Bonesrønning, Hans; Iversen, Jon Marius Vaag

Working Paper

The Importance of Tutors' Instructional Practices: Evidence from a Norwegian Field Experiment

CESifo Working Paper, No. 10878

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Bonesrønning, Hans; Iversen, Jon Marius Vaag (2024) : The Importance of Tutors' Instructional Practices: Evidence from a Norwegian Field Experiment, CESifo Working Paper, No. 10878, Center for Economic Studies and ifo Institute (CESifo), Munich

This Version is available at:
https://hdl.handle.net/10419/282566

CESifo WORKING PAPERS

10878
2024

January 2024

The Importance of Tutors' Instructional Practices: Evidence from a Norwegian Field Experiment

Hans Bonesrønning, Jon Marius Vaag Iversen

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. 10878*

The Importance of Tutors' Instructional Practices: Evidence from a Norwegian Field Experiment

Abstract

We use data from a large field experiment where young students were pulled out of their regular classes and offered mathematics instruction in small homogenous groups, to investigate the importance of the tutors' instructional practices. The analyzes are limited to low achievers, and the instructional practices are characterized by the degree of individualization and the tutors' allocation of attention between students. Tutors who spent much time with avoidant students were associated with a treatment effect of approximately 0.20 SD while tutors who spent little time with these students were associated with no significant treatment effects.

JEL-Codes: I200, I210.

Keywords: tutoring, tutor quality.

Hans Bonesrønning
Norwegian University of Science and Technology, Trondheim / Norway
hans.bonesronning@ntnu.no

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

We are grateful to Henning Finseraas, Ines Hardoy, Ole Henning Nyhus, Vibeke Opheim, Kari Vea Salvanes, Astrid Marie Jorde Sandsør, and Pål Schøne for valuable contributions to the design and execution of the intervention. Comments from the Scientific Advisory Board appointed by the Research Council of Norway: Peter Fredriksson, Peter Blatchford, and Dorteh Bleses are highly appreciated. 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.

1. Introduction

Student heterogeneity is a persistent challenge in all mass education systems. While some people worry that high achievers are being held back in heterogeneous and noisy classrooms, many more are concerned about students who struggle and leave school with poor basic skills. These worries are reflected in educational research. There is a burgeoning empirical literature on the effects of tutoring for struggling students (see reviews by Dietrichson, Bøg, Figes, and Jørgensen, 2017, and Nickow, Oreopoulos, and Quan, 2020).

Much of the recent empirical research is carried out as field experiments, targeting struggling students in one-on-one or two-on-one tutoring. Large treatment effects of almost 0.4 SD are reported from many of these experiments. The attractiveness of such interventions is tempered by the high costs related to the generous student-to-tutor ratio. Costs can be reduced by using low-paid tutors or by using somewhat higher student-to-tutor ratios. This paper provides evidence for the latter alternative.

We use data from a Norwegian field experiment where young students are pulled out of their regular mathematics classes to be taught in small homogenous groups of 4-6 students for two periods of 4-6 weeks during a school year. Bonesrønning et al (2022) show that the intention-to-treat effects from this intervention are significant positive, but that the effect size is only a third of those reported from one-on-one tutoring. The hypotheses to be investigated here are that the lower average treatment effects reflect challenges that are not present in one-on-one tutoring, leading to variations in the tutors' instructional practices, and substantial variations in treatment effects between tutors and schools.

These hypotheses are motivated by the fact that small groups and one-on-one tutoring differ in important ways. It is widely believed that the effectiveness of one-on-one tutoring is due to individualization and customization of instruction. Small groups are like one-on-one tutoring in that individualization of instruction is within reach, but they deviate from one-on-one tutoring in the sense that the tutors - if they choose to practice individualization of instruction - must decide on their allocation of attention across students[2]. An important contribution from this analysis is that it highlights what it takes to transform a high teacher-to-student ratio into significant student achievement.

In addition, the hypotheses are motivated by Norwegian institutions. In line with national regulations, only teachers formally qualified to teach mathematics were recruited as tutors. The teachers who were hired were informed about the characteristics of effective instruction prior to the intervention but without being trained as tutors. The lack of tutor training makes the current intervention different from most US tutoring experiments where tutors in many cases are paraprofessionals or volunteers who are given intensive training beforehand.

We ask the following questions about the instructional practices: Did the tutors take advantage of the small groups to provide individualization of instruction? How did the tutors distribute their attention across the students? And we ask the following questions about the effects: Were the tutors' decisions about instructional practices important for the size of the treatment effects

[2] There are not many papers on the allocation of teachers' time. Notable exceptions are Brown and Saks (1986, 1987). They find that "teachers tend to prefer narrower distributions of learning across students than wider ones."

and the cost effectiveness of the intervention? To the extent that the tutors practiced biased allocations of attention; did some students benefit more than others from the intervention?

We show that the tutors disagreed about individualization of instruction as well as about allocation of attention, leading to four types of tutors. One group consists of tutors who assisted all students and spent much time with avoidant students[3]. Avoidant students are (here) defined as students who need help but do not ask for help. A second group consists of tutors who let the students spend a lot of time working together or alone to solve math problems - but without providing much assistance to avoidant students. The two remaining groups consisted of tutors who did not individualize much, except from assisting avoidant students, and tutors who preferred individualizing of instruction but without paying much assistance to avoidant students.

To be relevant to the existing empirical tutoring literature, we have restricted the analyses of treatment effects to low achievers, defined as the students in the two lowest quintiles in the pretest score distribution[4]. We find that tutors who assisted avoidant students were associated with treatment-on-the-treated effects of 0.21-0.22 SD, while the two subgroups of tutors who provided little assistance to avoidant students were associated with small and insignificantly treatment effects. To put these differences in effectiveness across tutors into perspective, the medium-term treatment effects of 0.07 SD from the ITT-analyses (Bonesrønning et al., 2022) correspond to 0.14 SD per 1,000 USD, which is almost equal to the effect-cost ratio reported by Guryan et al. (2021) evaluating the Saga tutoring program in the US. The difference in treatment effects between the most and least effective tutors of about 0.20 SD is therefore of significant economic magnitude. To be clear, we do not claim that these differences across tutors originated solely with their instructional practices as characterized here. Towards the end of the paper, we discuss at length other factors - correlated with the instructional practices - that might contribute.

Our second major contribution is that we highlight the importance of peers. 2nd quintile students who were placed in groups with other 2nd quintile students experienced treatment effects of 0.17 - 021 SD when exposed to the most effective tutors. 2nd quintile students who were placed in groups with peers from the 1st quintile and exposed to the same tutors experienced no treatment effect – even though the 1st quintile students in these groups experienced significant positive treatment effects. In one interpretation, these findings show that the small groups with the lowest achievers were overcrowded in the sense that even the most effective tutors could not provide all students with adequate assistance. A complementary interpretation is that the most effective tutors allocated their assistance to the most struggling students in these groups.

The rest of the paper is organized as follows. The data are presented in the next section and are followed by three types of analyses. In the first analyses we show how the tutors can be

[3] Ryan, Patrick, and Shim (2005) have investigated the help seeking behavior of 6th grade math students to find that the students display appropriate (65%), avoidant (22%) or dependent (13%) help seeking behavior. While help avoidance is likely to increase across the grade levels, we expect that the proportion of avoidant students is higher among low achievers than among other student subgroups.

[4] In this paper 40% of the students are defined as low achievers, but we consistently distinguish between students in quintiles 1 and 2 in the analyses. Our rationale for choosing such a large group of low achievers is that 20% of Norwegian students are defined as low achievers in the PISA test 2018, and that about 30% of the students in the vocational track in the upper secondary school are dropouts. Much of the empirical tutoring literature targets a narrower group of struggling students.

classified into four types based on major characteristics of their instruction. In the second part we show how the treatment effects varied across the four types of tutors. In the third part we consider the endogeneity of the tutor characterizations and the importance of factors that potentially were correlated with the instructional practices. We discuss our contributions and conclude in the final section[5].

II. Data

We use student-level data for two cohorts of students (the 2008- and 2009 cohorts) covering one school year (2016/17) for the 2009-cohort and two years (2016/17 and 2017/18) for the 2008-cohort, a total of 16 276 students in the two cohorts. Appendix Table 2 provides information about the cohorts, treatment length, and pre- and post-tests. Privacy concerns dictate that the survey data cannot be mixed with register data (notably student and family characteristics) in the analyses, implying that individual students can only be characterized by pre- and post-test results in the present study.

Frequent reporting to the project manager was part of the job description for the tutors[6]. In each report, they were asked to identify the students in their current small group by performance level (low achievers, middle achievers, high achievers, and mixed composition) and to report their instructional practices for this group, especially their choice between tutor-student and student-student approaches and the allocation of available instructional time between the students. They were also asked about the allocation of time between presentations, seatwork, guided practice, and feedback, as well as their emphasis on automatization versus problem solving. Moreover, they were asked about the number of students in the group, the dosage of treatment measured by the number of weeks, and the number of lessons per week so that the quantitative parts of the treatment could be described in detail. All mathematics teachers involved in the experiment received questionnaires about their background (education and experience.

The students were tested in mathematics early in the fall of 2016 - a few weeks after start of the semester. Ideally, the pre-tests should have been taken prior to treatment, but this could not be accomplished due to a strict timeline imposed on the project. The first post-test was given at the end of the first year of treatment. All these tests were closely connected to the curricula for the respective grades and developed for the project by professionals who were familiar with test design and teaching in the early grades and piloted in schools outside the project. The tests were conducted by a company that specialized in testing, the tests were online, and the results were scored automatically.

Table 1 shows that the two cohorts had approximately equal sized small groups with an average of 5.0 students. The standard deviations were in the interval 1.3-1.7, indicating that quite a few small groups exceeded the upper limit of 6 students. The average dosage was 7.6 weeks for both cohorts, with standard deviations about 2.6, indicating that quite many students received less than the minimum of 4x2 weeks of small group instruction per year. Even though most

[5] Bonesrønning et al (2022) present important Norwegian institutions, the choice of participating schools, randomization, and implementation. This information is presented in Appendix 1.

[6] Kane et al (2011) provide evidence that "evaluations based on well-executed classroom observation do identify effective teachers and teaching practices." Our approach is to rely on the tutors' responses to surveys. One reason for this is to safeguard the validity of the field experiment.

schools were well within the limits set for size and dosage, some schools did not meet the minimum requirements for treatment. We consider the consequences of deviations from the requirements when discussing the robustness of the findings.

The last row in Table 1 shows that most tutors agreed that the small groups were homogenous with respect to the pretest score. To describe the composition of the small groups more precisely we have ranked all students and all small groups by quintiles based on pretest scores. For the small groups, the rank is based on the average pretest score in the group. If all schools were equal (the average pretest scores being equal to the sample mean and equal distributions), and if the students were perfectly sorted, the difference between group rank and individual rank would be zero for all students. In Appendix Table 4 we show that 86-87 percent of the students belong to groups with ranks -1, 0 or 1.

Table 1
Descriptive statistics for the 2008- and 2009-cohorts. Small group size, dosage, and homogeneity

	2008-cohort		2009-cohort	
	Mean/SD	N	Mean/SD	N
School year 2016/17:				
Number of weeks in small group instruction	7.64	3104	7.60	3193
	(2.40)		(2.47)	
Average small group size	4.99	3104	5.02	3193
	(1.28)		(1.65)	
Total number of minutes in small group instruction	1103	3104	1075	3193
	(418)		(410)	
School year 2017/18:				
Number of weeks in small group instruction	8.23	3082	7.98	3153
	(2.80)		(2.85)	
Average small group size	4.61	3082	4.65	3153
	(1.27)		(1.31)	
Total number of minutes in small group instruction	1184	3082	1077	3153
	(501)		(449)	
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		4.48	
	(0.89)		(0.77)	

Teachers who were formally qualified to teach in elementary school were employed as tutors by the schools. Observable characteristics of the tutors and the regular mathematics teachers are reported in Table 2. The tutors were more likely to be males, slightly younger and slightly less experienced compared to the mathematics teachers in the regular classes. The tutors had more credits in mathematics from the teachers' college and had taken more courses in mathematics in upper secondary school compared to the regular teachers. This reflects the recruitment criteria set by the project on recruiting tutors that were qualified to teach mathematics. Note that the number of tutors exceeded 78, reflecting that in some schools the tutor position is shared between two teachers. In these cases, the two tutors were assigned to different cohorts.

Table 2
Characteristics of tutors and regular math teachers. Treatment schools

Teacher characteristics	Average	St.Dev.	Min	Max	N
Gender (female=1):					
Tutor	1,28	0,449	1	2	98
Regular teacher	1,13	0,337	1	2	195
Age:					
Tutor	40,1	11,23	24	66	94
Regular teacher	41,7	11,42	24	67	191
Experience:					
Tutor	11,1	9,11	0	36	99
Regular teacher	12,3	9,62	0	40	207
Credits:					
Tutor	58,0	37	0	240	94
Regular teacher	36,9	29,7	0	240	200
>2 yrs. math secondary school*:*					
Tutors	0,469	0,502	0	1	96
Regular	0,401	0,491	0	1	200

Existing empirical research shows that teacher quality varies widely but among teacher credentials only teacher experience has a statistically significant effect on achievement (see for example Rockoff, 2004, Kraft and Papay, 2014). These findings point to the importance of unobservable teacher characteristics. In the next section we present our measures of the tutors' instructional practices.

III. The tutors' instructional practices

A. Essential characteristics

As stated above, it is widely believed that customization – teaching at the right level – is an essential mechanism behind the large treatment effects in one-on-one tutoring. In addition, Bloom (1984) argue that "feedback-corrective procedures", which are important ingredients in mastery learning, are integral parts of tutoring. Customization and feedback are harder to achieve in small groups than in one-on-one tutoring as there is less instructional time available

per student, and the tutor must decide on the allocation of time across the students[7]. On the other hand, the small group tutor can benefit from student-student interactions where the students assist each other in problem solving. In this case, there is less competition for the tutor's instructional time.

We assume that small group tutoring involves two sequential decisions. First, the tutor must choose between tutor-student and student-student interactions as the "basic model", and thereafter the tutor must choose how to allocate instructional time across the students. Moreover, and because the within-group variation in pretest scores is small, we assume that the tutors react to other student characteristics, especially the students' help-seeking behavior. Ryan, Patrick, and Shim (2005) separate between appropriate, avoidant, and dependent help-seeking students, and find in their study that the proportions of 6th graders in math classes in the respective categories were 65%, 22% and 13%. The behavior of the latter category (dependent help-seekers) lies somewhere between the other two categories. Their evidence thus indicate that many 6th grade students do not seek help when help is needed. Although our students are younger, they are a select group of students with low achievement results. A focus on students' help-seeking behavior may therefore be relevant. This line of reasoning lies behind the questionnaires to the tutors about their teaching practices[8].

Individualization of instruction is measured by the tutors' response to the following statement: "I supervise students who need help" (indiv1), and the allocation of attention across students is measured by the following statement: "I supervise individual students I know need help, even if they do not ask for help" (indiv2)[9].We thus investigate whether avoidant students receive less attention in the group.

We have added a content dimension by distinguishing between routine practice and drill on the one hand (Automat) and working with problems that can be solved in different ways (Problem) on the other. Content may interact with the instructional practices: our hypotheses are that routine practice and drill increases the effects of individualized instruction, while providing problems that can be solved in different ways increase the effects of student collaboration.

All measures are derived from statements rated by tutors on a 1-5 scale where 1 is "strongly disagree" and 5 is "strongly agree". Table 3, the bottom panel, shows that the indiv1-measure has an average of 4.44 and a relatively small standard deviation of 0.58 when reported for low achievers' small groups, indicating that many tutors agree or strongly agree that they supervise students who need help. The proportion of tutors who agree or strongly agree that they supervise students who do not ask for help (indiv2), is much smaller, and the variation substantially higher, compared to the indiv1-measure.[10] The tutors spend somewhat more time on drill and less on problem solving for low achievers compared to all students. Note also that

[7] Betts and Shkolnik (1999) find that "teachers shift time away from group instruction and towards individual instruction" when class size decreases.

[8] We realize that our characterization of the tutors' instructional practiced deviate much from the rich characterizations found in the empirical educational literature (see for instance Clements et al., 2013, Morgen et al., 2015). We discuss the consequences of omitted variables towards the end of the paper.

[9] We use the concept avoidant students for students the tutor knows need help, even if they do not ask for help. Note that this definition is significantly different from - and simpler than - the one we find in psychological literature.

[10] We have considered adding a variable capturing the time tutors spend with students that do not work unless controlled by the tutor, but it appears that this variable does not capture the tutors' allocation of time well.

the description of their own instructional practice for low achievers does not differ much from the average practice for all students, which indicates that many tutors do not differentiate their instruction much across student subgroups.

These examinations of the tutors instructional practices were taken four times during the second year of intervention. Recall that we have been able to sort out the practices for subgroups of students because the tutors simultaneously were asked to characterize the small group they currently were working with into low achievers, medium achieving students, high achievers, or mixed groups. In the analyses presented below we separate students into quintiles based on their pretest scores, and we assume that low achievers cover students in the 1st and 2nd quintiles of the pretest score distribution. We report separate estimates for the two quintiles.

Table 3
Tutors' instructional practices. Descriptive statistics. Teacher observation data

Variable	Observations	Mean	St.Dev.
All students:			
Problems	260	3.55	0.91
Automat	259	2.84	1.01
Indiv1	266	4.34	0.59
Indiv2	266	3.80	0.93
Low achievers:			
Problems	67	3.30	0.97
Automat	66	3.15	1.04
Indiv1	68	4.44	0.58
Indiv2	68	3.91	0.94

Note: 1-4 observations per tutor

B. Tutor types

We have established tutor types based on the degree of individualization of instruction and the allocation of attention across student subgroups.

The correlation between Indiv1 and Indiv2 is equal to 0.442, indicating that quite a few of the tutors who agreed that they spent a lot of time tutoring students also agreed that they helped students who did not ask for help. But there are deviations. We have separated the tutors into four categories based on their answers to the indiv1 and indiv2 statements. The tutors in HH-category reported high values (4 or 5) for both statements, the tutors in category HL reported a high value for indiv1 and a low value (1, 2 or 3) for indiv2, and so on. We label the instructional practices of HH- and LH-tutors as inclusive individualization.

Table 4 shows how the tutors who have reported their instructional practices for low achievers are distributed across the four categories. Three of the cells contain about 30 observations. The tutor type LH is rare.

Table 4
The distribution of tutors according to their instructional practices for low achievers

		Indiv1	
		High	Low
Indiv2	High	33	8
	Low	30	30

Note: The total number of observations is 101, reflecting that in some schools the tutoring man-year is divided between two teachers (each tutoring one cohort).

This division of tutors into 4 categories does not come without weaknesses. Notably, we do not know the numbers of avoidant students in the small groups, and thus we do not know whether the LL- and HL-tutors report spending little or no time with avoidant students because no avoidant students are present in their small groups or because avoidant students are present, but the tutors prefer to use no or little instructional time with these students. This seems to be more of a problem with the HL- than the LL-tutors because the latter subgroup prefers student-oriented practices, while the HL-tutors prefer to assist the students (but seemingly, not the avoidant ones).

IV. The tutors' instructional practices and treatment effects

A. Hypotheses

In this section we investigate whether the treatment effects for low achievers vary between the four tutor types. We expect the HH-tutors to be the most effective because they utilize the small groups to provide individual teaching to all group members, and we expect the LL tutors to be the least effective because low achievers are unlikely to benefit much from student-student interactions or from unassisted seat work. The ranking of the other two subgroups is less obvious. Much depends on the importance of assisting avoidant students. If the returns to individualized instruction for these students are small relatively to other low achievers, and there are no negative externalities associated with unassisted avoidant students, the HL-tutors might be more effective than the LH-tutors. If it is the other way around – the returns to assistance are relatively high and negative externalities are dampened, then the LH-tutors should be more effective than the HL-tutors.

An implicit assumption in this reasoning is that tutors only affect students' achievement gains through their instructional practices. This assumption is potentially restrictive. For instance, the treatment effects associated with the HH-tutors might reflect that they systematically choose smaller groups and more hours of instruction for the lowest achievers. We evaluate the importance of decisions other than the instructional practices in section V.

B. Treatment effects by quintiles

We estimate standard treatment-on-treated (TOT) equations by quintiles based on the students' pretest results. Initially we do not differentiate between tutors. The estimated equations are:

$$y_{ist} = \beta_0 + \beta_1 y_{is,t-1} + \beta_2 \bar{y}_{-is,t-1} + \beta_3 T + \beta_4 CS + \theta_M + \vartheta_c + \varepsilon_{ist} \qquad (1)$$

where y_{ist} and $y_{is,t-1}$ are the post-test and pre-test scores for student *i* in school *s*, $\bar{y}_{-is,t-1}$ is the average pretest score in the grade, and *T* is the treatment indicator, θ_M is a municipal fixed effect, and ϑ_c is a cohort dummy. The students included are those that are present in the small groups at the pre- and posttests. The results are presented in Table 5.

First, note thar the estimates from the TOT-analyses presented here are substantially larger than the average of 0.07 SD reported by Bonesrønning et al (2022) in their intent-to-treat (ITT) analyses. These differences across analyses reflect that the latter include quite a few students who were randomized to treatment without receiving treatment, and that different posttests were used. While the ITT-analyses used posttests taken approximately five months after the end of treatment, the TOT-analyses used posttests taken at the end of treatment.

The treatment effects are precisely estimated to 0.193 SD and 0.183 SD for students in quintiles 1 and 2 respectively. Low achievers benefit from the intervention as much as medium high and high achievers, i.e., students in quintiles 4 and 5. Note also that the estimates for the average pretest scores are significantly negative, and the estimates for the individual pretest scores are significantly positive - and increasing - throughout Table 5. These variables are included in the TOT-equations estimated below, but the estimates for these variables are not reported in the subsequent tables.

Table 5
Treatment-on-treated effects across quintiles of students

	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Treatment	0.193***	0.183***	0.236***	0.190***	0.151***
	(0.0353)	(0.0311)	(0.0302)	(0.0285)	(0.0287)
Pretest, ind.	0.484***	0.632***	0.649***	0.683***	0.872***
	(0.0399)	(0.0698)	(0.0882)	(0.0890)	(0.0655)
Pretest, average	-0.270***	-0.240***	-0.178***	-0.211***	-0.116***
	(0.0495)	(0.0437)	(0.0424)	(0.0389)	(0.0377)
Class size	-0.00482*	-0.00307	-0.00160	-0.00183	-0.00179
	(0.00265)	(0.00226)	(0.00207)	(0.00197)	(0.00180)
Cohort	0.0638*	0.0568*	-0.0188	-0.0311	-0.135***
	(0.0349)	(0.0307)	(0.0303)	(0.0273)	(0.0269)
Constant	-0.00791	0.0962	0.117	0.123	-0.111
	(0.107)	(0.0879)	(0.0793)	(0.0946)	(0.108)
Municipal fixed effects	Yes	Yes	Yes	Yes	Yes
Observations	2,495	2,749	2,820	2,797	2,591
R-squared	0.087	0.061	0.067	0.059	0.091

Notes: Dependent variable is standardized individual posttest score. Robust standard errors in parentheses. ***p<0.001, **p<0.05, *p<0.

C. Treatment effects by quintiles and tutor types

Next, we have estimated equation (1) by quintiles 1 and 2 and for subcategories of schools based on the tutors' instructional practices. The treatment schools are initially divided into two subgroups of tutors based on the tutor's own reports about their approach to avoidant students, that is, the two subgroups are made up of (HH and LH)- and (LL and HL)-tutors[11]. Table 6 - which provides the main results in this paper - shows that low achieving students in schools with HH- and LH- tutors experienced treatment effects of 0.21-0.22 SD, while low achievers in schools with tutors who reported that they spent little time with avoidant students (LL- and HL-tutors) experienced very small and insignificant treatment effects[12]. Thus, these analyses show that the effects of small group tutoring for low achievers vary substantially between tutor types, and that the LL- and HL-tutors were unable to transform a high teacher-to-student ratio into better achievement. However, as indicated by the results reported in column 1, which report the average treatment effects across all quintiles, the tutors that were ineffective for low achievers generated positive treatment effects for middle and high achievers.

Table 6
Treatment effects by quintiles for schools with (HH and LH)- and (LL and HL)- tutors

	All (1)	Quintile 1 (2)	Quintile 2 (3)
HH- and LH-tutors:			
Treatment	0.197***	0.221***	0.211***
	(0.0350)	(0.0686)	(0.0533)
LL- and HL-tutors:			
Treatment	0.0950***	0.0528	-0.00645
	(0.0439)	(0.0886)	(0.0787)

Note: The dependent variable is a standardized individual posttest score. Independent variables in addition to the treatment indicator are standardized pretest score, mean pretest score, regular class size, cohort, and fixed municipality effects ***p<0.001, **p<0.05, *p<0.10.

We investigate two further issues related to the findings in Table 6. The first issue is whether LH-tutors were more effective than HH-tutors. These subgroups differed with respect to their priority of assistance to all students. While the HH-tutors aimed at assisting all students, the LH-tutors reported that they targeted their assistance to avoidant students. If assistance to avoidant students is essential, we should expect that LH-tutors were more effective than HH-tutors.

In Table 7 we provide separate estimations for schools with HH- and LH-tutors to find weak evidence only that the LH-tutors were associated with larger treatment effects than the HH-

[11] In Appendix Table 4 we provide results from estimation of education production functions which show that the tutors who report to spend much time with avoidant students - the HH- and LH-tutors - are associated with significantly larger achievement gains than the LL- and HL-tutors. These findings motivate the division of tutors into two groups.

[12] Estimating a TOT-equation with an interaction between treatment and the subgroup of HH- and LH-tutors the estimate for the treatment indicator is 0.055 and statistically insignificant while the estimate for the interaction term is 0.15 and significant at the 5 percent level.

tutors, the differences in effectiveness being somewhat larger for 1st quintile students than for 2nd quintile students. In neither case are the differences statistically significant.

Table 7
Treatment effects for schools with different types of effective tutors

	All (1)	Quintile 1 (2)	Quintile 2 (3)
HH-tutors:			
Treatment	0.238***	0.217***	0.196***
	(0.0356)	(0.0654)	(0.0507)
LH-tutors:			
Treatment	0.212***	0.268*	0.232***
	(0.0400)	(0.0795)	(0.0584)

Note: The dependent variable is standardized individual posttest score. Independent variables in addition to the treatment indicator are standardized pretest score, mean pretest score, class size, fixed municipality effects and cohort. ***p<0.001, **p<0.05, *p<0.1

The second issue is more subtle. Tutors who have reported the same type of instruction for low achievers might have done this either because they have identical preferences for instruction, or because of high adaptability of instruction to the student body composition. The emphasis on preferences versus adaptability can differ across the tutors. Thus, a tutor who is of the HH-type for low achievers can be a HH-tutor for high achievers, or say, of the LL-type for high achievers[13]. Adaptability is usually considered to be a requisite for ability sorting to work well for all student subgroups.

We have separated tutors who are of the (effective) HH- and LH-types for low achievers into two groups by their adaptability to high achievers. Tutors who remain HH- and LH-type also for high achievers are labeled non-adaptive tutors. It turns out that most tutors are of the non-adaptive type.

The results from estimating treatment effects for non-adaptive HH- and LH-tutors are reported in Table 8[14] and show that this subgroup of tutors is associated with treatment effects of 0.17 SD and 0.25 SD for 1st and 2nd quintile students respectively. Comparing with the estimates reported in Table 6 (reproduced in the lower panel in Table 8), it is evident that non-adaptive tutors were less capable than adaptive tutors of dealing with the challenges in the small groups consisting of 1st quintile students. We hasten to emphasize that these are preliminary findings and that tailoring of instruction will be the subject of a future paper.

[13] Existing empirical research indicate that teachers on average respond to changes in the student body composition by making only small adjustments in their instructional practices (Tomlinson et al., 2003, Tomlinson, 2015). In our case there is substantial between-group variation in student body composition within schools implying that adaptive individual tutor's instructional practices might vary quite a lot across the small groups.

[14] As indicated by the descriptive statistics in Table 3, very few tutors are of the adaptive type. We therefore report results only for the non-adaptive type.

Table 8
Treatment effects for non-adaptive tutors

	Quintile 1	Quintile 2
Non-adaptive HH- and LH-tutors:		
Treatment	0.166***	0.249***
	(0.0903)	(0.0678)
All HH- and LH-tutors:		
Treatment	0.221***	0.211***
	(0.0686)	(0.0533)

Note: The dependent variable is a standardized individual posttest score. Independent variables in addition to the treatment indicator are standardized pretest score, mean pretest score, regular class size, and cohort. The number of students in the two subgroups for non-adaptive tutors are within the interval [1340, 1739]. ***p<0.001, **p<0.05, *p<0.10.

V. Omitted variables.

A. Instructional practices and tutors' credentials

Tutor-type reflects the tutors' preferences and skills, which might be correlated with observables such as experience and education. To investigate this hypothesis, we have estimated equations with tutor-type as the dependent variable and tutors' background characteristics as independent variables. The tutor-types are as identified for low achievers. The results are reported in Appendix Table 3. As shown there, tutors who have chosen advanced mathematics courses in high school are overrepresented among HH- and LH-tutors and underrepresented among LL- and HL-tutors. These findings indicate that preferences for - and skills in - mathematics are to a certain degree decisive for the tutors' instructional practice. We return to these findings when discussing policy implications in the conclusion.

B. The importance of peers

While most low achievers were placed with students from their own quintile, quite many 2nd quintile students were placed in groups that otherwise contained 1st quintile students, and vice versa. Our hypothesis is that 2nd quintile students who were placed in small groups with other 2nd quintile students experienced larger treatment effects than 2nd quintile students who were placed in groups with 1st quintile students[15]. This hypothesis is motivated by the findings reported above which show that effective tutors prioritized assistance to avoidant students. Thus, if 2nd quintile students are relatively less likely to be of the avoidant type compared to 1st quintile students, they might receive less assistance in groups with lower performing peers.

To examine this hypothesis, we need to address the concern that the tutors may have allocated the most struggling 2nd quintile students to the 1st quintile students' small groups (and not necessarily the 2nd quintile with the lowest pretest scores). We do this by excluding all observations of 2nd quintile students in 1st quintile students' groups who did not satisfy the

[15] Note that this approach has some similarities with the burgeoning empirical rank order literature (Murphy and Weinhardt, 2020, Denning et al., 2021, Delaney and Devereux, 2021, Elsner et al., 2021). While this literature analyses medium- and long-run outcomes, our analyses are short run and more related to the traditional peer group literature.

requirement that their pretest score is lower than the lowest pretest score in the small groups that contain mainly 2nd quintile students. That is, we have excluded all observations that deviate from a strict application of the homogeneity recommendation.

The results from TOT-analyses for 2nd quintile students are reported in Table 9. In short, this table shows that placement is important. Among students who were exposed to HH-tutors, 2nd quintile students placed in 2nd quintile students' groups experienced a significant treatment effect equal to 0.17 SD while 2nd quintile students who were placed with 1st quintile students experienced an insignificant effect of 0.03 SD – indicating that many HH-tutors did not reach out to the 2nd quintile students in the 1st quintile groups. For LH-tutors the numbers are 0.21 SD and zero, respectively. That is, these findings are consistent with the hypothesis that the 1st quintile groups were overcrowded and that the highest achievers were least prioritized by tutors[16]. However, below we show that at least the HH-tutors chose smaller dosages of treatment for low achievers, indicating that we should await strong conclusions about the exact magnitude of the instruction's impact on student achievement.

Table 9
Treatment effects for 2nd quintile students with different peers and tutors

	2nd quintile students	
	HH	LH
In 2nd quintile group:		
Treatment	0.168**	0.214***
	(0.104)	(0.0632)
In 1st quintile group:		
Treatment	0.0266	-0.0005
	(0.0940)	(0.0699)

Note: The dependent variable is standardized individual posttest score. Independent variables in addition to the treatment indicator are standardized pretest score, mean pretest score, class size, and cohort. ***p<0.001, **p<0.05, *p<0.1

VI. Correlated decisions

In addition to their own instructional practices, the tutors decided on group size within the interval [4,6] students, on the number of weeks in treatment in the interval of [4,6] weeks, on the allocation of students to small groups, and on the scope of cooperation with the regular math teachers. To the extent that some of the outcomes from these decisions affected the students' performance and were correlated with the instructional practices, the analyses presented above provide biased evidence about the importance of tutors' instruction. To evaluate whether such biases are of a certain magnitude, we estimate equations with each of these factors against the tutor types.

[16] This argument has some familiarity with Lazear (2001) who models teaching as a public good with congestion and assumes that congestion increases more rapidly with increasing groups size if the students require much attention from the teacher.

A. Small group size

Equations with small group size as the dependent variable and tutor characteristics as the independent variables are estimated while controlling for the size of the regular class and cohort. These estimations are carried out for quintiles of students and the results are reported in Table 10, columns 1 and 2. As shown, none of the tutor types have chosen small group sizes that deviate from the sizes chosen by the omitted category of LH-tutors. The estimates for HL-tutors stand out by being large and very imprecise, indicating that at least some of these tutors have practiced larger groups for low achievers.

Table 10
Associations between small group size, weeks in treatment and tutor characteristics

	Small group size		Length of treatment	
	Quintile1	Quintile 2	Quintile 1	Quintile 2
	(1)	(2)	(3)	(4)
HH	0.0599	-0.0865	-2.207**	-1.568
	(0.224)	(0.259)	(1.043)	(1.133)
HL	0.591	1.383	2.572	-2.567
	(1.311)	(1.989)	(4.857)	(2.949)
LL	-0.397	-0.339	-0.916	-0.483
	(0.254)	(0.283)	(1.349)	(1.477)
Class size	0.0413*	0.0574***	-0.143*	-0.0261
	(0.0213)	(0.0184)	(0.0790)	(0.0425)
Cohort	0.206	0.172	-0.182	-1.144
	(0.175)	(0.203)	(0.795)	(0.819)
Constant	3.491***	3.401***	28.92***	25.12***
	(0.494)	(0.450)	(2.036)	(1.632)
Observations	735	714	735	714
R-squared	0.111	0.157	0.038	0.017

Note: Tutor type LH is the reference category. ***p<0.001, **p<0.05, *p<0.1

Table 10, columns 3 and 4, report results from estimations of equations with length of treatment per year measured in hours (not weeks) as the dependent variable and independent variables as in columns 1 and 2. Compared to LH-tutors, HH-tutors spent significantly less hours with students in the 1st quintile. The dosages provided to low achievers by the two other subgroups of tutors are not significantly different from the dosages chosen by LH-tutors. Thus, if length of treatment is a determinant for the treatment effect, the estimates for the effectiveness of HH-tutors' instructional practices for students in the 1st quintile as reported above will be biased: it could have contributed to the treatment effect associated with HH-tutors being smaller than the comparable estimate for LH-tutors in Table 7.

B. Assignment of students to small groups

The probability that an individual student was placed in a group with peers who belonged to the same quintile as the student depended primarily on the composition of the regular class and the student's own pretest score. In addition, the teachers/tutors could overrule the recommendation to form homogenous groups. We have investigated whether the compositions of the small groups - after controlling for composition of the regular class and the student's own pretest score - varied across the tutors. To do this we have established a measure of homogeneity, d-rank (difference in rank), which is defined as the difference between the rank of the individual student and the rank of the representative student in the small group to which the individual student belongs. A student with rank 2 - indicating that (s)he belongs to quintile 2 - who sit in a group where the representative student belongs to quintile 1 has d-rank equal to 1.

Table 11 reports results from estimations of an equation where the individual student's d-rank is the dependent variable, and the independent variables are tutor characteristics together with the individual' pretest score, the average pretest score in the regular class, class size, and cohort.

Table 11
Correlations between d-rank and tutor characteristics

	d-rank All	d-rank ϵ [-1,1] All	d-rank ϵ [-1,1], 2nd quintile
Pretest, ind.	0.436***	-0.139***	-0.622***
	(0.0376)	(0.0270)	(0.105)
Pretest, average	-0.720***	-0.299***	-0.175***
	(0.0544)	(0.0405)	(0.0654)
HH	0.0501	0.0878*	0.199**
	(0.0684)	(0.0502)	(0.0818)
LL	0.00550	-0.0327	0.164*
	(0.0789)	(0.0575)	(0.0943)
HL	0.713***	0.636***	0.607***
	(0.171)	(0.128)	(0.216)
Class size	1.70e-05	-0.000674	-0.00436*
	(0.00214)	(0.00161)	(0.00245)
Cohort	-0.225***	-0.155***	-0.251***
	(0.0384)	(0.0281)	(0.0465)
Constant	-0.138**	-0.00564	-0.198**
	(0.0664)	(0.0482)	(0.0901)
Observations	3,610	2,942	1,001
R-squared	0.082	0.058	0.083

Note: The dependent variable is d-rank. Tutor type LH is the reference category.
***p<0.001, **p<0.05, *p<0.1

Compared to the reference group of LH-tutors, it is evident from Table 11, column 3, that the other tutor types have assigned significantly more students from the 2nd quintile to the 1st quintile students' small groups. The HL-tutors deviate the most from the LH-tutors. These

findings might contribute to the explanation why LH-tutors are the most effective, while HL-tutors belong to the subgroup of the least effective tutors for low achievers – and especially for 2nd quintile students. Note also that the differences in assignment practices among the HH- and LH-tutors is consistent with the finding that the LH-tutors tend to be slightly more effective than the HH-tutors.

C. Collaboration between the tutor and the regular teacher

The teachers and tutors were encouraged to ensure smooth transitions between the small group and the regular class, implying that the teaching plans had to be coordinated. We have asked the regular teachers and the tutors how often the teaching plan is discussed: before each teaching session, before next week, or for the entire period of 4-6 weeks. 87% of the tutors report that they discuss the plan for the next week, about 40% that they discuss before each teaching session, and just as many that they discuss the plan for the entire period. These categories are not mutually exclusive. The shares of *regular* teachers reporting that they belong to one of the two latter categories are 50% and 34%, respectively.

We have estimated equations with the three frequencies as reported by the tutors as dependent variables and the tutor types as independent variables. The results reported in Table 12 show that there are no statistically significant differences in any of the collaboration measures for the HH-, LL-, and LH-tutors. HL-tutors stand out by reporting significantly more collaboration about individual lessons and for the entire period.

Table 12
Correlations between collaboration about the teaching plan and tutor characteristics as reported by the tutors.

	Individual lessons	One week at a time	For the entire period
Variable	(1)	(2)	(3)
HH	0.309	0.156	0.328
	(0.216)	(0.253)	(0.204)
LL	0.113	-0.0464	0.111
	(0.246)	(0.282)	(0.232)
HL	0.955*	0.803	1.064*
	(0.570)	(0.639)	(0.565)
Pretest, average	0.0253	0.0355	0.0102
	(0.171)	(0.177)	(0.170)
Class size	-0.00836	-0.00979	-0.0112**
	(0.00617)	(0.00654)	(0.00474)
Constant	0.439*	0.593**	0.467**
	(0.222)	(0.259)	(0.181)
Observations	63	63	63
R-squared	0.086	0.054	0.118

Note: Tutor type LH is the reference category. ***p<0.001, **p<0.05, *p<0.1

We have estimated the same equations using the frequencies reported by the regular teachers. These teachers are asked the questions repeatedly, so there are more observations, contributing to more precise estimates. The results are reported in Table 13. According to the regular teachers, LH-tutors are associated with significantly less collaboration than the three other categories, and HL-tutors are associated with slightly more collaboration (but not significantly so) about individual lessons and weekly planning - which are quite consistent with the results reported in Table 12.

Table 13
Correlations between planning of lessons and tutor characteristics. As reported by the regular teachers

	Individual lessons	One week at a time	For the entire period
Variable	(1)	(2)	(3)
HH	0.319***	0.409***	0.316***
	(0.0751)	(0.0797)	(0.0755)
LL	0.439***	0.618***	0.488***
	(0.109)	(0.114)	(0.108)
HL	0.640**	0.661**	0.475
	(0.322)	(0.332)	(0.306)
Pretest, average	-0.0218	-0.0469	-0.0468
	(0.0420)	(0.0457)	(0.0392)
Class size	0.00952**	0.00675*	0.00681**
	(0.00369)	(0.00374)	(0.00312)
Constant	-0.113	-0.0270	-0.0854
	(0.0836)	(0.0848)	(0.0718)
Observations	661	661	661
R-squared	0.122	0.152	0.133

Note: Tutor type LH is the reference category. ***p<0.001, **p<0.05, *p<0.1

These findings indicate that ineffective tutors collaborate more with the regular teachers than do the effective LH-tutors. Thus, if collaboration has a positive effect on the quality of instruction, the estimated treatment effects associated with HL-tutors and LH-tutors are biased upwards and downwards, respectively.

We have highlighted four types of decisions that the tutors have made, or have contributed to, together with the regular teachers. The size of the small groups, the dosages of treatment, the student body composition of the small groups and the degree of collaboration between the two teachers might all contribute to the treatment effects. If they contribute, and are correlated with the tutors' instructional practices, our estimates of the importance of the instructional practices are biased. For HH- and LL-tutors the correlations are weak and insignificant. The HL-tutors have chosen larger groups, smaller dosages, and some of them have chosen heterogenous small groups. These choices might have contributed to the small treatment effects associated with these tutors. The LH-tutors seem to have collaborated least with the regular teachers, potentially implying that the estimates for their instructional practices are biased downwards.

VII. Conclusions

This paper provides evidence from a field experiment of small group mathematics instruction for very young students. The focus is on low achievers and the tutors' instructional practices. While there is no consensus about the essential characteristics of effective math teachers in general, we have focused on how the tutors have utilized the high teacher-to-student ratio: Did they individualize their instruction? Did they allocate their attention equally between the students? We find that the tutors who spent much instructional time with avoidant students were associated with treatment effects slightly larger than 0.20 SD, while tutors who spent little time with this student subgroup were associated with statistically insignificant treatment effects. The latter findings imply that interventions with small groups of 4-6 students for 4-6 weeks twice a year are no better than teaching in regular classes if the tutors do not assist the avoidant students.

Our second main finding is that even the tutors who otherwise were associated with large treatment effects, failed to generate significantly positive treatment effects for 2nd quintile students who were placed in small groups of 1st quintile students. Because 1st quintile students experienced positive treatment effects in these groups, a suggestive explanation is that the 2nd quintile students fell victims of the tutors' preferred allocation of attention in crowded groups. That is, when the tutors have preferences for assisting the avoidant and most struggling students, the highest achievers in low achievers' small groups might have experienced little assistance and small or no treatment effects. This interpretation echoes Brown and Saks (1987) who state that "teachers tend to prefer narrower distributions of learning across students".

To be clear, we do not claim that the differences in tutor effectiveness are fully due to individualization and allocation of assistance. There may be important omitted tutor characteristics. For example, individualizing tutors typically provide much feedback and tailored instruction. These are well-established characteristics of effective teaching (see Bloom, 1984, for a seminal contribution), but they are not explicitly focused here. High expectations are another example of a tutor trait that can be correlated with the instructional practices. Also, the tutors can have made decisions about other factors that might have a significant impact on the treatment effects. The most obvious example is the placements of students to small groups. In particular, the poor performance of the small group of tutors who ignored the recommendation to form homogeneous groups might be a combination of this and their lack of assistance to avoidant students. Other examples are the decisions about small group size, dosage, and teacher collaboration. To the extent that some of these factors affect the size of the treatment effect and are correlated with the applied tutor characteristics, the empirical estimates for the tutors instructional practices are biased.

That said, the main findings reported above motivates several hypotheses for future analyses. We mention a few. First, if crowding is an essential mechanism, we should expect that the highest achievers in the less crowded middle and high achievers' small group experience positive treatment effects for tutors who rely on tutor-student interactions. Second, if the effective tutors prefer assisting the poorest performers in a group, we should expect that for example 2nd quintile students in 3rd quintile groups perform better than 2nd quintile students placed in other groups. Third, we should expect that crowding is less of a problem in groups of 4 than in groups of 6 students. However, Clarke et al (2017) report no differences in treatment effects between groups of two and five students in their kindergarten mathematics intervention. They explain this by the greater potential for student-student interactions in the groups of five

students. Translated to the intervention analyzed here, we should expect that the tutors who are ineffective for low achievers are effective for high achievers because they rely on student-student interactions where group size is less of an issue.

Should the model evaluated here be scaled up? Assuming that the effect-cost ratio is crucially dependent on tutor quality, this raises questions about the supply of high-quality tutors (see Davis et al. (2017) for a throughout discussion). The tutors in the field experiment were recruited from urban areas and to schools that offered tutoring to all students – not only to the low achievers who are the focus of this paper. Policymakers should keep in mind that rural areas have thin teacher labor markets, and that the attractiveness of tutoring to potential tutors is not necessarily maintained if small group instruction is offered only to students who struggle in mathematics.

We have also provided evidence that many tutors with only compulsory math courses in high school relied on student-student interaction as their preferred instructional approach – regardless of the student body composition of the small groups. An unanswered question is whether these tutors would choose a teacher-student approach to low achievers if they were given training in small group instruction in advance. Barnes et al (2016) and Guryan et al (2021) report that the tutors participating in their experiments underwent rigorous training processes. They also report that the tutors were followed closely by site managers or research assistants during the interventions. If rigorous training and close monitoring is required, a pressing question is whether successful tutoring can be carried out "in-school" as in the Norwegian experiment or must be "out-sourced" as in some of the US experiments.

Appendix

A. Institutional context and the intervention

The experiment was carried out within the framework of ordinary mathematics teaching in the public schools (enrolling 96.3% of all students in 2016). The public schools are governed by a two-tier system (national and local governments). The national government sets goals, curriculum, distributes instructional time across subjects, defines minimum standards for teachers' formal qualifications and the maximum number of students per teacher. Inclusion is strongly emphasized. Thus, no student subgroups can be excluded from regular classrooms, except for shorter periods of time. The experiment, being a combination of in-school delivery and a pull-out strategy is adapted to these institutions. The local governments run the schools, that is, they decide on the school structure and provide the schools' budgets but have no discretion on teacher qualifications and inclusion of students - important issues in the current experiment.

To increase the length of treatment while maintaining inclusion, it was decided that treatment should be divided in two periods of small group intervention per school year, each period of 4-6 weeks. 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 on average for students in 1^{st}- 4^{th} grade. Small group instruction was given in parallel to all regular mathematics classes.

The public schools are run by local municipalities. The municipalities differ much in size, implying that the number of schools and students per municipality differs much, from one elementary school in the smallest municipalities to 107 schools in the capital Oslo in 2016. In 2016 the national average number of schools per municipality was 6.6.

Since the lion's share of field experiments with tutoring are carried out in the US, it should be noted that there is more between-school segregation by ability in the US than in Norway, where the variation in student performance is much larger within than between schools.

B. Randomization

10 large or quite large municipalities spread around Norway were invited to participate in the field experiment. Large municipalities were chosen because they have relatively well-functioning local labor markets for teachers and reasonable well-staffed municipal administrations, implying that they had the capacity to recruit new teachers and keep control schools going for 4 years with taking tests and providing necessary information. In addition, this approach was chosen because it could shed some light on the local governing system as a moderator for treatment effects. The 10 superintendents were informed that participation would give half of the elementary schools in the municipality one extra teacher man-year (an average of 8 man-years per municipality).

We conducted stratified randomization in the following manner. Within each municipality the schools were ranked based on their mean test score in the national math tests at the fifth grade

(no tests are taken at earlier stages). We averaged over the mean score in the two preceding school years (2014, 2015) to reduce measurement error. Next, we constructed a set of strata of at least four schools in each stratum. In doing so, we followed the recommendation by Imbens (2011) to have at least two treatment and control schools in each stratum, so that one can derive a within-strata variance in the treatment effect. Most strata consist of four or six schools. In three municipalities, we had an uneven number of schools who volunteered to participate in the project, which resulted in one stratum in each municipality with seven schools. Next, we randomized schools to the treatment or the control group by using a random number generator. A total of 159 schools participated, 81 in the control group, 78 in the treatment group. Appendix Table 1, reproduced from our first paper (Bonesrønning et al., 2022) reporting results from the project, shows that randomization was successful.

Having informed municipalities and schools about the outcomes of the randomization process, the researchers visited all participating municipalities to present the intervention for municipal officers and school leaders in treatment and control schools. All schools were informed about the intervention. The leaders in control schools were told not to make changes in the use of resources, to participate in pre- and post-tests and to report on characteristics of teaching such as the size of the regular classes and the formal qualifications of teachers. The treatment schools received information about how to form small groups (size, composition, duration of small group treatment), about cooperation and coordination between the ordinary teachers and the small group teacher(s), and about the routines for reporting about small group participation. These meetings ensured that the information reached the schools widely, helped to clarify misunderstandings and mobilize the schools for implementation.

C. The implementation

Implementation was discussed with municipal officers and school principals in all the participating municipalities. Some compromises were made. The project leadership accepted that the school principals in treatment schools could decide whether to allocate the new teacher to small group instruction or substitute the new teacher for an existing staff member who then was allocated to small group instruction. A few schools asked to split the teacher man-year into two parts. In this case, the two teachers should be responsible for the small group teaching in one of the two cohorts. Importantly, agreement was reached that there should be only one tutor per cohort in each school.

All schools - control schools as well as treatment schools - in the 10 municipalities were instructed to keep the number of teacher assistants in the intervention grades unchanged and not change the use of school resources due to the schools' participation in the project. The allocated teaching year was not fully filled with small group teaching in most schools. The schools were therefore required to use the rest of the man-year for grades that did not participate in the experiment.

In small schools (with one class per grade) or medium sized treatment schools (with two classes per grade) all students in the chosen grades were included. In schools with more than two classes in each grade one teacher man-year was not enough to provide treatment to all students. In these cases, the project leader randomized two classes to treatment. In our earlier intention-to-treat analyses (Bonesrønning et al., 2022) all classes in treatment schools with more than

two classes were included as treated. In the present treatment-on-treated analyses only the treated classes are included.

A Handbook, targeting participating teachers and containing much of the information from the introduction meetings, was distributed to all schools. Here the teachers were recommended to form small groups that were homogenous with respect to pre-test scores. It was emphasized that the two teachers - in the regular class and the small group respectively - should cooperate to coordinate the teaching, to ensure seamless returns to the home class. Assessments should be used to guide areas for focus, provide feedback to students and track student progress. Connections should be made between out-of-classroom learning (in small groups) and classroom teaching.

Empirical evidence about the characteristics of effective instruction in mathematics, based on reviews of existing research made by What Works Clearinghouse (Gersten et al., 2009, Gersten et al., 2015) and the National Mathematics Advisory Panel (2008) were presented in the Handbook.

Appendix tables

Table A1
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

Table A2
Starting age and treatment duration

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

Notes: The table shows the treatment age and duration of the two cohorts that were part of the present analyses as well as the timing of the different mathematics tests. PRE refers to the pre-test (baseline), POST refers to post-tests after treatment and Test refers to the National test for all 5^{th} graders in Norway.

Table A3
Associations between tutor types and their credentials

	HH and LH (1)	LL and HL (2)
Female	0.243***	0.00984
	(0.0792)	(0.0749)
Experience	-0.00406	0.00103
	(0.00336)	(0.00322)
> 1 yrs.math upper secondary	0.175*	-0.238***
	(0.0905)	(0.0887)
Credits higher education	0.000836	-4.67e-06
	(0.00112)	(0.000549)
Constant	0.137	0.441***
	(0.140)	(0.116)
Observations	231	231
R-squared	0.057	0.046

Notes: The number of observations reflects that the tutors have responded to the surveys several times. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table A4
The distribution of individual students' ranks

Rank	Numbers	2008-cohort	2009-cohort
-4	5	0	5
-3	65	21	44
-2	292	106	186
-1	1,190	499	691
0	2,827	1,397	1,430
1	1,238	700	538
2	389	207	182
3	58	33	25
4	6	4	2

Table A5
Associations between student achievement gains and tutors' instruction

	Low achievers	Medium achievers	High achievers
Pretest, ind.	0.643***	0.411***	0.742***
	(0.0457)	(0.132)	(0.0502)
Pretest, average	-0.353***	-0.342***	-0.341***
	(0.0986)	(0.0838)	(0.0726)
Studcent1	0.0641	-0.0111	0.0003
	(0.0433)	(0.0389)	(0.0295)
Problems	-0.0159	-0.0707*	0.0234
	(0.0358)	(0.0402)	(0.0316)
Automat	-0.107**	0.0301	-0.0581*
	(0.0419)	(0.0402)	(0.0299)
HH	-0.0911	-0.0408	0.135*
	(0.0916)	(0.0818)	(0.0811)
HL	-0.671***	0.0826	-0.107
	(0.192)	(0.287)	(0.129)
LL	-0.255*	0.0397	0.00694
	(0.129)	(0.102)	(0.0729)
Constant	0.360	0.350*	0.0741
	(0.220)	(0.314)	(0.156)
N	1283	1093	1343
R^2	0.213	0.044	0.132

Notes: Dependent variable is standardized individual posttest score. The tutor types are as identified by their approach to low achievers. Tutor type LH is the reference category for the individualization variables. The variables "Problems" and "Automat" indicate whether the students spend much time on problem solving and automatization of arithmetic operations respectively. Robust standard errors in parentheses. ***p<0.001, **p<0.05, *p<0.1

References

Barnes, Marcia A., Alice Klein, Paul Swank, Prentice Starkey, Bruce McCandliss, Kylie Flynn, Tricia Zucker, Chun-Wei Huang, Anna-Maria Fall, and Greg Roberts. 2016. Effects of tutorial interventions in mathematics and attention for low-performing preschool children. *Journal of Research on Educational Effectiveness*, *9*, no.4: 577-606.

Bertoni, Marco, and Marco Nisticò, R. 2023. Ordinal rank and the structure of ability peer effects. *Journal of Public Economics*, *217*, 104797

Betts, Julian R., and Jamie L. Shkolnik. 1999. The behavioral effects of variations in class size: The case of math teachers. *Educational Evaluation and Policy Analysis* 21, no. 2: 193-213.

Bloom, Benjamin S. 1984. The 2-sigma problem: The search for methods of group instruction as effective as one-on-one tutoring. *Educational Researcher*, 13, no.6: 4–16

Bonesrønning, Hans, Henning Finseraas, Ines Hardoy, Jon Marius V. Iversen, Ole Henning Nyhus, Vibeke Opheim, Kari V. Salvanes, Astrid M. J. Sandsør, and Pål Schøne. 2022. Small group instruction to improve student performance in mathematics in early grades: Results from a randomized field experiment, *Journal of Public Economics*, 216, 104765

Brown, Byron W., and Daniel H. Saks. 1986. Measuring the effects of instructional time on student learning: Evidence from the beginning teacher evaluation study. *American Journal of Education***,** no. 94: 480-500

Brown, Byron W., and Daniel H. Saks. 1987. The microeconomics of the allocation of teachers' time and student learning. *Economics of Education Review*, 6, no. 4: 319-332

Clarke, Ben, Christian T. Doabler, Derek Kosty, Evangeline Kurtz Nelson, Keith Smolkowski, Hank Fien, and Jessica Turtura. 2017. Testing the efficacy of a kindergarten mathematics intervention by small group size. AERA open, 3, no,2.

Clements, Douglas H., Roberto Agodini, and Barbra Harris. 2013. *Instructional practices and student achievement: Correlations from a study of math curricula* (NCEE Evaluation Brief). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Davis, Jonathan M.V., Jonathan Guryan, Kelly Hallberg, and Jens Ludwig. 2017. *The economics of scale-up*, NBER Working Papers 23925, National Bureau of Economic Research, Inc.

Delaney, Judith M., and Paul J. Devereux. 2021. High school rank in math and English and the gender gap in STEM. *Labour Economics* 69.

Denning, Jeffrey T., Richard Murphy, and Felix Weinhardt. 2021. Class rank and long-run outcomes. *The Review of Economics and Statistics*, 1-45.

Dietrichson, Jens, Martin Bøg, Trine Filges, and Anne-Marie Klint Jørgensen. 2017. Academic interventions for elementary and middle school students with low socioeconomic status: A systematic review and meta-analysis. *Review of Educational Research*, 87, no. 2: 243-282.

Elsner, Benjamin, Ingo E. Isphording, Ulf Zölitz. 2021. Achievement rank affects performance and major choices in college. *The Economic Journal*, *131*(640), 3182-3206.

Gersten, Russell, Sybilla Beckmann, Benjamin Clarke, Anne Foegen, Laurel Marsh, Jon R. Star, and Bradley Witzel. 2009. Assisting students struggling with mathematics: Response to intervention (RtI) for elementary and middle schools. Institute of Education Sciences (ED), National Center for Education Evaluation and Regional Assistance; What Works Clearinghouse (ED).

Gersten, Russell, Eric Rolfhus, Ben Clarke, Laura E. Decker, Chuck Wilkins, and Joseph Dimino. 2015. Intervention for first graders with limited number knowledge: Large-scale replication of a randomized controlled trial. *American Educational Research Journal*, 52, no. 3: 516-546.

Guryan, Jonathan, Jens Ludwig, Monica P. Bhatt, Philip J. Cook, Jonathan M.V. Davis, Kenneth Dodge, George Farkas, Roland G. Fryer Jr., Susan Mayer, Harold Pollack, and Laurence Steinberg. 2021. Not too late: Improving academic outcomes among adolescents. NBER Working Paper No. 28531.

Imbens, Guido. 2011. Experimental Design for Unit and Cluster Randomized Trials. International Initiative for Impact Evaluation Paper.

Kane, Thomas J., Eric S. Taylor, John H. Tyler, and Amy L. Wooten. 2011. Identifying Effective Classroom Practices Using Student Achievement Data. *Journal of Human Resources*, 46, no. 3: 587-613.

Kraft, Matthew, and John P. Papay. 2014. Can professional environments in schools promote teacher development? Explaining heterogeneity in returns to teaching experience. *Educational Effectiveness and Policy Analysis*, 36, no. 4 :476-500.

Lazear, Edward. 2001. Educational production. *The Quarterly Journal of Economics*, 116 no. 3: 777–803.

Morgan, Paul L., George Farkas, and Steve Maczuga. 2015. Which instructional practices most help first-grade students with and without mathematics difficulties*? Educational Evaluation and Policy Analysis*, 37 no. 2: 184–205.

Murphy, Richard, and Felix Weinhardt. (2020). Top of the class: The importance of ordinal rank. *The Review of Economic Studies*, 87(6), 2777-2826.

Nickow, Andre, Philip Oreopoulos, and Vincent Quan. 2020. The impressive effects of tutoring of preK12 learning: A systematic review and meta-analysis of the experimental evidence. NBER Working Paper No. 27476.

Rockoff, Jonah E. 2004. The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review*, 94, no. 2: 247–52.

Ryan, Allison M., Helen Patrick, and Sungok S. Shim. 2005. Differential profiles of students identified by their teacher as having avoidant, appropriate, or dependent help-seeking tendencies in the classroom. *Journal of Educational Psychology*, 97, no. 2: 275–285.

The National Mathematics Advisory Panel. 2008. Foundations for success: The final report of the National Mathematics Advisory Panel. US Department of Education.

Tomlinson, Carol Ann, Catherine Brighton, Holly Hertberg, Carolyn M. Callahan, Tanya R.Moon, Kay Brimijoin, Lynda A. Conover, and Tomothy Reynolds. 2003. Differentiating instruction in response to student readiness, interest, and learning profile in academically diverse classrooms: A review of literature. Journal for the Education of the Gifted. 27 (2/3): 119–145

Tomlinson, Carol Ann. 2015. Teaching for excellence in academically diverse classrooms. Society 52: 203–209.