study is particularly relevant for policy-makers seeking to use additional teaching resources to target a heterogeneous student population efficiently.

Keywords: education economics; small group instruction; tutoring; tracking; class size; field experiment; intervention; randomized controlled trial; teacher-student ratio, mathematics instruction.

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

[*] We are grateful for comments from Colin Green, Simon Calmar Andersen, Diane Schanzenbach, participants at AERA, the LEER Workshop on Experimental Evidence in Education Economics and the Oslo workshop on Educational Resources and Student Performance, as well as comments from the Scientific Advisory Board appointed by the Research Council of Norway: Peter Fredriksson, Peter Blatchford, and Dorteh Bleses. We would like to thank Ester Bøckmann for excellent research assistance. This research is part of the 1+1 Project, supported by the Norwegian Research Council under Grant 256217.

[†] Corresponding author: Astrid Marie Jorde Sandsør, University of Oslo and Nordic Institute for Studies in Innovation, Research and Education (NIFU) (a.m.j.sandsor@isp.uio.no, +4797615822). Hans Bonesrønning: Norwegian University of Science and Technology (NTNU); Henning Finseraas: NTNU and Institute for Social Research (ISF); Ines Hardoy, Pål Schøne: ISF; Jon Marius Vaag Iversen, Ole Henning Nyhus: NTNU Social Research; Vibeke Opheim, Kari Vea Salvanes; NIFU. Shared first authorship with authors listed in alphabetical order.

1. Introduction

Student heterogeneity is a persistent and fundamental challenge in all school systems. For decades, smaller classes[1], more assistants[2], and special education have been the preferred solutions to improve educational achievement across ability groups. The evidence in favor of these policies is at best mixed, leading actors within the education sector and researchers to look for alternatives. One of the most prominent alternatives is tutoring – defined as one-on-one or small group instruction – which has been shown to substantially positively affect student learning (Dietrichson et al. 2017; Nickow et al. 2020). It has also emerged as a promising strategy for addressing COVID-related learning loss (Robinson & Loeb, 2021).

We present new evidence from an experiment of low-dosage tutoring in mathematics in a setting where additional teachers are used to provide small group instruction as an alternative to classroom-based teaching in the same subject for a shorter period of time. Tutoring is directed at students of all ability levels in mostly homogenous groups, allowing us to target the effect of a customized learning approach for all ability levels while holding instruction time fixed.

The experiment was conducted as a large-scale pre-registered randomized controlled trial (RCT) using additional teachers to tutor small groups of students during mathematics classes from 2016/17 to 2019/20. About 7,500 students aged 7–9 in 159 Norwegian elementary schools were each year pulled out from their regular mathematics classes for two periods of 4-6 weeks per school year to receive mathematics instruction in small groups of 4-6 students. To allow for tailoring of instruction, teachers were advised to construct small groups with students of similar ability levels in mathematics. From surveys, we know that most teachers chose this strategy.

[1] Leuven & Oosterbeeek (2018) and Schanzenbach (2020) provide recent reviews of the literature on class size.

[2] Finn & Achilles (1999), Muijs & Reynolds (2003), Blatchford et al. (2012) and Webster et al. (2013) find no beneficial effect from having teacher assistants whereas Andersen et al. (2020) report beneficial effects from teacher aides.

The Norwegian government made this field experiment possible by allocating around 20 million Euros to hire 80 qualified teacher person-years for four school years. Four cohorts of students born between 2008-2011 participated with variation in starting age and treatment length across cohorts. 78 treatment schools received funding to hire an additional teacher, while 81 schools served as the control group. About 30,000 students within ten local governments participated in the RCT. We closely follow the pre-registration plan published before gaining access to administrative data (Bonesrønning et al., 2018).

We find sizable average treatment effects on student performance. In our intention-to-treat analysis, we find that students in treatment schools increased their performance on national tests in mathematics by .06 standard deviations 0.5 years after the intervention. We also find that all student subgroups benefit from treatment, regardless of pre-ability level and gender.

Our paper adds to the literature on the effect of increased teacher-student (TS) ratio on student performance (e.g., Schanzenbach 2006; Angrist et al. 2019; Hoxby 2000; Browning and Heinesen 2007; Fredriksson and Öckert 2008; Leuven et al. 2008; Iversen and Bonesrønning 2013). While the evidence is mixed (see Leuven & Oosterbeeek (2018) and Schanzenbach (2020) for recent reviews), previous research has shown no or small effects in the resource rich Norwegian context (Leuven et al. 2008; Iversen & Bonesrønning 2013; Falch et al. 2017; Leuven & Løkken 2018; Haaland et al. 2021; Borgen et al. 2022). Most of this literature investigates the impact of increased TS ratio through reduced class size, suggesting that more flexible approaches to increasing the TS ratio may be key (Solheim & Opheim 2018). Alternative strategies to reduce the TS ratio include adding additional teachers to the classroom. A recent paper by Haaland et al. (2021) finds that additional teachers in literacy instruction only yield positive effects in combination with teacher professional development. In contrast, our study suggests that using extra teachers to provide low-dose tutoring at young ages yields positive effects for all students.[3] This is a more flexible and potentially less costly way of reducing the TS ratio, as it allows schools to target subjects or students needing additional support.

As such, our paper also adds to the literature on tutoring. A review by Nickow et al. (2020) shows that tutoring programs yield consistent and substantial effects on learning outcomes, typically in the area of .30-.40 of a standard deviation. Further, a recent meta-analysis by Dietrichson et al. (2017) found tutoring to be both the most common and most effective intervention to improve the educational achievement for low socioeconomic status students.[4] However, the reviewed tutoring programs are typically high dosage, targeted at low-ability students, and in many cases may entail increased instruction time–replacing recreational activities, unfilled time, or potentially crowding out instruction time in other subjects. Little is known about the performance of the type of low-dosage tutoring investigated in this paper, where instruction time in the subject is held fixed. Such knowledge is in high demand from policy-makers since they are less costly to implement at full scale.

Finally, our paper adds to the literature on ability grouping, as small groups were largely comprised of students of similar ability levels in mathematics. A scarce literature credibly identifies the impact of tracking within schools on student outcomes. Duflo et al. (2011) show that within school ability tracking in a developing country (Kenya) benefits all students. However, as they note in their paper, this is not necessarily directly transferable to countries with different educational contexts, teacher incentives, and distribution of student ability. Similarly, Zimmer (2003) finds that within school tracking is beneficial for lower achieving students in the US, suggesting that tailored instruction outweighed any potential adverse effects from low-ability students losing their high-ability peers, although e.g. Matthewes (2021) finds

[3] Both studies were funded by the Research Council of Norway to implement a randomized controlled trial on the effect of additional teachers on student performance.

[4] Recent papers that evaluate different tutoring programs include e.g. Gersten et al. (2015), Fryer (2014), Dobbie and Fryer (2013), Fryer (2017) and Fryer and Howard-Noveck (2020).

the opposite for Germany where between-school tracking is harmful for low-achieving students. This paper measures the impact of a less comprehensive form of tracking with ability grouping in one subject only for a limited period of time, implying less of an impact from peer effects than more comprehensive forms of tracking. Our results, with beneficial effects across all student ability levels, suggest that the impact of customized instruction may be an important mechanism through which ability grouping can increase student outcomes.

The rest of the paper is organized as follows: The institutional context and intervention are presented in section 2, while section 3 discusses the randomization process, data, and balance. Section 4 presents the empirical specification, whereas the estimated treatment effects of the small-group instruction are presented in section 5. Finally, section 6 offers some concluding remarks and discusses our results and previous findings in the literature.

2. Institutional context and the intervention

a. Institutional context

Compulsory education is free of charge, and less than 4 percent of students attend private schools. The public sector at the municipal level is responsible for providing compulsory education. There are three stages: lower primary education, grades 1-4 (ages 6-10); upper primary education, grades 5-7 (ages 10-13) and lower secondary education, grades 8-10 (ages 13-16). Compulsory education is comprehensive with a common curriculum for all students, and there is no tracking. The grade cutoff date is January 1, and grade promotion or retention is very uncommon, ensuring that nearly all students follow their cohort and graduate from lower secondary school the year they turn 16. The school year lasts from August to June, from about 8:30 to 1:30. All children in grades 1-4 are entitled to enroll in voluntary before/after school programs, with most children enrolling particularly for the lowest grades. Enrollment in after school programs has increased in recent years due to an increase in subsidies to cover parental fees.

About 5 percent of students in grades 1-4 were eligible for special education in 2017. 37% of the special education students received assistance in their regular classes, the rest were taught alone or in small groups of eligible students.

While the Norwegian Education act allows for small group instruction, results from our teacher survey indicate that there was no wide-spread use of small group instruction in Norwegian primary schools. During the intervention, when asked about the number of students that participated in small group instruction during the previous mathematics lesson, teachers at treatment schools reported an average of 3.78 students whereas the corresponding result for control schools was 0.37 – likely reflecting special needs students receiving assistance outside of the regular class.

b. Treatment description

School leaders in the intervention schools were allocated an additional teacher person-year in the school years 2016/17-2019/20, which they were instructed to use for small group tutoring in mathematics in specific grades. Due to the combination of in-school delivery and a pull-out strategy, the design of the intervention had to comply with the national legislation for public elementary schools. First, permanent tracking is not allowed, but small homogenous student groups can be pulled out of their regular class for shorter periods. It was accepted that six weeks is within the limit for a short period. Second, the treatment dosage is determined by legislation saying that the students will be taught mathematics for 560 hours during grades 1-4, or on average 140 hours per year, implying a planned dosage for treated students of minimum 30 hours (1800 minutes) of small group instruction per year. The sessions differed in length, as there are local variations in the schools' organization of the regular mathematics instruction. While some schools have long sessions (up to 90 minutes), others have shorter sessions, often 60 or 45 minutes, but always adding up to 140 hours per year. Instruction was given in parallel to all regular mathematics classes. See the online appendix A or the pre-analysis plan (Bonesrønning et al. 2018) for further details on the intervention.

From a registration form sent to small group instructors[5] we have information on the number of minutes spent in small group instruction for all students, excluding time used on breaks. Calculating the averages for students in the treatment group by cohort and treatment year, we know that the average small group consisted of about 5 students and that students received about 8 weeks of small group instruction per treatment year–amounting to between 1075 to 1184 minutes of instruction time–depending on the cohort and treatment year. This amounts to between 60 and 66 percent of the planned minimum treatment.[6]

National legislation requires that teachers are formally qualified to teach mathematics at the elementary level so that only formally qualified teachers are hired. From a teacher survey we have some information on how small group instructors were recruited as well as information on characteristics of small group instructors and regular teachers.[7] 31% of small group instructors had previously worked at the same school, meaning that the majority were externally recruited. Compared to regular mathematics teachers, a larger fraction was male 28% compared to 13% for regular teachers, they were on average 40 years old compared to 42 for regular teachers, had 12 years of teaching experience compared to 19 years for regular teachers, and had more credits in mathematics, 58 credits compared to 37 for regular teachers, which is equivalent to about 2/3 of a semester in higher education. There was no difference in the share that had completed teacher education, which was 98% for both groups.[8]

The small group teachers received no training as tutors, but they (together with the regular teachers) received a handbook including detailed instructions on how to implement the intervention–i.e small group size, duration etc.–information on data collection as well as

[5] See online appendix B and online appendix H, Table A5 for details on data and implementation.

[6] Note that planned and received treatment may not be directly comparable as received treatment deducts time spent on breaks.

[7] See online appendix H for details.

[8] The teachers survey also provides some information on how the class teacher experience the small group intervention: Teachers were asked to rate, on a likert scale, where 1 is strongly disagree and 5 corresponds to strongly agree, the following statement: “If a group of students participate in small group instruction, I (the class teacher) am able to follow up the students much better”. The average score is 4.4, which indicates that teachers agree or highly agree with this statement.

recommendations based on previous research. The handbook contained information about the characteristics of previous successful interventions using additional teachers and, importantly, encouraged the teachers to create small groups with students of similar mathematical abilities.[9] Based on survey data from small group instructors we know that the majority of small group instructors followed this recommendation as 97% reported that they agreed or strongly agreed with the statement that "Small groups were composed of students of nearly equal ability level in mathematics".

One birth cohort (2010) was treated only in 4th grade (2019/20). The cohorts 2008 and 2011 were treated for two years, starting in 3rd grade (2016/17) and 2nd grade (2018/19), respectively. Those born in 2009 were treated for three years, starting in 2nd grade (2016/17). In this paper, we mainly restrict the analysis to cohorts for which we have data on the national tests in 5th grade, i.e., the 2008 and 2009 cohorts. These are also the only two cohorts unaffected by the Covid-19 pandemic when completing the national tests.

Throughout the project, small group teachers reported which students received small group instruction and the instruction length for each session. Additionally, the project group met with small group teachers and school leaders yearly, all teachers and school leaders received yearly surveys, and visits were carried out at some treatment schools. Together, this allowed us to follow implementation closely and quickly detect whether schools were having any problems with implementation due to e.g. misunderstandings, teacher absence, or teacher turnover. The school visits comprised classroom observation, interviews with school principals, as well as interviews with math teachers (both the main teachers and small group teachers). An important finding was that small group instruction generally was much appreciated (Bubikova-Moan & Opheim 2020).

[9] For further details on the content of the handbook se online appendix B.

3. Randomization, data, and balance

a. Randomization

Randomization was carried out at the school level within each of the ten municipalities participating in the project.[10] We randomized at the school level to avoid resistance from schools and parents due to similar students being treated differently within schools. Also, school-level randomization ensured that the control group was less likely to be affected by the treatment through spill-over effects.

We conducted stratified randomization in the following manner: Schools with at least 20 students per grade were eligible to participate within each municipality. We ranked the schools based on their mean test score in the 5th-grade national test in mathematics, averaging over the mean score in the two preceding school years to reduce measurement error. Next, we constructed a set of strata of at least four schools in each stratum. In doing so, we follow Imbens' (2011) recommendation to have at least two treatment and control schools in each stratum to derive a within-strata variance in the treatment effect. Most strata consist of four or six schools. We randomized schools to the treatment or the control group by using a random number generator. One school refused to participate after their treatment status was revealed. Following the pre-analysis plan, we exclude all schools in the respective strata.

All treatment schools received one additional teacher person-year regardless of cohort size. This implied that the smallest schools in our sample have a larger increase in the student-teacher ratio than the larger schools, with about 70 students in each grade. Additionally, as larger schools would not obtain sufficient treatment intensity for all students, we randomized classes or groups to treatment at these schools.

[10] The ten municipalities are geographically spread from the southern to the northern part of Norway, all fairly densely populated.

b. Data

The main data source is administrative data collected and organized by Statistics Norway. We have background information about the students and test scores from the national tests in 5th grade from administrative registers (see online appendix for details). We use this data to identify the main treatment effects and to assess balance across treatment and control groups. In addition, we analyze pre-test and post-test data collected by the project. We developed math tests in collaboration with teachers and math educators. For most cohorts, the pre-tests were conducted late in the school year prior to entering the project.[11] The post-tests were conducted at the end of the school year (May-June). We use this data to identify short-term treatment effects at a younger age than the national tests and to examine treatment heterogeneity on baseline test scores.

A small percentage of students have no reported test score on the national test. We find no evidence of a correlation between missing test scores and treatment status (see online appendix Table A1). This is important since it indicates that missing test scores will not bias our results and will have a negligible impact on statistical power. In the online appendix (Table A3), we also show that there is no important treatment-control difference in geographic mobility, measured as whether they completed the national test in another school than the baseline test.

c. Balance tests

Following the pre-analysis plan, we study balance on gender, parental level of education, the share of first or second-generation immigrants, and school size (see online appendix for details on background variables).[12] Table 1 shows that treatment and control schools are balanced across these variables, except for a slightly higher share of students in the treatment group with

[11] The exception is the first year of the project (the 2016/2017 school year), for which we did the pre-tests early in the school year (August).

[12] The pre-analysis plan says that we will study balance on the teacher-student ratio as well, but we have been unable to obtain that information broken down by cohort and school class.

parents in the highest education level category. Reassuringly, the F-test of joint significance produces a large p-value of .41. We therefore conclude that randomization was successful.

Table 1. Balance test.

		Control		Treatment		Difference
		N/[Schools]	Mean/SE	N/[Schools]	Mean/SE	(1)-(2)
Female		8128	0.481	8148	0.488	-0.007
		[81]	(0.006)	[78]	(0.007)	
Parental edu:	Primary	8128	0.055	8148	0.054	0.001
		[81]	(0.007)	[78]	(0.007)	
Parental edu:	Secondary	8128	0.213	8148	0.196	0.017
		[81]	(0.012)	[78]	(0.013)	
Parental edu:	College, low	8128	0.390	8148	0.373	0.017
		[81]	(0.009)	[78]	(0.009)	
Parental edu:	College, high	8128	0.308	8148	0.339	-0.031*
		[81]	(0.019)	[78]	(0.019)	
Parental edu:	Missing	8128	0.035	8148	0.039	-0.004
		[81]	(0.003)	[78]	(0.004)	
Foreign-born		8128	0.063	8148	0.064	-0.000
		[81]	(0.005)	[78]	(0.004)	
Second generation		8128	0.100	8148	0.101	-0.002
		[81]	(0.011)	[78]	(0.013)	
School size		8128	56.615	8148	58.579	-1.964
		[81]	(2.153)	[78]	(2.238)	
F-stat joint significance, p-value						1.04, .41

Notes: Standard errors are clustered at school. Strata and cohort FE are included in all estimations. *** p<0.01, ** p<0.05, * p<0.1

4. Empirical specification

We identify the intention-to-treat (ITT) effects using the following regression models:

$$y_i = \beta TREATED_g + \alpha_s + \mu_c + X_i'\gamma + \epsilon_i$$

where *i* indexes individuals, *g* schools, *s* randomization strata, and *c* cohorts. *y* is the test score and *TREATED* is a binary indicator of whether the student was enrolled in a school in the treatment group when entering the project. We define all students in a treatment school as treated despite randomizing classes or groups to treatment or control in larger schools. This is due to potential spill-over effects from the treated classes and because schools might have changed the class compositions in response to the class randomization. Thus, our classification ensures that β is the cleanest ITT estimate, although likely representing a lower bound estimate of the treatment effect. Because randomization was performed within strata, we include strata fixed effects α. Cohort fixed effects, μ, and a vector X with socio-economic background variables, are included to improve statistical power. Standard errors are adjusted for clustering at the school level, the level of treatment assignment and delivery.

5. Treatment effects

This section presents the estimated treatment effects. Section *a* presents treatment effects on our main outcome, test scores on a national test in mathematics in 5th grade, while section *b* discusses effects on national tests in reading and English. Section *c* supplements the estimated treatment effects from section *a* with analyses of test scores on own tests in mathematics carried out at the end of the treated school years. Treatment effect heterogeneity is analyzed in section *d*.

a. Medium-term effects – national test scores in mathematics

The main intention to treat (ITT) estimates are presented in Table 2. The first column is without individual level controls, while the second includes the vector of controls used in the balance tests. Without controls, we find that students in the treatment schools increase their performance by .066 standard deviations relative to students in the control group.[13] When we add SES controls, the estimate declines to .058 standard deviations. For comparison, we find that students with a university-educated father perform about .14 standard deviations better than other students. Thus, the effect amounts to about one-third of the education difference. Our estimates are in-between the high-dosage (.31) and small-dosage (.015) treatment estimates in Fryer (2017).

[13] To rule out that any treatment effects are driven by researchers' interactions with several treatment schools visited during the intervention period, we have run a specification check where we re-run the estimation in column (1) on a sample excluding the 14 strata containing schools visited. Reassuringly, the results are unaltered—for results see online appendix Table A1.

Table 2. Baseline results. Dependent variable is standardized national test scores.

	(1)	(2)
	Mathematics	
Treatment school	.066** (.031)	.058** (.026)
Observations	14,891	14,891
Strata FE	Yes	Yes
Cohort FE	Yes	Yes
SES controls	No	Yes
RI p-value	.05	.05
IWE	.067** (.031)	.057** (.027)

Note: OLS regression with robust standard errors adjusted for clustering on school in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Our conclusions are robust to using randomization inference (RI) to derive p-values (Imbens and Rubin, 2015; Hess 2017), which is reassuring since RI avoids assumptions regarding resampling, the parametric distribution of t-values, and is valid irrespective of the sample size. It is potentially useful to avoid these assumptions since the intervention only involves 159 schools, which might imply that asymptotic characteristics do not apply.

The ITT estimate using conventional fixed effects models can be misleading if there is important treatment heterogeneity (Gibbons et al., 2018), as such models place more weight on averages from the groups (in our case strata) with the most within-group variance. This does not seem to be a problem in our case, as the treatment effect estimates are identical if we follow Gibbons et al. (2018) and interact the treatment indicator with the strata fixed effects and derive the average treatment effect from these interaction terms.

b. Effects on national test scores in reading and English

Table A4 in the online appendix presents the ITT estimates on national test scores in 5th-grade reading and English. These outcomes are not true placebo outcomes since there might be spill-overs from small group instructions in mathematics, e.g., from cognitive development or improved motivation for school work. However, the intervention aims to improve skills in Mathematics, so we should not expect similar-sized treatment effects on these outcomes. For English, the ITT is essentially zero, while the ITT for reading is .029, less than half of the effect on mathematics. The difference between the ITT for math and reading is, however, not statistically significant.

c. Short-term effects

Next, we use our own pre- and post-tests to estimate short-term effects. These short-term estimates are useful because we can examine whether the treatment effect increases or declines with time since treatment. However, the interpretation of the ITT effects on the post-test scores is complicated by a lower test completion rate in the control group. The share of missing test scores is about six percentage points lower in the treatment group on average across cohorts (see online appendix Table A2). The treatment-control difference in completion likely reflects lower teacher motivation in the comparison schools to carry out additional testing for students that missed the first test due to absence.

In Table 3, we analyze post-test scores for the 2008 and 2009 cohorts at the end of third grade, and the 2011 cohort at the end of second grade. We include the 2011 cohort since comparisons across cohorts provide information on the importance of length and timing of treatment. When our tests were completed, the 2009 cohort had been treated for two years (second and third grade), whereas the 2008 and 2011 cohorts had been treated for one year (respectively in third and second grade). When we pool data from all cohorts, we find a treatment effect of .158, which is about three times larger than the treatment effect on the national tests.[14] The treatment effects are quite similar across cohorts, despite differences in age, years of treatment, and teacher experience in small group instructions. Thus, we find no substantial benefits from being treated for two years compared to one year.

[14] The estimates in Table 3 are precisely estimated, but due to the difference in missing test scores between treatment and control schools they do not accurately reflect the uncertainty in the treatment effect estimate. Therefore we also estimate so-called Lee trimming bounds on the treatment effects (Lee, 2009), which suggest that the pooled treatment effect is between .04 and .30 for the Always-Reporters.

Table 3. Short-term effects. Dependent variable is standardized score from project tests.

	Pooled	Cohort 2008	Cohort 2009	Cohort 2011
Treatment school	.158*** (.031)	.144*** (.049)	.169*** (.046)	.164*** (.051)
Observations	21,983	7,790	7,179	7,014
Strata FE	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes
SES controls	No	No	No	No
Years treatment		1	2	1
Test grade		3	3	2

Note: Robust standard errors adjusted for clustering on school in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

d. Treatment effect heterogeneity

In this final section, we study treatment heterogeneity. First, we study heterogeneity on the national test score across cohorts and gender. In the first column in Table 4 we present results when we include an interaction term between an indicator for the 2009 cohort and the treatment indicator. This interaction term is negative and indicates that the 2008 cohort drives the treatment effect in Table 2. This result is unexpected since the 2009 cohort was treated longer and from a younger age. The difference might reflect extraordinary motivation among teachers at the beginning of the project that decreased over time (Dietrichson et al., 2017). However, the interaction term is not statistically significant, so we cannot rule out that the effect is the same for both cohorts.

The second column in Table 4 shows a large gender gap in the test score, as male students perform much better on the national test. The intervention appears to reduce this gap slightly since the treatment effect is larger for female students. However, the treatment effect difference across gender is not statistically significant.

Table 4. Cohort-specific effects. Dependent variable is standardized national test score.

	Cohorts 2008 & 2009	Gender
Treatment	.073** (.036)	.046 (.029)
Treatment x 2009-cohort	-.031 (.045)	
2009-cohort	-.009 (.031)	
Treatment x Female		.024 (.030)
Female		-.251*** (.021)

Observations	14,891	14,891
Strata FE	Yes	Yes
Cohort FE	Yes	Yes
SES controls	Yes	Yes

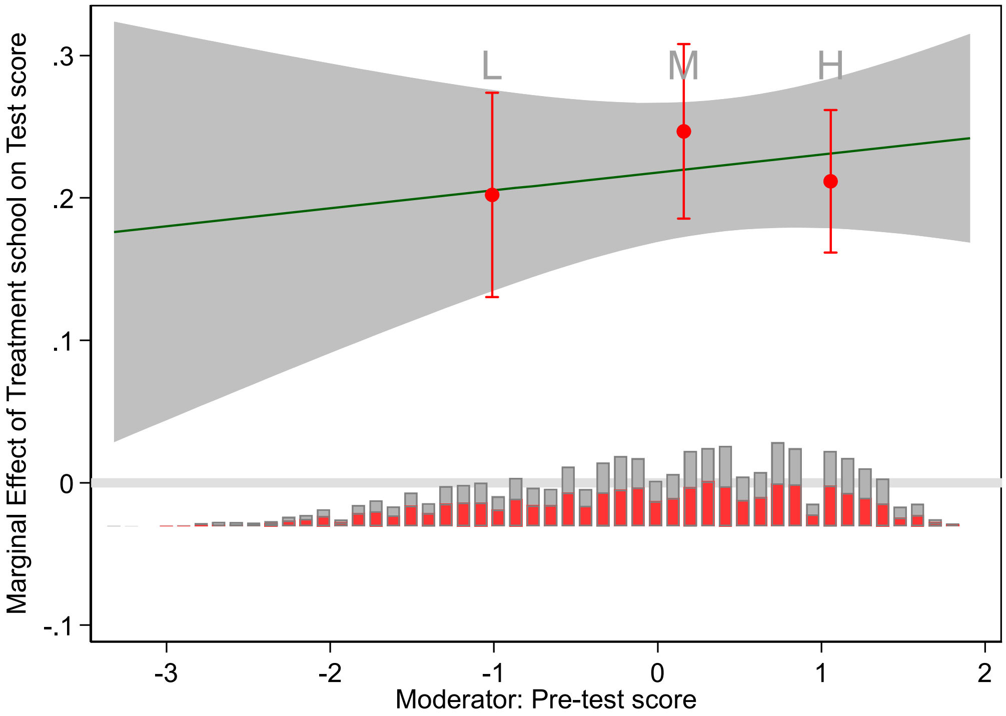
Note: OLS regression with robust standard errors adjusted for clustering on school in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Next, we use our own pre- and post-tests to study treatment heterogeneity depending on *i*) baseline ability, *ii*) average baseline score of the school, and *iii*) within school heterogeneity in baseline test scores. The test of treatment heterogeneity by average pre-test score in the school and within-school heterogeneity was not pre-registered and should be considered as exploratory. To examine heterogeneity on baseline ability, we interact the baseline test score with the treatment indicator. As mentioned above, there is a difference between treatment and control schools in the share of students that conducted the test. To reduce the bias from selection to the test, we follow the pre-registration plan and conduct entropy balancing (Hainmueller, 2012) to reweight the sample so that the treatment-control difference in the baseline test score is zero.

Figure 1 shows a positive correlation between the treatment effect and baseline test score, but the interaction term is not statistically significant (coeff = .01, p=.51). 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.[15] These estimates indicate that there is a weak non-linearity in the marginal effects. The treatment effect is slightly larger for the mid-level achievers on the baseline test.

[15] See Hainmueller et al. (2019) for details.

Figure 1. Treatment heterogeneity by pre-test score.



Note: The plot 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 pre-test score) in the pooled sample, and the red and white shaded bars refer to the distributions in the treatment and control groups, respectively.

The online appendix presents treatment effects across average baseline scores and within-school heterogeneity. Figure A.1 is based on a regression model with an interaction between the treatment effect and the mean test score of the school, controlling for the individual level test score. We find that the marginal effect of treatment declines with school test scores in the linear model (p=.06). However, the linear model does not seem like the most appropriate specification since the estimated treatment effect is much larger for schools in the mid-range of the pre-test score distribution, as indicated by the point estimate for the medium group (in red). This result suggests that when compared to schools with medium average baseline test scores, schools with respectively low and high average baseline scores are somewhat less able to utilize the benefits of the treatment. Schools with high average baseline scores might also face ceiling effects.

In Figure A.2, we interact the treatment indicator with the school's standard deviation of the baseline test score. Here we find that the linear model produces flat marginal treatment effects. We again see that schools in the middle of the distribution perform slightly better, but the differences across the bins are not significant. Thus, there is no evidence that the intervention has larger effects in heterogeneous schools where small homogenous groups would represent a stronger deviation from the normal situation.

Perhaps the most surprising finding from the heterogeneity analyses is that low-performing students seem to benefit as much from the intervention as high-performing ones. Guryan et al., (2021) in a recent paper provide evidence that individualization of instruction can explain much of the benefits from tutoring for struggling students. According to Duflo et al. (2011) who report from a tracking experiment in Kenya, such findings most likely reflect that the teachers successfully tailor their instruction to the students at hand.

In future work we will investigate whether the treatment effects are due to tutors who tailor their instruction to the average ability level in the small groups, and/or whether the tutors take advantage of the small group to provide individualization of instruction within the small groups.

Conclusion

Our results show that low-dosage tutoring in mathematics for primary school students, can increase learning outcomes for students of all ability levels, even without increasing instruction time. We find sizable effects on performance in mathematics. Treatment schools score on average .16 standard deviations better than control schools after completing a school year with tutoring (a short-term effect). However, this effect drops to .06 standard deviations on the national test (a longer-term effect).

It is also important to address the intervention's effectiveness compared to costs. The descriptive statistics in Table A5 in the online appendix suggest that every student received a unique treatment equivalent to 222 minutes yearly. On average, the students were treated for

2.5 years. This constitutes 1.25 percent of a teacher's person-year. The unit cost for a teacher person-year was NOK 705,000 in 2017, resulting in a total per student cost equal to NOK 8,800 or 1,064 USD. Following this approach, the intervention resulted in an ITT effect of around .056SD per 1000 USD. However, supported by findings in Section 5b, treatment might yield as much as .14SD per 1,000 USD, given that effects are similar for 1 as 2.5 years of treatment. This implies that our intervention is slightly more effective than preliminary findings from a small group instruction intervention targeting low-performing 8th graders in Norway (Kirkebøen et al. 2019) evaluated by the conservative estimate of 2.5 years duration. The effect-cost ratio is quite similar to those found in Andersen et al. (2020) evaluating extra teacher's aides in Demark (.076-.11SD per 1000 USD) and Guryan et al. (2021) evaluating the Saga tutoring program in the US, whereas the yield is significantly higher than what Schanzenbach (2006) estimates for Project STAR.

The effect sizes are smaller than those in the high-dosage literature but larger than those found in previous low-dosage experiments (Fryer, 2017; Nickow et al., 2020). Limited to experiments with young students and mathematics, Smith et al. (2013) and Gersten et al. (2015) report much stronger effects than we do for young struggling students. A recent meta-analysis on tutoring (Nickow et al., 2020) shows larger positive effects than reported here, typically around .30-.40 standard deviations. The majority of the included programs are relatively high dosage and aimed at low-ability students. Tutoring typically lasts between 10 weeks and a school year, involves one-on-one tutoring and is catered for students who performed at or below a given threshold. A weakness in much of the literature is that it is unclear what activities students would have engaged in had they not been tutored – implying that increased instruction time is a potential confounding factor. Increased instruction time could either replace recreational activities, other unfilled time or crowd out instruction time in other subjects. In our study, instruction time is held constant by design.

These findings add to the tutoring and tracking literature by showing that a pull-out strategy using small homogenous groups in mathematics while keeping instruction time constant can benefit all students. It is also worth noting that we find effects of additional teacher resources on student performance in a resource rich context where previous research has shown no or small effects of reduced student-teacher ratio (Leuven et al., 2008; Iversen & Bonesrønning, 2013; Falch et al., 2017; Leuven & Løkken 2018; Haaland et al., 2021, Borgen et al., 2021). This makes our study particularly relevant for policy-makers seeking additional teaching resources to target a heterogeneous student population efficiently.

References

Andersen, S. C., Beuchert, L., Nielsen, H. S., & 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. https://doi.org/10.1093/jeea/jvy048

Angrist, J. D., Lavy, V., Leder-Luis, J., & Shany, A. (2019). Maimonides' Rule Redux. *American Economic Review: Insights*, *1*(3), 309-24. https://doi.org/10.1257/aeri.20180120

Betts, J. R. (2011). The Economics of Tracking in Education. In E. A. Hanushek, S. Machin & L. Woessmann (Eds.), *Handbook of the Economics of Education* (pp. 341-381). Vol. 3 Amsterdam: North Holland.

Blatchford, P., Russell A., and Webster R. (2012). *Reassessing the Impact of Teaching Assistants: How Research Challenges Practice and Policy*. New York: Routledge. https://doi.org/10.4324/9780203151969

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. https://doi.org/10.17605/OSF.IO/YWQVC

Borgen, N. T., Kirkebøen, L. J., Kotsadam, A., & Raaum, O. (2021). Do funds for more teachers improve student performance? CESifo Area Conferences 2021, Economics of Education. Working paper.

Browning, M., & Heinesen, E. (2007). Class Size, Teacher Hours and Educational Attainment. *Scandinavian Journal of Economics, 109*(2): 415-438. https://doi.org/10.1111/j.1467-9442.2007.00492.x

Bubikova-Moan, J. & Opheim, V. (2020): 'It's a jigsaw puzzle and a challenge': critical perspectives on the enactment of an RCT on small-group tuition in mathematics in Norwegian lower-elementary schools*, Journal of Education Policy*, https://doi.org/10.1080/02680939.2020.1856931

Dietrichson et al. (2017)…

Dobbie, W., & Fryer Jr, R. G. (2013). Getting Beneath the Veil of Effective Schools: Evidence from New York City. *American Economic Journal: Applied Economics, 5*(4): 28-60. https://doi.org/10.1257/app.5.4.28

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. https://doi.org/10.1257/aer.101.5.1739

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

Finn, J. D., & Achilles, C. M. (1999). Tennessee's Class Size Study: Findings, Implications, Misconceptions. *Educational Evaluation and Policy Analysis, 21*(2): 97-109. https://doi.org/10.3102/01623737021002097

Fredriksson, P., & Öckert, B. (2008). Resources and Student Achievement: Evidence from a Swedish Policy Reform. *Scandinavian Journal of Economics, 110*(2): 277-296. https://doi.org/10.1111/j.1467-9442.2008.00538.x

Fryer Jr, R. G. (2014). Injecting Charter School Best Practices into Traditional Public Schools: Evidence from Field Experiments. *Quarterly Journal of Economics 129*(3): 1355-1407. https://doi.org/10.1093/qje/qju011

Fryer Jr, R. G. (2017). The Production of Human Capital in Developed Countries: Evidence from 196 Randomized Field Experiments. In E. Duflo, & A. Banerjee (Eds.), *Handbook of Field Experiments (*pp. 95-322). Vol. 2 Amsterdam: North-Holland. https://doi.org/10.3386/w22130

Fryer Jr, R. G., & Howard-Noveck, M. (2020). High-Dosage Tutoring and Reading Achievement: Evidence from New York City. *Journal of Labor Economics, 38*(2): 421-452. https://doi.org/10.1086/705882

Gersten, R., Rolfhus, E., Clarke, B., Decker, L. E., Wilkins, C., & Dimino, J. (2015). Intervention for First Graders With Limited Number Knowledge: Large-Scale Replication of a Randomized Controlled Trial. *American Educational Research Journal, 52*(3): 516-546. https://doi.org/10.3102/0002831214565787

Gibbons, C. E., Serrato, J. C. S., & Urbancic, M. B. (2019). Broken or Fixed Effects? *Journal of Econometric Methods 8*(1): 1-12. https://doi.org/10.1515/jem-2017-0002

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.* https://doi.org/10.3386/w28531

Haaland, V. F., Rege, M., & Solheim, O. J. (2021). Do Students Learn More with an Additional Teacher in the Classroom? Evidence from a Field Experiment. *mimeo*

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. https://doi.org/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. https://doi.org/10.1017/pan.2018.46

Hess, S. (2017). Randomization inference with Stata: A guide and software. *Stata Journal, 17*(3): 630-51. https://doi.org/10.1177/1536867X1701700306

Hoxby, C. M. (2000). The Effects of Class Size on Student Achievement: New Evidence from Population Variation. *Quarterly Journal of Economics, 115*(4): 1239-1285. https://doi.org/10.1162/003355300555060

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

Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference for Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press. https://doi.org/10.1017/CBO9781139025751

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

Lee, D. S. (2009). Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. *Review of Economic Studies, 76*(3): 1071-1102. https://doi.org/10.1111/j.1467-937X.2009.00536.x

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, Research Department.

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

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): 663-693. https://doi.org/10.1111/j.1467-9442.2008.00556.x

Leuven, E., & Oosterbeek, H. (2018). Class size and student outcomes in Europe. *EENEE, Analytischer Bericht*, (33).

Matthewes (2021).

Muijs, D., & Reynolds, D. (2003). The Effectiveness of the Use of Learning Support Assistants in Improving the Mathematics Achievement of Low Achieving Pupils in

Primary School. *Educational Research, 45*(3): 219-230. https://doi.org/10.1080/0013188032000137229

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.* https://doi.org/10.3386/w27476

Robinson & Loeb, 2021…

Schanzenbach, D. W. (2006). What Have Researchers Learned From Project STAR? *Brookings Papers on Education Policy,* (9): 205-228. https://doi.org/10.1353/pep.2007.0007

Schanzenbach, D. W. (2020). The economics of class size. In S. Bradley & C. Green (Eds.), *The Economics of Education (Second Edition)* (pp. 321-331). Academic Press. https://doi.org/10.1016/B978-0-12-815391-8.00023-9

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

Webster, R., Blatchford, P., & Russell, A. (2013). Challenging and Changing How Schools Use Teaching Assistants: Findings from the Effective Deployment of Teaching Assistants Project. *School Leadership & Management, 33*(1): 78-96. https://doi.org/10.1080/13632434.2012.724672

Online appendix material

Small Group Instruction to Improve Student Performance in Mathematics in Early Grades: Results from a Randomized Field Experiment

A: Treatment[16]

Treated schools were allocated an additional teacher man-year with the instruction of using it for small-group tutoring in mathematics in grades specified by the project (see Table A1). To ease the organization of the school timetable, schools were allowed to use the additional resource to recruit (no more than) two small group instructors—implying that some teachers divided their time between small group instruction and other teaching responsibilities at the school.

Schools were instructed that students of all ability levels, not only struggling students, should participate. The only exception being special needs students–in the case where small group instruction would interfere with their rights to special needs education. It was left to the schools to make decisions on the participation of these students on a case-by-case basis, making sure that their rights remained intact.

Two cohorts were treated each school year. Schools were instructed that as many students as possible, in each grade level, should receive at least two periods of small-group instruction during each school year, with each period lasting 4-6 weeks. Implying that to the extent possible–all students should receive the same number of hours of tutoring. If it was not feasible to give all students this target dosage, schools were instructed that instead of reducing the duration of small group instruction for all students, one group should receive a little less while the rest should receive the target dosage.

Table A1 shows the planned use of one teacher-man yea during the intervention period.

[16] This part is based the treatment description in the pre-analysis plan (see Boesrønning et al. 2018).

Table A1: Project plan for use of one teacher man-year for four years (T=treatment)

School year: **Cohort*:**	**2016/17**	**2017/18**	**2018/19**	**2019/20**
2008	T (grade 3)	T (grade 4)		
2009	T (grade 2)	T (grade 3)	T (grade 4)	
2010				T (grade 4)
2011			T (grade 2)	T (grade 3)

*Cohort refers to the birth cohort.

In the first year the 2nd and 3rd grades were treated by the additional teacher man-year. The second year, the 3rd and 4th grades were treated. In the third year, 2nd and 4th grades were treated, and in the fourth year, 3rd and 4th grades were treated. Thus, according to the initial plan, after four years, one cohort should have been treated for one year (4th grade in the final year of the intervention), two cohorts should have been treated for two years (starting in the 3rd grade year 1, and starting in the 2nd grade year 3), and one cohort should have been treated for three years (starting in the 2nd grade year 1).

Given that the dosage was intended to be the same in each treatment year–the expected treatment dosage between cohorts differs by the number of treatment years.[17] The expected treatment dosage per year is determined by the minimum treatment requirement, stating that each student should receive at least two periods of small group instruction, with each period lasting a minimum of 4 weeks. Schools were instructed to give small group instruction in parallel to all regular mathematics classes. There is legislation in place dictating that students are to receive mathematics instruction for 560 hours during grades 1-4, or on average 140 hours per year. This implies that treated students should receive instruction in small groups for a minimum of 30 hours (1800 minutes) per year[18]. The sessions differed in length, as there are local variations in the schools' organization of the regular mathematics instruction. While some schools have long sessions (up to 90 minutes), others have shorter sessions, often 60 or 45 minutes, but always adding up to 140 hours per year.

Due to differences in cohort size between schools, we limited the number of included students for each cohort at each school. In small or medium sized schools all student at each grade level could be included. In large schools (more than 48 students or more than two

[17] In total four cohorts were included in the intervention–with the intended treatment duration ranging from 1 to 3 years. However, due to the pandemic, small group instruction was severely disrupted in the last months of the intervention – implying that the 2010 and 2011 cohorts received less treatment than we originally planned for. In this article only the 2008 and 2009 cohorts are included in the analysis.

[18] A school year lasts for 38 weeks.

classes at each grade level), only a share of the students can participate. In these cases, the project selects students to the treatment group either by randomly selecting classes or singular students (of the students that are not organized in fixed classes). The selection of students in large schools was necessary to ensure that each student receives sufficient 'dose' of treatment during a school year (2 x 4 weeks minimum, in a group of maximum 6 students).

B: Handbook with information to teachers

To inform treated schools about different aspects of project participation, the participating school leaders and teachers received a "handbook" containing instructions on how the additional teaching resource *should* be used, information about the data collection[19] as well as some advice based on previous research–including information about characteristics of previous successful interventions using additional teachers. The distinctions between what should be done (mandatory) and advice (optional) was made clear.

Information on how the resource should be used is described in Appendix A. Information on data collection included information about test-taking, information about the electronic registration form sent to small group instructors as well as information about the different surveys.

To be able to monitor the received treatment dosage, all small group instructors received an electronic registration form where they were asked to provide specific details on how the small group teaching in mathematics was carried out. The form essentially included lists with names of students. The small group instructor was asked to indicate which students had participated in small group teaching in each of the mathematics lessons along with the duration of each session. They were asked to report the actual time spent on small group instruction–thereby excluding time used for other things such as moving from one classroom to the room where small group instruction was taking place, breaks and so on. On a couple of occasions we also used this registration form to gather information on small group instructors' own time use in the teaching situation. The final information about the data collection contained a brief description of the scope of the annual surveys to all teachers and school leaders–surveys aimed at both treatment and control schools, but with additional questions to treatment schools regarding the implementation of the intervention.

[19] The information on data collection pertaining to the control group was sent to control group schools in a separate document.

The advice-section of the “handbook” described elements that could help make the intervention effective and can be summarized as follows:

- Group students into small groups by ability level.
- Customize instruction according to the ability composition of the small group.
- Close monitoring of student learning.
- Close cooperation between the small-group instructor and classroom teacher.
- The class teacher and small group instructor should work together to make a plan for each mathematics lesson containing clear learning objectives.
- The class teacher and small group instructor should together carry out ongoing assessments/monitoring to identify which students are to be grouped together in the small groups, identify academic focus areas and regularly assess students' academic progress.
- The topics covered in the small group instruction should be closely tied to the topics covered in the regular mathematics lesson.

For further reading we provided the following references:

Gersten, R., Beckmann, S., Clarke, B., Foegen, A., Marsh, L., Star, J. R., & Witzel, B. (2009). Assisting Students Struggling with Mathematics: Response to Intervention (RtI) for Elementary and Middle Schools. *What Works Clearinghouse*. http://ies.ed.gov/ncee/wwc/pdf/practice_guides/rti_math_pg_042109.pdf

Higgins, S., Katsipataki, M., Kokotsaki, D., Coleman, R., Major, L.E., & Coe, R. (2014). The Sutton Trust-Education Endowment Foundation Teaching and Learning Toolkit. London: Education Endowment Foundation. https://educationendowmentfoundation.org.uk/evidence/teaching-learning-toolkit

Kulik, J., og Kulik, C. (1987), Effects of Ability Grouping on Student Achievement, *Equity & Excellence in Education.* Vol. 23:1-2, side 22-30.

National Mathematics Advisory Panel (2008). *Foundations for Success. The Final Report of the National Mathematics Advisory Panel.* Washington, DC: U.S. Department of Education.

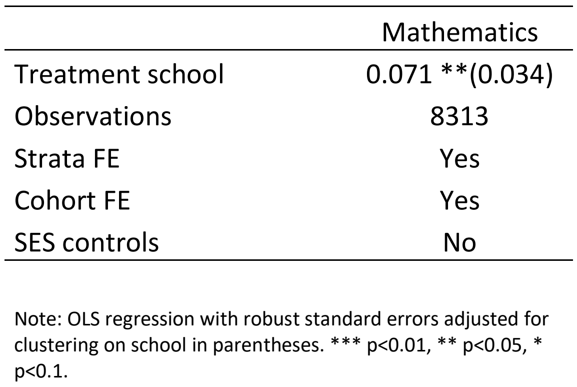
Sharples, J., Webster, R., & Blatchford, P. (2015). *Making Best Use of Teaching Assistants Guidance Report*. London: Education Endowment Foundation.

Slavin, R. (1987), Ability Grouping and Student Achievement in Elementary Schools: A Best-Evidence Synthesis, *Review of Educational Research.* Vol. 57:3, side 293-336.

C: School visits

In total 14 schools were visited during the trial period, where 11 schools were visited one time, while 3 schools were visited on two occasions.

Table A1. Excluding strata containing schools visited during the intervention period



	Mathematics
Treatment school	0.071 **(0.034)
Observations	8313
Strata FE	Yes
Cohort FE	Yes
SES controls	No

Note: OLS regression with robust standard errors adjusted for clustering on school in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

C: Missing test scores

Table A1. Share of missing national test scores by treatment status.

	(1)	(2)
Treatment school	.000	.001
	(.006)	(.005)
Observations	16,276	16,276
Strata FE	Yes	Yes
Cohort FE	Yes	Yes
SES controls	No	Yes

Mean Y	.09	.09

Note: OLS regression where the outcome variable is an indicator of missing national test scores. Robust standard errors adjusted for clustering on school in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A2. Share of missing short-term test scores by treatment status.

	Pooled	Cohort 2008	Cohort 2009	Cohort 2011
Treatment school	-.063***	-.042***	-.056**	-.091***
	(.014)	(.012)	(.027)	(.025)
Observations	25,337	8,491	8,736	8,110
Strata FE	Yes	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes	Yes
SES controls	No	No	No	No
Mean Y	.13	.08	.18	.14

Note: OLS regression where the outcome variable is an indicator of missing project test score. Robust standard errors adjusted for clustering on school in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

D: Geographic mobility

Table A3. Geographic mobility by treatment status.

	(1)	(2)
Treatment school	-.027	-.023
	(.025)	(.024)
Observations	16,276	16,276
Strata FE	Yes	Yes
Cohort FE	Yes	Yes
SES controls	No	Yes
Mean Y	.08	.08

Note: OLS regression where the outcome variable is an indicator of whether the student takes the national test in another school than s/he completed the baseline test. Robust standard errors adjusted for clustering on school in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

E: Treatment effects on Reading and English

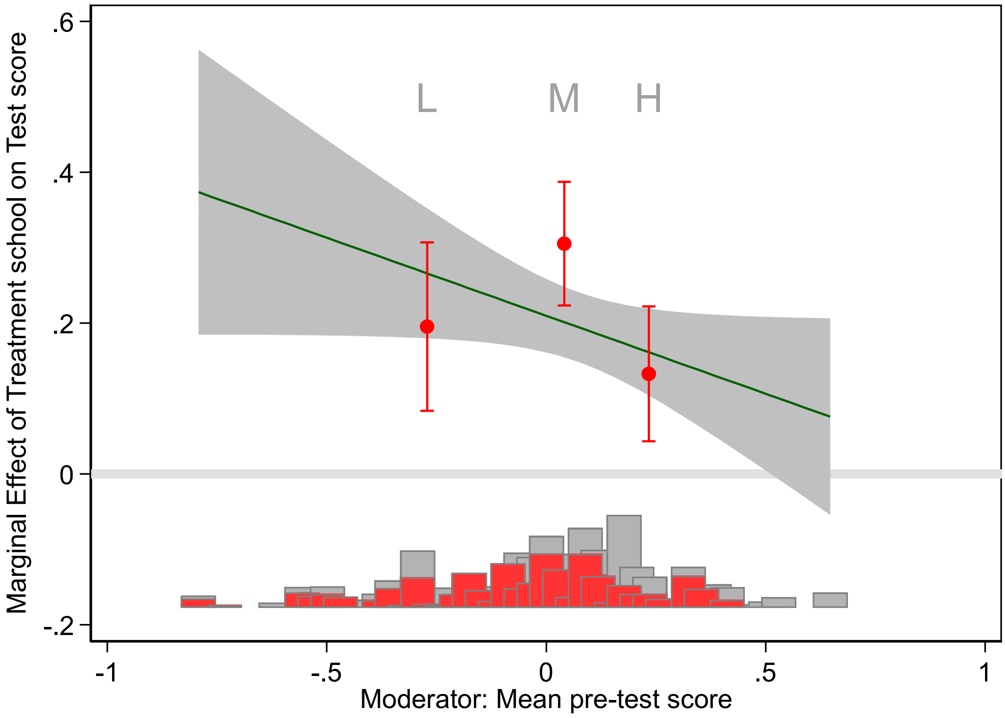
Table A4. Baseline results for reading and English. Dependent variable is standardized national test scores.

	(3)	(4)
	Reading	English
Treatment school	.029 (.027)	-.009 (.026)
Observations	14,735	14,985
Strata FE	Yes	Yes
Cohort FE	Yes	Yes
SES controls	No	No

Note: OLS regression with robust standard errors adjusted for clustering on school in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

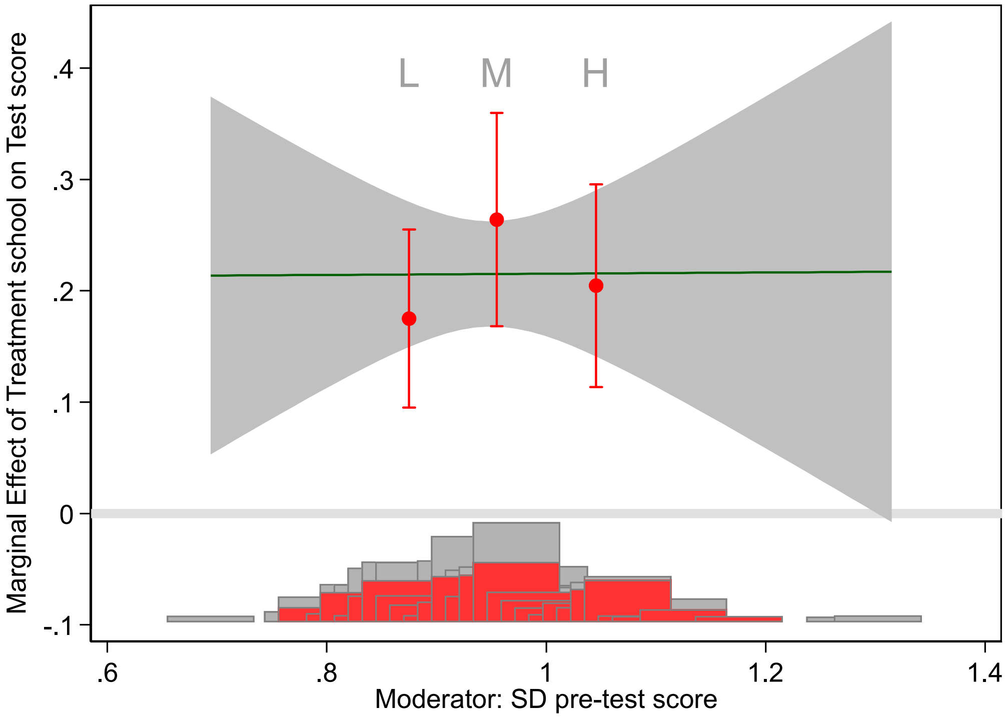
F: Treatment heterogeneity

Figure A1. Treatment heterogeneity by mean pre-test score by school



Note: The plot 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 (pre-test score by school) in the pooled sample, and the red and white shaded bars refer to the distributions in the treatment and control groups, respectively.

Figure A.2. Treatment heterogeneity by within school heterogeneity



Note: The plot 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 (SD of pre-test score) in the pooled sample, and the red and white shaded bars refer to the distributions in the treatment and control groups, respectively.

Table AX. Treatment heterogeneity on parental level of education and class size. Dependent variable is standardized national test scores.

	(1)	(2)	(3)
		Mathematics	
Treatment school	.054 (.038)	.055** (.027)	.060*** (0.27)
Treatment x College edu.	.016 (.038)		
College edu.	.497 (.023)		
Treatment x Class size		-.002 (.002)	
Class size		.003 (.002)	
Treatment x School size			.001 (.001)
School size			-.002 (.001)
Observations	14,891	14,891	14,891
Strata FE	Yes	Yes	Yes
Cohort FE	Yes	Yes	Yes
SES controls	Yes	Yes	Yes

Note: OLS regression with robust standard errors adjusted for clustering on school in parentheses. College edu is a binary indicator of whether one of the parents have college education. Class size and school size (and interactions with treatment) are mean-centered so that the treatment school coefficient refers to the average treatment effect. *** p<0.01, ** p<0.05, * p<0.1.

G: Detailed description of administrative data sources and project tests

From the registers, we have information on gender, country of birth, test results from the National tests in the 5th grade, as well as parental level of education and parental country of birth. We also have project tests that are both pre-tests and outcome variables.

Outcome variables and pre-tests

National test 5th grade: We use test scores from national achievement tests in mathematics from 5th grade as our main outcome. Compulsory national tests in 5th grade have been administered since 2007 in reading, mathematics, and English. The Directorate of Education and Training commissions test development from subject experts at universities in Norway and psychometric experts in the directorate (see https://www.udir.no/eksamen-og-prover/prover/nasjonale-prover/om-nasjonale-prover/). The tests are designed to capture the full range of skills in these subjects among students within each grade. About 96% of all students in Norway take the test; students with special needs and those following introductory language courses may be exempt. Data from 2007 and onwards are available as a score summing up correct responses. In addition, from 2014, a scaled score based on a two-parameter IRT model is available (for details, see https://www.udir.no/globalassets/filer/vurdering/nasjonaleprover/metodegrunnlag-for-nasjonale-prover-august-2018.pdf). The test results are mainly used to track school development over time. Results are conveyed to teachers and parents but have no direct consequence for students. In the present study, we standardize the summed test scores within test and year.

Project tests: The project collected pre-tests at the beginning of the treatment periods and post-tests at the end of each school year for participating cohorts in 2nd and 3rd grade. These tests were developed by the research team in collaboration with teachers and math educators. The tests were digital and meant to mimic national tests while making the difficulty level appropriate for lower grades. They were also piloted before implantation. Teachers received detailed instructions on how to carry out the tests. In the second year, the software added the option to listen to the question read aloud. Psychometric analyses revealed that the tests were adequately unidimensional. Following recommended fit statistics (Maydeu-Olivares, 2013), the Rasch model fitted the data reasonably well in both grade 2 (M2(170) = 1090, $p < .001$, CFI = 0.951, RMSEA 95% CI = [0.040 - 0.044], SRMSR = 0.06), and in grade 3 (M2(170) = 1045, $p < .001$, CFI = 0.945, RMSEA 95% CI = [0.041 - 0.046], SRMSR = 0.057). We used empirical item characteristic curves to inspect item misfit, and no extreme discrepancies or anomalies were observed. Yen (1984)'s Q3 statistic was examined for both tests, but no local item dependency was indicated. As for test information and reliability, both tests adequately covered the lower-to-average ability level, with marginal reliability around .70-.80.[20]

Measurement of background variables

Girl: Dummy equal to 1 if the student is a girl.

[20] For more information, see Haverkamp (2020).

Parental level of education: Five dummy variables representing the highest education level of the parents. The Norwegian Standard Classification of Education has 10 categories: No education (0), Primary education (1), Lower secondary education (2), Upper secondary education, basic (3), Upper secondary education, final (4), Post-secondary non-tertiary education (5), First stage of tertiary education, undergraduate level (6), First stage of tertiary education, graduate level (7), Second stage of tertiary education, postgraduate level (8), Unspecified (9). We recode categories 0,1,2 to primary education, 3,4,5 to upper secondary education, 6 to higher education lower level, 7, 8 to higher education higher level, and 9 to unknown education.

Foreign born: Dummy equal to 1 if the student is not born in Norway.

Second generation immigrant: Dummy equal to 1 if both parents are born abroad while the student is born in Norway.

School size: Measured as the total number of students in the grade.

H. Implementation of treatment

Table A5 summarizes descriptive statistics of received treatment dosage for students in the treatment group. The data source is information given by the small group instructors in the registration form (described in Appendix B). Note that the information is calculated based on fewer observations than those included when estimating the ITT effects in Table 2. The discrepancy between number of students at treatment schools and the number of students receiving treatment is mainly due to not all students being allowed to participate in small group instruction in large treatment schools[21], as described in Appendix A. It can also be partly explained by some students moving between the time of randomization and time of treatment as well as special needs students not participating if small group instruction was perceived to interfere with their rights to special needs education. Overall, about 73-74% of students at treatment schools participated in small group instruction in the school year 2016/17.

Table A5: Implementation for 2008 and 2009 cohorts

	2008-cohort		2009-cohort	
	Mean/SD	N	Mean/SD	N
School year 2016/17:				
Number of weeks in small group instruction	7.64	3104	7.60	3193
	(2.40)		(2.47)	
Average small group size	4.99	3104	5.02	3193
	(1.28)		(1.65)	
Total number of minutes in small group instruction	1103	3104	1075	3193
	(418)		(410)	
School year 2017/18:				
Number of weeks in small group instruction	8.23	3082	7.98	3153
	(2.80)		(2.85)	
Average small group size	4.61	3082	4.65	3153
	(1.27)		(1.31)	
Total number of minutes in small group instruction	1184	3082	1077	3153
	(501)		(449)	
School year 2018/19:				
Number of weeks in small group instruction			7.97	2800[22]
			(2.43)	
Average small group size			5.24	2800
			(1.68)	

[21] For the 2008 and 2009 cohorts this amounts to 842 and 718 students, respectively.

[22] Note that the big drop in the number of students receiving small group instruction in 2018/19 compared to 2017/18 is due to some technical glitch in the registration form making us unable to merge information from the registration forms for all students in the treatment group.

Total number of minutes in small group instruction	1187 (461)	2800

Note: Only the 2009 cohort continued receiving treatment in school year 2018/19. The number of included students decrease over time because students move to different schools.

Figure A3: Distribution of received treatment by cohort and school year

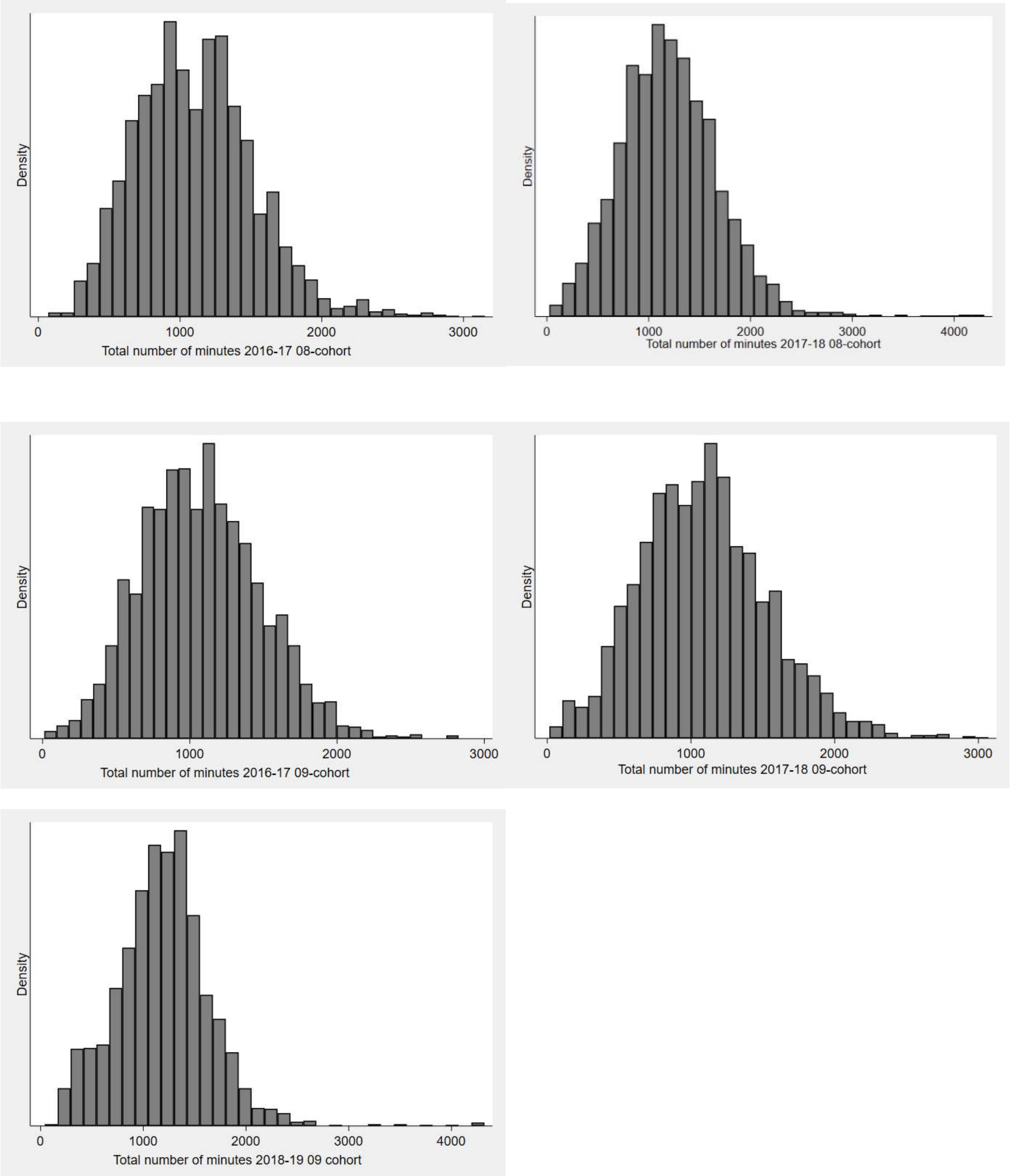


Table provides descriptive characteristics on small group instructors and regular mathematics teachers. The information comes from the annual teacher survey.

Table A6: Teacher characteristics

	Small group teacher	Regular teacher
Male teacher	0.28	0.13
Age	40	41.8
Experience as teacher at this school	7.12	14.53
Experience as teacher in total	11.51	19.29
Number of credits in math at university	58.2	36.8
Have you completed teacher education?	0.98	0.98

In addition, from the teacher survey we know that 31% of small group instructors had previously worked at the school, implying that the majority of instructors were external hires. Finally, when asked to rate on a liker scale of 1-5 (where 5 is strongly agree) to whether "Small groups were composed of students of nearly equal ability level in mathematics" where 97% of small group instructors agreed or strongly agreed with this statement.

Online appendix references

Haverkamp, Y. E. (2020). Investigating the underlying item characteristics in NIFU's 1+1 tests for elementary mathematics. Master thesis. University of Oslo.

Maydeu-Olivares, A. (2013). Goodness-of-Fit Assessment of Item Response Theory Models. Measurement: Interdisciplinary Research and Perspectives, 11(3), 71–101. https://doi.org/10.1080/15366367.2013.831680

Yen, W. M. (1984). Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model. Applied Psychological Measurement, 8(2), 125–145. https://doi.org/10.1177/014662168400800201