Journal of Public Economics 216 (2022) 104765

Contents lists available at ScienceDirect

Journal of Public Economics

journal homepage: www.elsevier.com/locate/jpube

Small-group instruction to improve student performance in mathematics in early grades: Results from a randomized field experiment ☆

Hans Bonesrønning [b], Henning Finseraas [c], Ines Hardoy [f], Jon Marius Vaag Iversen [d], Ole Henning Nyhus [d], Vibeke Opheim [e], Kari Vea Salvanes [e], Astrid Marie Jorde Sandsør [a,*], Pål Schøne [f]

[a] University of Oslo and Nordic Institute for Studies in Innovation, Research and Education (NIFU), Norway
[b] Norwegian University of Science and Technology (NTNU), Norway
[c] NTNU and Institute for Social Research (ISF), Norway
[d] NTNU Social Research, Norway
[e] NIFU, Norway
[f] ISF, Norway

ARTICLE INFO

Article history:
Received 28 November 2021
Revised 19 September 2022
Accepted 17 October 2022

Keywords:
Education economics
Small-group instruction
Tutoring
Tracking
Class size
Field experiment
Intervention
Randomized controlled trial
Teacher-student ratio, mathematics
C93 (Field Experiments)
H52 (Government Expenditures and Education)
I21 (Analysis of Education)

ABSTRACT

We investigate whether small-group instruction improves student performance in mathematics in the early grades using a large-scale RCT 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 of 4–6 students for mathematics instruction for 3 to 4 h 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 0.06 of a standard deviation in national tests, with no differential effect by baseline test score level, parental education, or gender. Our study is particularly relevant for policy-makers seeking to use additional teaching resources to target a heterogeneous student population efficiently.

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

☆ We are grateful for comments from Simon Calmar Andersen, Colin Green, Diane Schanzenbach, participants at AERA, the LEER Workshop on Experimental Evidence in Education Economics, the Educational Resources and Student Performance Workshop and the CESifo Area Conference on Economics of Education 2022, anonymous reviewers, as well as comments from the Scientific Advisory Board appointed by the Research Council of Norway. 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. Shared first authorship with authors listed in alphabetical order.

* Corresponding author.
E-mail address: a.m.j.sandsor@isp.uio.no (A.M.J. Sandsør).

[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.

https://doi.org/10.1016/j.jpubeco.2022.104765
0047-2727/

or small-group instruction – which has been shown to substantially improve student learning (Dietrichson et al. 2017; Nickow et al. 2020), but more knowledge is needed on the program characteristics that allow for high-impact tutoring for all and at scale (Robinson & Loeb 2021). Tutoring has also emerged as a promising strategy for addressing learning loss related to Covid-19.

We present new evidence from a large-scale experiment providing tutoring in mathematics to students of all ability levels during early grades. Additional teachers are used to provide small-group instruction in parallel to regular classroom instruction using a pull-out strategy where each group returns to regular instruction once their tutoring period is over. The small-group instruction is carried out 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 pre-registered randomized controlled trial (RCT) in 159 Norwegian primary schools (78 treatment schools and 81 comparison schools) 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 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. The regular teacher and small-group teacher coordinated their teaching so that the same topics were covered in the main class and the small group. To allow for tailoring of instruction, teachers were advised to construct small groups with students of similar ability levels in mathematics, a strategy chosen by most teachers according to surveys.

The field experiment was made possible through a Norwegian government grant of 20 million Euros to hire 80 qualified teacher person-years for four school years. Four cohorts of students born between 2008 and 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. Across all four years, approximately 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, students in treatment schools increased their performance on national tests in mathematics by 0.06 of a standard deviation half a year after the intervention ended. We also find that all student subgroups benefit from treatment, regardless of baseline test score level, parental education and gender.

Our paper adds to the literature on the effect of increased teacher resources on student performance (e.g., Schanzenbach 2006; Angrist et al. 2019; Hoxby 2000; Browning and Heinesen 2007; Fredriksson & Öckert 2008; Leuven et al. 2008; Iversen & Bonesrønning 2013). While the evidence is mixed (see Leuven and Oosterbeek (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 increasing the teacher-student (TS) ratio through reduced class size, suggesting that more flexible approaches to increasing the TS ratio may be key. Alternative strategies to reduce the TS ratio include having more than one teacher involved in teaching the same student group (Solheim and Opheim, 2019). 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. A recent paper by Haaland et al. (2021), using extra teachers mostly within the classroom for literacy instruction, only finds positive effects when combined with teacher professional development (TPD). In contrast, our study suggests that using extra teachers to provide tutoring, thereby bringing students out of the classroom, yields positive effects for all students with no additional TPD needed.[3]

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 0.3 to 0.4 of a standard deviation. Furthermore, 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 (the majority between 10 weeks and one school year in Nickow et al. (2020)), targeted at low-ability students, one-on-one tutoring 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 lower dosage tutoring (two sessions of 4–6 weeks per year) 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 and can be integrated into a standard school day.

Finally, our paper adds to the literature on ability grouping, as small groups were largely comprised of students of similar ability levels in mathematics. Ability grouping is less restrictive than tracking in both scale and permanence, but involves the same potential trade-off between the gains of allowing for more targeted instruction and the losses for lower-achieving students feeling stigmatized and not being able to benefit from positive peer effects (Figlio & Page 2002; Duflo et al. 2011; Oakes 1985). A small literature credibly identifies the impact of tracking on student outcomes (Betts 2011, Duflo et al. 2011). 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 the opposite for Germany where between-school tracking is harmful for low-achieving students. Duflo et al. (2011), however, show that within school ability tracking in a developing country (Kenya) benefits all students. We measure the impact of a less comprehensive form of tracking with ability grouping in one subject only for a limited period of time, implying less impact from peers 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. Section 5 also includes a cost-benefit analysis and a discussion of scalability. Finally, Section 6 offers some concluding remarks and discusses our results in relation to previous findings in the literature.

2. Institutional context and the intervention

2.1. Institutional context

In Norway, compulsory education is free, and less than 4 percent of students attend private schools. The public sector at the

[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 during the first years of schooling.

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

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 without 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 received special-needs education in 2017. 37 % of these students received assistance in their regular classes and the rest were taught alone or in small groups of eligible students (The Norwegian Directorate for Education and Training 2021).

While the Education Act (1998, § 8–2) in Norway allows for small-group instruction, results from our teacher survey indicate that there was no wide-spread use in Norwegian primary schools prior to our intervention. During the intervention, when asked about the number of students that participated in small-group instruction during the previous mathematics lesson, teachers in treated grades 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.

2.2. Treatment description

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 primary schools. First, permanent tracking is not allowed, but small homogenous student groups can be pulled out of their regular class for shorter periods. Second, the treatment dosage is determined by legislation saying that the students will be taught mathematics for 560 h during grades 1–4, i.e. about 140 h per year, allowing for a planned dosage of minimum 30 h of small-group instruction per year (see online appendix A for details). The sessions differed in length, as there are local variations in how schools organize mathematics instruction. While some schools have long sessions (up to 90 min), others have shorter sessions, often 60 or 45 min. Instruction was given in parallel to all regular mathematics classes. See online appendix A or the pre-analysis plan (Bonesrønning et al. 2018) for further details on the intervention.

Throughout the project, small-group teachers provided detailed information (via a registration form) on which students received smallgroup instruction and the instruction length for each session, excluding time spent for breaks.[5] Calculations show that the average small group consisted of about 5 students that received about 8 weeks of small-group instruction per treatment year, amounting to between 1075 and 1184 min depending on the cohort and treatment year. This is between 60 and 66 percent of the planned minimum treatment. Note that planned and received treatment may not be directly comparable as received treatment deducts time spent on breaks.

National legislation requires that teachers are formally qualified to teach mathematics at the primary level so that only formally qualified teachers are hired. From a teacher survey, we have information on how small-group instructors were recruited and the characteristics of small-group instructors and regular teachers (see online appendix C, Table C2 for details). 31 % of small-group instructors were recruited from within the school, meaning that most were externally recruited. Compared to regular mathematics teachers, a larger fraction was male (28 % versus 13 %), they were on average two years younger (40 versus 42), and they had seven years less teaching experience (12 versus 19 years). However, they had more credits in mathematics (58 versus 37), equivalent to about 2/3 of a semester in higher education. There was no difference in the share that had completed teacher education, about 98 % for both groups.[6]

The small-group teachers received no training as tutors, but both small-group and regular teachers received a handbook including detailed instructions on how to implement the intervention – i.e smallgroup size, duration, etc. – information on data collection as well as recommendations based on previous research. The latter included characteristics of previous successful interventions using additional teachers and, importantly, encouraged the teachers to create small groups with students of similar mathematical abilities.[7] Based on survey data from small-group instructors, we know that most of them followed this recommendation as 97 % reported that they agreed or strongly agreed with the statement that "Smallgroups 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). Table 1 shows the treatment age and duration for all cohorts that participated in the intervention and the timing of the national test and project administered pre- and post-tests in mathematics.

In this paper, we mainly restrict the analysis to birth cohorts unaffected by the Covid-19 pandemic when completing the national tests (2008 and 2009), although results for the remaining cohorts are consistent and reported in the online appendix F, Table F1.

In addition to the registration forms submitted electronically at the end of each month, teachers and principals received yearly surveys, and visits were carried out at some treatment schools. Also, the project group met yearly with teachers and principals at treatment and control schools. Together, this allowed us to follow the implementation and data collection 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 highly appreciated (Bubikova-Moan & Opheim 2020).

[5] See online appendix B and online appendix C, Table C1 and Figure C2 for details on the registration data and implementation.

[6] The teachers survey also provides some information on how the regular 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, indicating that teachers agree or highly agree with this statement.

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

Table 1
Starting age and treatment duration.

	Cohort			
School year	**2008**	**2009**	**2010**	**2011**
2016/17	3rd grade$^{PRE,\ POST}$	2nd grade$^{PRE,\ POST}$		
2017/18	4th grade	3rd grade^{POST}		
2018/19	Test (5th grade)	4th grade		2nd grade$^{PRE^*,\ POST}$
2019/20		Test (5th grade)	4th grade$^{PRE^*}$	3rd grade
2020/21			Test (5th grade)	
2021/22				Test (5th grade)

Notes: The table shows the treatment age and duration of the four cohorts that were part of the 1 + 1 project and the timing of the different mathematics tests. PRE refers to the baseline ability test, POST refers to post-tests after treatment, and Test refers to the National test for all 5th graders in Norway. * Carried out in the spring at the end of the previous school year.

3. Randomization, data, and balance

3.1. Randomization

Randomization was carried out at the school level within each of the ten municipalities participating in the project, geographically spread from the southern to the northern part of Norway and all fairly densely populated. 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 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 teacher-student ratio than larger schools. Smaller schools were able to have smaller groups for a longer duration but were instructed to use any surplus teaching hours beyond this for cohorts that were never treated by the intervention. In larger schools, we randomized classes or groups to treatment to ensure sufficient treatment intensity. Overall, about 73–74 % of students at treatment schools participated in small-group instruction.

3.2. 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 D 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 baseline ability and post-test data collected by the project. We developed math tests in collaboration with teachers and math educators. For the first two cohorts, the focus of this paper, baseline ability tests were carried out early in the school year (August), while baseline ability tests were carried out in the spring of the previous school year for the next two cohorts (see Table 1).[8] 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 G1). This is important since missing test scores will not bias our results and will have a negligible impact on statistical power. In the online appendix (Table I1), 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.

3.3. 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 D for details on background variables).[9] Table D1 in the online appendix presents descriptive statistics on SES for the population of students and balance tests. We find that treatment and control schools are balanced across gender, immigration status, and school size. Treated students have a slightly higher share of parents grouped in the highest education level category (graduate and post-graduate level of tertiary education), whereas the shares are slightly lower (not significantly) for the two education groups upper secondary education and undergraduate level of tertiary education. Reassuringly, the F-test of joint significance produces a large p-value of 0.41, meaning that randomization was successful.

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

[8] The exception is the first year of the project (the 2016/2017 school year), for which we did the.

[9] 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.

randomization. Thus, our classification ensures that β is a clean ITT estimate, although likely representing a lower bound estimate of the treatment effect. As 73–74 % of students were treated, we can scale our treatment effects by 1.3 to estimate the local average treatment effect (LATE). 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*, while cost-benefit and scalability are discussed in section *e*.

5.1. 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 0.066 standard deviations relative to students in the control group.[10] When we add SES controls, the estimate declines to 0.058 of a standard deviation. For comparison, we find that students with a university-educated father perform about 0.14 of a standard deviation better than other students. Thus, the effect amounts to about one-third of the education difference. Our estimates are in-between the high-dosage (0.31) and small-dosage (0.015) treatment estimates in Fryer (2017).

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., 2019), 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. (2019) in interacting the treatment indicator with the strata fixed effects and deriving the average treatment effect from these interaction terms.

5.2. Effects on national test scores in reading and English

Table H1 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 spillovers from small-group mathematics instruction, e.g., 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 0.029, less than half of the effect on mathematics. The difference between the ITT for math and reading is, however, not statistically significant.

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

	(1)	(2)
	Mathematics	
Treatment school	0.066** (0.031)	0.058** (0.026)
Observations	14,891	14,891
Strata FE	Yes	Yes
Cohort FE	Yes	Yes
SES controls	No	Yes
RI p-value	0.05	0.05
IWE	0.067** (0.031)	0.057** (0.027)

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

5.3. 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 G2). 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 0.158, which is about three times larger than the treatment effect on the national tests. 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.

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 0.04 and 0.30 for the Always-Reporters, i.e. positive and with our national test estimates within the bound.

5.4. Treatment effect heterogeneity

Finally, we study treatment 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 negative interaction term indicates that the 2008 cohort drives the

[10] 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 J, Table J1.

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

	Pooled	Cohort 2008	Cohort 2009	Cohort 2011
Treatment school	0.158*** (0.031)	0.144*** (0.049)	0.169*** (0.046)	0.164*** (0.051)
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.

treatment effect in Table 4. 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 since the treatment effect is larger for female students. However, the difference across gender is not statistically significant. Table E1 in the online appendix investigates heterogeneity by parental education, class size, and school size, all of which show small differences and none that are statistically significant.

We also 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 baseline ability score in the school and within-school heterogeneity was not pre-registered and should be considered exploratory. To examine heterogeneity in 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.

Fig. 1 shows a positive correlation between the treatment effect and baseline test score, but the interaction term is not statistically significant (coeff = 0.01, p = 0.51). 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.[11] These estimates indicate that there is a weak non-linearity in the marginal effects with the treatment effect being slightly larger for the mid-level achievers on the baseline test (see Duflo et al. (2011) for similar results, and see Smith et al. 2013, Gersten et al. 2015, and Guryan et al. 2021 for studies that find larger effects for struggling students). If classroom teaching targets mid-performing students, one expectation may be that instruction in homogenous groups will benefit low- and high-performing students. However, we find no support for this reasoning. The coefficients for L, M, and H are not significantly different, so the main impression from this analysis is that all students benefitted about the same from the treatment.

Online appendix E presents treatment effects across average baseline scores and within-school heterogeneity. Figure E.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 = 0.06). 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 baseline ability score distribution, as indicated by the point estimate for the medium group (in red). This result suggests that compared to schools with medium average baseline test scores, schools with low and high average baseline scores are somewhat less able to utilize the benefits of the treatment. However, only the coefficient on H is significantly different from M. Schools with high average baseline scores might also face ceiling effects.

In Figure E.2 in online appendix E, we interact the treatment indicator with the school's standard deviation of the baseline test

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

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

	Cohorts 2008 & 2009	Gender
Treatment	0.073** (0.036)	0.046 (0.029)
Treatment × 2009-cohort	-0.031 (0.045)	
2009-cohort	-0.009 (0.031)	
Treatment × Female		0.024 (0.030)
Female		-0.251*** (0.021)
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.

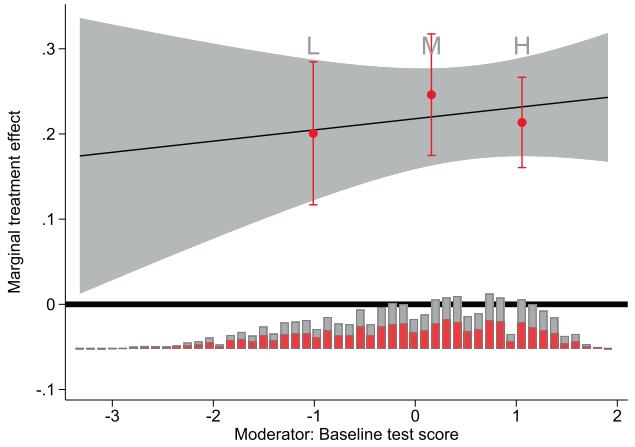

Fig. 1. Treatment heterogeneity by baseline ability score. Note: The plot shows the estimated marginal effects using both a conventional linear interaction model and a binning estimator. The total height of the stacked bars refers to the distribution of the moderator (individual baseline ability score) in the pooled sample, and the red and white shaded bars refer to the distributions in the treatment and control groups, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

score. Here we find that the linear model produces flat marginal treatment effects. 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 benefit as much from the intervention as high-performing ones. In a recent paper, Guryan et al. (2021) 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.[12]

5.5. Cost efficiency and scalability.

Regarding cost-efficiency, the descriptive statistics in Table C1 in the online appendix suggest that every student received a unique treatment equivalent to 222 min yearly. On average, the students were treated for 2.5 years. This constitutes 1.25 percent of a teacher's person-year.[13] 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 one and 2.5 years of treatment. This implies that our intervention is slightly more efficient than findings from a small-group instruction intervention targeting low-performing 8th graders in Norway (Kirkebøen et al. 2021) 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 somewhat higher than Project STAR (Schanzenbach 2006). As 73–74 % of students were treated, our LATE effects are 30 % higher.

For policy relevance, it is important to consider the scalability of the intervention. Statistical inference, population representativeness, and representativeness of the situation are vital elements concerning scalability (Al-Ubaydli et al. 2017). While our pre-registered analysis plan eases concerns regarding statistical inference, the included school and school owners are not a representative sample in the Norwegian setting. The main differences are population and school size, settlement patterns, parental education, and school spending. School spending per student is about 40 % higher among school owners not in our experiment, mainly because of scattered settlement patterns and economies of scale. However, the heterogeneity analyses in Table E1 in the online appendix and Fig. 1 suggest that treatment effects do not vary by parental education, class size, school size, or prior ability levels. The negative association between treatment effects and mean prior ability measured at the school level might suggest a downward bias since schools in the treated municipalities had a .085SD higher score on the national test in numeracy in pre-intervention years. The largest concern regarding scalability is if schools in more rural areas in Norway would be able to recruit teachers with the same competence as those hired in treatment schools during the experiment.

6. Conclusion

Our results show that lower dosage tutoring in mathematics for primary school students can increase learning outcomes for students of all ability levels, even without increasing overall instruction time. We find sizable effects on performance in mathematics. Students in treatment schools increased their performance on the national test by 0.06 of a standard deviation half a year after the intervention.

The effect sizes are smaller than those in the high-dosage literature but larger than those in previous lower 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. The recent *meta*-analysis on tutoring by Nickow et al. (2020) shows larger positive effects than reported here, typically around 0.30-0.40 standard deviations. The majority of the included programs are relatively high dosage and aimed at low-ability students, where 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.

Our findings add to the tutoring and ability grouping 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. 2022). This makes our study particularly relevant for policy-makers seeking additional teaching resources to target a heterogeneous student population efficiently.

Data availability

The authors do not have permission to share data.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jpubeco.2022.104765.

[12] In future work we will investigate mechanisms by linking project data to test data for the parents/guardians who consented to data linkage (89–93 %), allowing us to study the effect of treatment dosage, small-group composition and school and teacher characteristics, among other things.

[13] In our ITT setting, each student received about 1,110 min of small-group instruction yearly. The average group size was 5. A teacher-person-year constitutes 44,460 min (741 h) of lecturing.

References

Education Act. (1998). Act relating to Primary and Secondary Education. (LOV-1998-07-17-61). Retreived from https://lovdata.no/dokument/NL/lov/1998-07-17-61.

Al-Ubaydli, O., List, J.A., Suskind, D.L., 2017. What Can We Learn from Experiments? Understanding the Threats to the Scalability of Experimental Results. American Economic Review Papers and Proceedings 107 (5), 282–286. https://doi.org/10.1257/aer.p20171115.

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), 309–324. 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. 10.1016/B978-0-444-53429-3.00007-7.

Blatchford, P., Russell A., and Webster R. (2012). *Reassessing the Impact of Teaching Assistants: How Research Challenges Practice and Policy*. New York: Routledge. 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. 10.17605/OSF.IO/YWQVC.

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

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, 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. https://doi.org/10.3102/0034654316687036.

The Norwegian Directorate for Education and Training (2021): Fakta om spesialpedagogisk hjelp og spesialundervisning. Retrived from webpage: Fakta om spesialpedagogisk hjelp og spesialundervisning (udir.no).

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., Sandsør, A.M.J., Strøm, B., 2017. Do smaller classes always improve students' long-run outcomes? Oxford Bull. Econ. Stat. 79 (5), 654–688. https://doi.org/10.1111/obes.12161.

Figlio, D.N., Page, M.E., 2002. School choice and the distributional effects of ability tracking: does separation increase inequality? Journal of Urban Economics 51 (3), 497–514. https://doi.org/10.1006/juec.2001.2255.

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. Quart. J. Econ. 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: Duflo, E., Banerjee, A. (Eds.), Handbook of Field Experiments, Vol. 2. North-Holland, Amsterdam, pp. 95–322. 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. Am. Educ. Res. J. 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? J. Econometr. 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. 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), 630–651. https://doi.org/10.1177/1536867X1701700306.

Hoxby, C.M., 2000. The Effects of Class Size on Student Achievement: New Evidence from Population Variation. Quart. J. Econ. 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. 10.1017/CBO9781139025751

Iversen, J.M.V., Bonesrønning, H., 2013. Disadvantaged Students in the Early Grades: Will Smaller Classes Help Them? Educ. Econ. 21 (4), 305–324. https://doi.org/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. Statistics Norway, Research Department. Discussion Papers 957,.

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

Leuven, E., Løkken, S.A., 2020. Long-term impacts of class size in compulsory school. J. Human Resour. 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 J. Econ. 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. Analytischer Bericht, EENEE, p. 33.

Matthewes, S.H., 2021. Better together? Heterogeneous effects of tracking on student achievement. Econ. J., 131(635), 1269-1307. 10.1093/ej/ueaa106.

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. Educ. Res. 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. 10.3386/w27476

Oakes, J., 1985. Keeping Track: How Schools Structure Inequality. Yale University Press, New Haven, CT.

Robinson, C.D., Loeb, S., 2021. High-impact tutoring: State of the research and priorities for future learning. *National Student Support Accelerator*, *2*1(284), 1–53.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., 2006. What have researchers learned from Project STAR? Brookings papers on education policy 9, 205–228.

Schanzenbach, D.W., 2020. The economics of class size. In: Bradley, S., Green, C. (Eds.), The Economics of Education. (Second Edition). Academic Press, pp. 321–331. 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. Am. Educ. Res. J. 50 (2), 397–428. https://doi.org/10.3102/0002831212469045.

Solheim, O.J., Opheim, V., 2019. Beyond class size reduction: Towards more flexible ways of implementing a reduced pupil–teacher ratio. Int. J. Educ. Res. 96, 146–153. https://doi.org/10.1016/j.ijer.2018.10.008.

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 Manage. 33 (1), 78–96. https://doi.org/10.1080/13632434.2012.724672.

Zimmer, R., 2003. A new twist in the educational tracking debate. Econ. Educ. Rev. 22 (3), 307–315. https://doi.org/10.1016/S0272-7757(02)00055-9.