CESifo WORKING PAPERS

9443
2021

November 2021

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, Pål Schøne

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 9443*

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

Abstract

We report results from a large-scale, pre-registered randomized field experiment in 159 Norwegian schools over four years. The intervention includes students aged 7-9 and consists of pulling students 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. All students, not only struggling students, are pulled out. We find that students in treatment schools increased their performance in mathematics by .16 standard deviations at the end of the school year and by .06 standard deviations in national tests 1-2 years later, 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.

JEL-Codes: C930, H520, I210.

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

*Hans Bonesrønning**
Norwegian University of Science and Technology
Trondheim / Norway
hansbo@svt.ntnu.no

Henning Finseraas
NTNU / Trondheim / Norway
henning.finseraas@ntnu.no

Ines Hardoy
ISF / Oslo / Norway
ines.hardoy@samfunnsforskning.no

Jon Marius Vaag Iversen
NTNU Social Research / Norway
jon.iversen@samforsk.no

Ole Henning Nyhus
NTNU Social Research / Norway
ole.nyhus@safforsk.no

Vibeke Opheim
NIFU / Oslo / Norway
vibeke.opheim@nifu.no

Kari Vea Salvanes
NIFU / Oslo / Norway
kari.vea.salvanes@nifu.no

Astrid Marie Jorde Sandsør
NIFU / Oslo / Norway
astrid.sandsor@nifu.no

Pål Schøne
ISF / Oslo / Norway
pal.schone@samfunnsforskning.no

*corresponding author

We are grateful for comments from Colin Green and Simon Calmar Andersen, as well as comments from the Scientific Advisory Board appointed by the Research Council of Norway: Peter Fredriksson, Peter Blatchford, and Dorteh Bleses. We are grateful to 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.

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. The evidence in favor of these policies is at best mixed, leading actors within the education sector as well as researchers to look for alternatives. One of the most prominent alternatives is tutoring – defined as one-on-one or small group instruction. A recent review (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.[3] The 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. Less is known about the performance of low-dosage tutoring, 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.

We present new evidence from an experiment of low-dosage tutoring in mathematics in a setting where tutoring is used as an alternative to classroom-based teaching in the same subject for a shorter period of time. Furthermore, tutoring is directed at students of all ability levels, 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

[1] See, 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). 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. A recent study by Haaland et al. (2021) finds that additional teachers in literacy instruction only yield positive effects in combination with teacher professional development.

[3] See also Fryer (2014), Dobbie and Fryer (2013), Fryer (2017), Fryer and Howard-Noveck (2020) for recent papers that evaluate different tutoring programs.

classes in 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 customized learning, 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. As such, this paper also adds to the literature on tracking. Duflo et al. (2011) show that students of all ability levels benefit from ability tracking in a randomized field experiment in Kenya. They argue that students might benefit when teachers can better tailor their instruction level, but credible empirical evidence on tracking from developed countries is still scarce (Betts, 2011).

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 in 2018 before gaining access to administrative data (Bonesrønning et al., 2018).

We find sizable average treatment effects on student performance. Students in treatment schools score on average .16 standard deviations better than students in control schools after completing a school year with tutoring. However, this effect diminishes but remains at .06 of a standard deviation on national math tests conducted at a later stage.

We also find that all student subgroups, not only struggling students, benefit from treatment – all students benefitted about the same from the treatment. Further, when investigating heterogeneity in terms of school level average baseline scores, we find results suggesting 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.

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

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 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.

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 that the treated students received instruction in small groups 30 to 44 hours 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 pre-analysis plan (Bonesrønning et al., 2018) for further details on the intervention.

National legislation requires that the teachers are formally qualified to teach mathematics at the elementary level so that only formally qualified teachers are hired. The small group teachers received no training as tutors, but they (together with the regular teachers) received a handbook with detailed instructions on how to implement the intervention. The handbook also 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.

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 were receiving 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. During the first two years of treatment, students had about 22 weeks of small group instruction, in groups of about 5 students, with each period lasting for a little more than 4 weeks (see Appendix Table A.5). These numbers confirm that teachers were largely able to provide the expected dosage to all eligible students. 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). The data also showed that most teachers were grouping students by ability level (Appendix Table A.5), which our pre-test group averages also confirmed.

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.[4] 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

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

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.

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 Appendix E 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.[5] 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 Appendix

[5] 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).

Table A.1). 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 Appendix Table A.3, 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 Appendix E for details on background variables).[6] 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 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

[6] 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.

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. 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).

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

Appendix Table A.4 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 Appendix Table A.2). 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.[7] 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.

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.

[7] 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.

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.

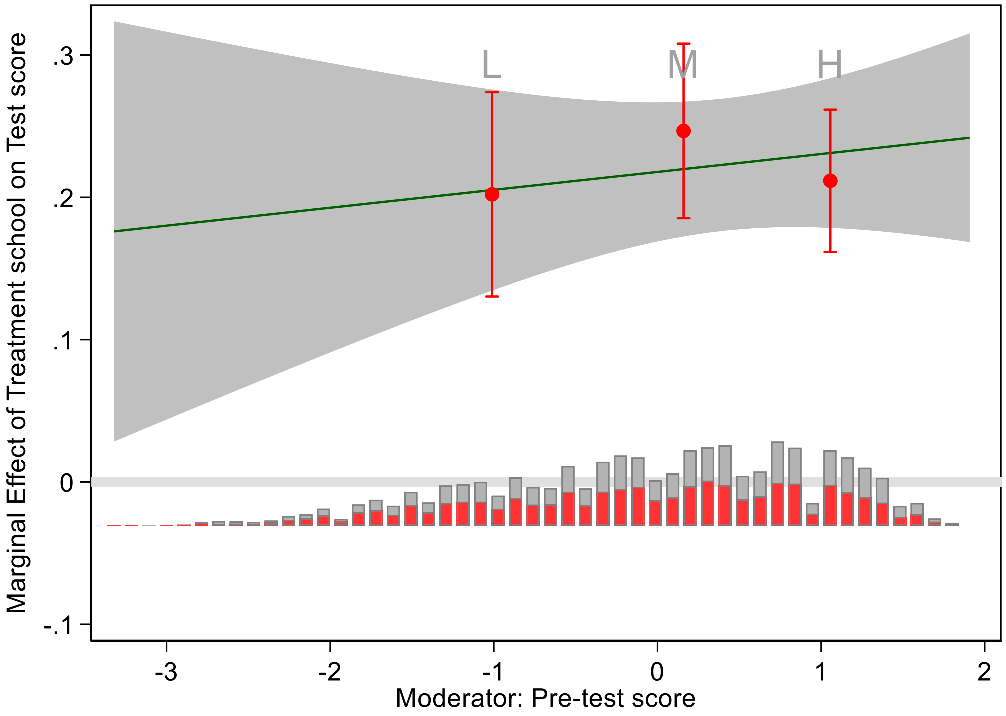
	Cohorts 2008 & 2009	Gender
Treatment	.073** (.036)	.046 (.029)
Treatment x 2009-cohort	-.031 (.045)	
2009-cohort	-.009 (.031)	
Treatment x Female		.024 (.030)
Female		-.251*** (.021)
Observations	14,891	14,891
Strata FE	Yes	Yes
Cohort FE	Yes	Yes
SES controls	Yes	Yes

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.[8] 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 (see Duflo et al. 2011 for similar results, but see Smith et al. 2013, Gersten et al. 2015, and Guryan et al. 2021 for studies that find larger effects for struggling students). Because classroom teaching is often targeted to the mid-performing students, one expectation is that instruction in homogenous groups will be more beneficial for low- and high-performing students. However, we find no support for this reasoning. The main impression from this analysis is that all students benefitted about the same from the treatment.

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 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

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

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.

6. Conclusion

Our results show that customized learning, provided through 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).

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 (see also Guryan et al. 2021). A recent meta-analysis on tutoring shows larger positive effects than reported here, typically around .30-.40 standard deviations (Nickow et al., 2020). 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.

We find that treatment effects are similar in magnitude across all ability levels and for both genders. 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., Salvan es, 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

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

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

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

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).

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

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

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

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

Appendix

A: Missing test scores

Table A.1. 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 A.2. 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.

B: Geographic mobility

Table A.3. 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.

C: Treatment effects on Reading and English

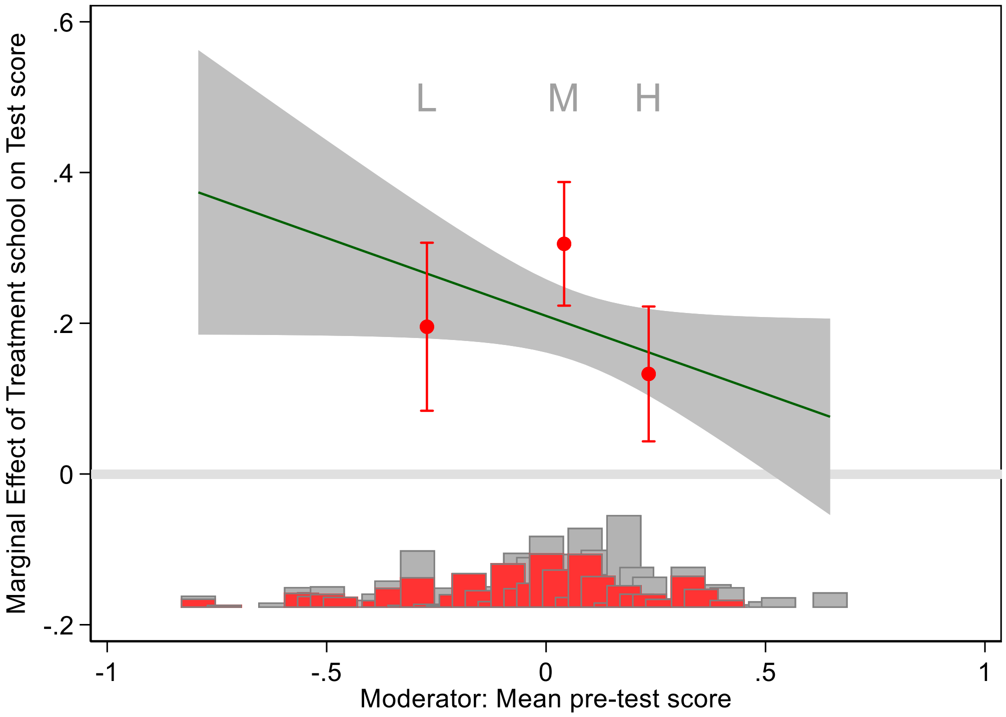
Table A.4. Baseline results for reading and English. Dependent variable is standardized national test scores.

	(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.

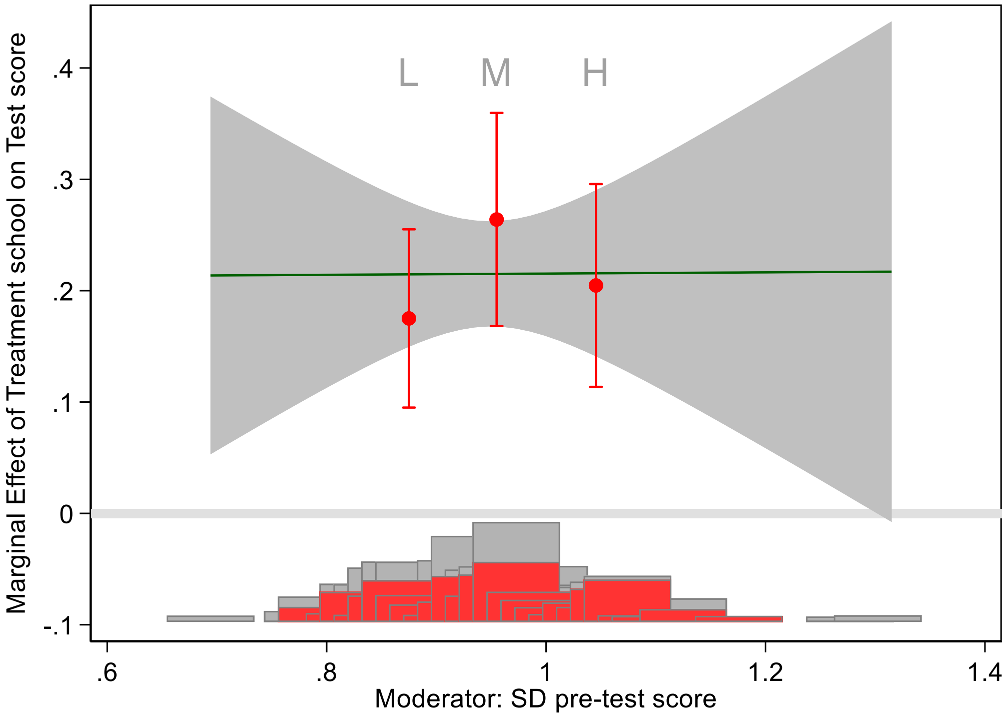
D: Treatment heterogeneity by pre-test scores

Figure A.1. 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.

E: 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.[9]

Measurement of background variables

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

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.

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

F. Implementation of treatment

Table A.5: Implementation for 2008 and 2009 cohorts, first two years of treatment.

	2008-cohort		2009-cohort	
	Mean	St.dev	Mean	St.dev
Number of weeks in small groups instruction	21.82	5.24	21.64	5.15
Average number of weeks in each small group period	4.1	0.73	4.08	0.76
Average small group size	4.9	0.86	4.8	0.84
Total number of minutes in small group instruction	3031	973	2959	950
To what extent do you agree with the following statement: Students are placed into small groups with students on the same ability level (1-5 scale)	4.37	0.83	4.48	0.77

Note: Numbers refer to the treatment years 2016/17 and 2017/18 when both the 2008 and 2009 cohorts were treated. The 2009 cohort continued receiving treatment in the year 2018/19.