Pre-Analysis Plan

The Effect of Small Group Instruction in Mathematics for Pupils in Lower Elementary School: Results from a Randomized Field Experiment

Bonesrønning, Hans (NTNU)
Finseraas, Henning (ISF, corresponding author: henninfi@samfunnsforskning.no)
Hardoy, Ines (ISF)
Iversen, Jon Marius Vaag (NTNU)
Nyhus, Ole Henning (NTNU)
Opheim, Vibeke (NIFU)
Salvanes, Kari Vea (NIFU)
Sandsør, Astrid Marie Jorde (NIFU)
Schøne, Pål (ISF)

Introduction

There is a large literature on the effect of educational resources on student performance, where a particular focus has been the role of class size. Despite marked improvements in data quality and the development of a range of new empirical strategies, the evidence on the importance of resources and class size on student performance remains inconclusive. While many studies using the now famous RCT Project STAR report positive long and short term effects of smaller class sizes (13-17 students versus 20-25), see e.g. Schanzenbach (2006) for an overview, studies using different natural experimental approaches find inconclusive results (e.g. Angrist and Lavy, 1999; Hoxby, 2000; Browning and Heinesen, 2007; Fredriksson and Öckert, 2008).[1] There is growing evidence, however, that students in early grades appear to gain more from smaller classes than older students (Jepsen and Rivkin 2009).

The class size literature is somewhat out of step with current trends in the staffing of classrooms. In many countries, Norway included, we have witnessed large increases in the use of additional teachers, teacher assistants and special education teachers (Bonesrønning et al. 2011). A pertinent question is whether these kinds of targeted resources lead to improvements in student performance.

The earliest evidence, reported from the STAR-experiment in the 1980's, suggests no beneficial effects on student performance from having a teacher assistant in kindergarten to Grade 3 (e.g. Finn and Achilles, 1999). A small number of recent empirical analyses report similar findings. For instance, Reynolds and Muijs (2003) use data from British primary schools to show that students who receive Numeracy Support Assistance do not make more progress in mathematics than those who do not receive assistance. Blatchford et al. (2012) and Webster et al. (2013) find that students who receive the most support from teaching assistants have less engagement with a qualified teacher and less achievement gains than similar students who receive less support from assistants.

Some of the researchers involved in the above mentioned research make the following two claims. First, the negative performance effects associated with teacher assistants reflect that the assistants are used in informal, unsupported instructional roles. Second, assistants

[1] A handful of studies analyze Norwegian data. Leuven et al. (2008) use a quasi-experimental approach to analyze the importance of class size in the Norwegian lower secondary school. Identification is based on maximum class-size rules and population variation. Results show no class-size effect. Iversen and Bonesrønning (2013) use data from the Norwegian elementary school to test whether students from disadvantaged backgrounds benefit from smaller classes. The previous Norwegian class size rule of maximum 28 students is used to generate credible exogenous class size variation. They find significant class size effects for the subgroup of students with parents who are educated at or below the upper secondary school level, and for the subgroup of students from dissolved families.

might contribute positively if the interventions are designed properly.[2] In summary, the literature suggests that interventions by additional teachers need to follow a number of specific guidelines to be successful. We mention a few here (see Sharples et al. 2015): A pull-out strategy should be used. The sessions should be brief (20-50 minutes) and maintained for several weeks. The intervention should have structured supporting resources and lesson plans. Assessments should be used to identify appropriate students, 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.

We conduct a large-scale RCT intervention of the effect of small group teaching which incorporates the insights and recommendations drawn from the existing literature. The RCT intervention covers 163 schools, almost equally divided between treatment and control group schools (80 in treatment group). The RCT takes place in ten large municipalities in Norway, geographically spread from the south-west to the northern region. The participating municipalities are: Asker, Bærum, Bodø, Drammen, Sandefjord, Sarpsborg, Stavanger, Tromsø, Trondheim and Ålesund. Concentrating the project in large municipalities which are mainly densely populated areas is to ensure sufficient supply of qualified teachers in mathematics. Moreover, it ensures a sufficient number of schools from each municipality, as randomization is conducted within each municipality. The total number of public primary schools permits robust matching and randomization processes, as described below. The implementation of the intervention will be closely monitored through surveys, monitoring forms and case studies.

Treatment

School leaders in treated schools will be allocated an additional teacher man-year which they are instructed to use for small-group tutoring in mathematics in grades specified by the project (see below). The schools will be instructed to use a pull-out strategy where small groups of

[2] Slavin et al. (2011) report that combining a strong focus on improving classroom instruction with providing one-to-one, phonetic tutoring to students who continue to experience difficulties in reading, leads to better student performance in reading for 5 – 10 year-old students. Moreover, Dobbie and Fryer (2013) and Fryer (2014) find beneficial effects of adding teaching assistants. First, they show that successful charter schools are characterized by using "high-dosage" tutoring. Second, in a field experiment implementing best practice from charter schools, including of high-dosage tutoring, they find positive effects for low-performing public schools. High-dosage tutoring occurs when small groups of students (less than six) are taken out of their ordinary classroom, and meet four or more times per week. Finally, Andersen et al. (2015) report positive average effects on reading scores, but not on math scores, from using teacher aides as classroom co-teachers for 13-year-old Danish students. They relate the insignificant results for math to differences in optimal teaching technology in reading and math and that successful math learning relies more on accurate instructions than learning to read (Murnane, 2002).

students attend tutorials in separate classrooms. The groups will vary over the school year so that all students will have received at least one period of small-group instruction in mathematics by the end of the school year. This pull-out strategy generates the necessary leeway for the additional teacher to customize teaching to the students.

Treatment will consist of sessions, where the students meet three to five times each week, and where treatment is maintained for a minimum of four weeks. The sessions will differ in length, as there are variations in the schools' organization of the mathematics instruction. While some schools have long sessions (up to 90 minutes), others have shorter sessions, most often sessions are of 60 or 45 minutes duration. The project collects data on the duration of each small-group session.

The ambition is that each student will participate in two such treatment rounds during a school year. This will ensure that each student receive a minimum of eight weeks of small-group instruction in mathematics during one school year. We will instruct school leaders to install formative assessments intended to inform the teachers' decisions about which students should attend tutorials, and to track progress. Pull-outs and assessments define the core of the interventions. The researchers will work closely with the school leaders to secure that these elements are implemented.

Existing evidence indicates that for the pull-out strategy to be effective, the additional teachers should coordinate the tutorials closely to regular teaching, and lesson plans should be worked out together. Moreover, the teachers should spend time evaluating the effects of the interventions, and, if necessary, make changes in group compositions during the school year. Close co-operation between the additional teacher and the regular mathematics teachers are among the central success factors.

We will inform the participants that these types of activities are likely to be crucial determinants of success. In addition, a small handbook with detailed instructions to teachers on how to implement the intervention and information about data collection will be provided. The handbook will also contain some information about the characteristics of previous successful interventions using additional teachers.[3]

Adding an additional teacher to a class does not come without complications. If the additional teacher is introduced suddenly, and the school is unable to make adjustments in existing special education or in the use of assistants, the school leader must decide whether i) students that already are offered tutorials should get alternative or additional treatment or ii)

[3] The handbook is available for all teachers in the project, including the teachers at the control schools.

to be left untreated by the project. In such cases, we will instruct the school leaders that the additional teacher man-year is not an alternative to existing arrangements.

If the school leader knows about the hiring of the additional teacher prior to the school year (most cases), s/he might respond by reallocating existing assistants from the "intervention grade" to other grades, implying that a wider range of ordinary students are indirectly affected by the additional teacher. In this case, treatment by the additional teacher will to some extent be an alternative to treatment by assistants. The teacher-student ratio will then differ only marginally between treated and untreated schools, and intervention effects depend solely on the quality differences and the choice of target groups between two types of tutoring.

All treatment schools are informed about the importance of not mixing the use of additional teacher with other teacher or assistant resources (i.e. not reducing the number of assistants in the "intervention grades" or in other ways changing the use of other school resources due to the schools' participation in 1+1). We considered hard restrictions in order to prevent such resource reallocations, but fear that hard restrictions will have detrimental effects on the motivation of the participants, and thus the outcomes of the experiment. Instead, we will analyze data on resource use in schools to facilitate interpretation. To the extent possible, data on the number of students and teacher-resources–from official sources like *Grunnskolens informasjonssystem* (GSI)[4]–will be analyzed to reveal such (and other) unintended effects. We will also collect data on resource allocations as well as use of assistants and additional teachers though annual surveys to mathematics teachers and school leaders. The surveys will be an instrument for analyzing similarities and differences in resource allocations between treatment and control schools during the four year intervention period.

There is a limitation to how many students at each grade level that can be included in the treatment each school year. In small or medium sized schools all student at each grade level will be included. In large schools (more than 48 students or more than two classes at each grade level), only a share of the students can participate. In these cases, the project selects students to the treatment group either by randomly selecting classes or singular students (of the students that are not organized in fixed classes). The selection of students in large schools is necessary to ensure that each student receives sufficient 'dose' of treatment during a school year (2 x 4 weeks minimum, in a group of maximum 6 students). The schools

[4] All schools are obliged to report this information annually to The Norwegian Directorate for Education and Training.

are informed of the regulations of the maximum number of students that may participate each school year.

Additional questions affecting the design of the intervention are how many years the students should be allocated to small group instructions by an additional teacher and when the treatment should start. Because the existing literature offers few guidelines, we experiment with the starting age and of treatment. Figure 1 shows how we plan to use one teacher person-year for four years.

One teacher person-year is allocated to each treatment school for the period of four years. According to standard working conditions the teacher assistant can deal with two classes per grade per year as a minimum (dependent on the size of the school and the number of math lessons per grade). The standard working conditions define a restriction of the intervention. We have given the schools strict instructions regarding the use of the teacher man-year if the work on the project is less than a teacher man-year: They are not allowed to use these hours on cohorts participating in the project.

In the first year the 2nd and 3rd grades are treated by the additional teacher man-year. The second year, the 3rd and 4th grades are treated. In the third year, 2nd and 4th grades are treated, and in the fourth year, 3rd and 4th grades are treated. Thus, after four years, one cohort is treated for one year (4th grade in the final year of the intervention), two cohorts are treated for two years (starting in the 3rd grade year 1, and starting in the 2nd grade year 3), and one cohort is treated for three years (starting in the 2nd grade year 1). In total four cohorts are included in the intervention. The design allows for ample analyses of students' age, duration of treatment and implementation quality during the four years of interventions.

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

Figure 1: Project plan for use of one teacher man-year for four years (T=treatment).

Hypotheses

Our main hypothesis is that small group instructions improve test scores in math. We will test the effect of small group instructions on the grade on the national test in fifth grade.[5] Our primary hypothesis is to test the treatment effect across all cohorts in a pooled analysis. In addition, we estimate four additional treatment effects, one for each cohort. We estimate one treatment effect per cohort since the age when receiving treatment and the duration of treatment varies across cohorts. We consider the analysis of cohort-differences in the treatment effect as a secondary hypothesis.

In addition, we have a secondary hypothesis of treatment heterogeneity according to baseline characteristics. More specifically, we will analyze heterogeneity according to i) baseline ability, as measured in pre-tests and ii) gender. We expect missing data on baseline test scores (see below). If missing data on baseline test scores is above 20 percent in the control schools, we will replace the analysis of heterogeneity on baseline ability with an analysis of heterogeneity on parental level of education, as reported in the register data. If so, we will measure parental level of education as a dummy of whether any of the parents have a post-secondary education.

Key data sources

Administrative register data will be linked to students at the individual level. The register data is collected and organized by Statistics Norway (SSB). From the registers we will have information on school, gender, country of birth, test results from the National test in the 5th grade, as well as parental level of education and parental country of birth. Parental level of education will be measured when the student was five years of age. The parental level of education will be measured by five dummy variables: Primary education, upper secondary education, higher education lower level, higher education higher level, and unknown education. Except for those born in Norway, place of birth will be measured by continent dummy variables.

In addition, the project collects its own data. Most importantly, the project collects pre-tests at the beginning of the treatment periods and post-tests at the end of the treatment periods. These tests are developed by the research team. To describe the field experiment and monitor how the schools implemented the treatment, we conduct surveys of teachers and

[5] It will be possible to examine effects on eight grade test scores as well. The project intends to apply for additional funding to conduct such analyses. If approved, the analysis will follow this pre-analysis plan as closely as possible.

principals. In these surveys we collect information to capture teacher quality; work experience, seniority, and length and type of education (e.g., whether they have mathematical specialization). We collect information about the teaching environment, job satisfaction, the recruitment process of the treatment teacher, and the cooperation between the main teacher and the treatment teacher. We further collect information about math teaching, how math instruction is organized, time use, attitudes towards teaching mathematics and recollections of the previous mathematics lesson.

Randomization

We conducted randomization at the school level such that each school is randomized into either the control group or the treatment group. The randomization is conducted within each municipality. We decided to conduct randomization at the school level for two main reasons. First, school leaders may be reluctant to participate in an experiment where similar students are treated differently within the school. Second, it is more challenging to keep the control group unaffected by the treatment if this group is within the same school as students receiving the treatment.

We conducted the stratified randomization in the following manner. Within each municipality we ranked the schools based on their mean test score in the national math tests at the fifth grade.[6] Next we constructed a set of strata of at least four schools in each strata. In doing so, we follow Imbens' (2011) recommendation to have at least two treatment and control schools in each strata, so that one can derive a within-strata variance in the treatment effect. The strata sizes range from 4 to 7. Most strata consist of 4 or 6 schools. In the municipalities of Asker, Sarpsborg and Ålesund we had an uneven number of schools who volunteered to participate in the project, which resulted in one strata in each municipality consisting of 7 schools. Next we randomized schools to the treatment or the control group by using the random number generator in Stata. One school refused to participate after their treatment status was revealed. Since we have outcome test scores from the register data, the only implication is that we miss pre-test scores for all students from the school that withdrew from the project. For the analyses where we rely on pre-test scores we will present results without this school's strata. If a school closes during the project period we will exclude data from all schools in that school's strata for the cohorts after closure.

[6] We averaged over the mean score in the two preceding school years to reduce measurement error.

The treatment schools were given one additional teacher-man-year.[7] As all schools received one teacher-man-year this implied that the smallest schools in our sample–with 20 pupils per grade level–will have a much larger increase in student-teacher ratio than the larger schools–the largest school has about 70 students in each grade. Larger schools will not be able to obtain sufficient treatment intensity if they spend the teacher-man-year on all students. We therefore decided to randomize groups of students into treatment at these schools. We do so by randomizing classes to treatment.

Identification of the treatment effects

We identify the pooled treatment effect using the following regression model:

$$Y_{icgs} = \beta Treated_{sg} + \alpha_s + \mu_c + \epsilon_{icgs} \qquad (1)$$

where *i* indexes individuals, *c* cohorts, *g* schools, and *s* indexes randomization strata. *Y* is the standardized test score in the national tests in fifth grade, *Treated* is an indicator of whether the student belongs to a school in the treatment group when s/he is enrolled in school, μ_c is cohort fixed effects, and α_s is strata fixed effects. The strata fixed effects are included because treatment status is random within strata. Standard errors are adjusted for clustering at the school level because treatment is at the school level. β represents the treatment effect.

β is an unbiased estimate of the treatment effect due to the randomization of treatment. However, to potentially increase the precision of the β estimate we in addition utilize our pre-test information:

$$Y_{icgs} = \beta Treated_{sg} + \gamma_1 Pre_{icgs} + \gamma_2 Miss_{icgs} + \gamma_3 Pre_{icgs} * Treated_{sg}$$
$$+\gamma_4 Miss_{icgs} * Treated_{sg} + \ \alpha_s + \mu_c + \epsilon_{icgs} \ (2)$$

We standardize the pre-test score and include it as a covariate *Pre*. We control for baseline test score as a continuous variable, but will also present results using a flexible specification where we divide the students into quartiles. In the first year of data collection we find that the share of students taking the pre-test differs between treatment and control schools. Our preliminary analysis suggests that, in our first cohort, on average 97 percent of students in the treatment schools took the test, while 90 percent in the control schools took the test. We hope to improve these numbers, but suspect that we will see differences in later cohorts as well. We address the missing test score issue by assigning all students with missing pre-test scores a zero on the test and include a dummy variable *Miss* for missing test-scores (see Gerber and

[7] Schools were instructed to split the teacher-man-year between no more than two teachers. About one third (26) of the schools did so.

Green 2012: 241, Lin et al. 2016).We do the same for students with missing parental consent to merge the pre- and post-test scores, i.e. we set the pre-test to missing for these students.[8] We follow Lin et al. (2016) and i) mean-center *Pre* and *Miss*, ii) include interactions between *Pre* and *Miss* and the treatment indicator, because the interactions might improve precision. The mean-centering of *Pre* and *Miss* ensures that β represents the average treatment effect (see Lin et al. 2016: 14). We follow this procedure also when we include other covariates.

The treatment effect estimated in equation (1) is the intention to treat (ITT) effect. The ITT will differ from the average treatment effect (ATE) if students in the treatment schools did not receive treatment or if students in the control group receive treatment. We identify the (local) ATE by using *Treated* as an instrumental variable for the receipt of treatment in a 2SLS set up. We have information on the receipt of treatment at the individual level from the teacher surveys.

The cohort-specific treatment effects will be estimated by running the analysis separately for each cohort.

Balance tests

The balance tests will be conducted using the same empirical specification we use to test the treatment effect (equation 1). We will conduct an F-test of the joint significance of the variables we study balance on using a regression with the treatment indicator as the dependent variable.

We will study balance on the following variables:

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

Parental level of education: The level of education will be measured by five dummy variables: Primary education, upper secondary education, higher education lower level, higher education higher level, unknown education.[9] We will rely on the highest education level of the parents.

First generation immigrant: Dummy equal to 1 if the student is not born in Norway.

[8] We need parental consent to merge the pre-test scores, the post-test scores, the national test scores and the register data. The preliminary, incomplete data on parental consent suggest that we will have higher level of parental consent in the treatment schools.

[9] 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, categories 3,4,5 to upper secondary education, category 6 to higher education lower level , categories 7, 8 to higher education higher level, and category 9 to unknown education.

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.

Teacher-to-student-ratio: Measured as ratio of teachers and assistants to the total number of students.

Robustness checks

We will conduct at least the following robustness checks:

i) include the set of pre-treatment variables for which we find imbalance.

ii) include all pre-treatment variables listed above.

iii) present results where we aggregate the results to the school level and weight the schools according to the number of students.

iv) present results using randomization inference to derive p-values.

Treatment heterogeneity

We will study heterogeneity depending on the baseline test score and gender. We do so by expanding equations (1) and (2) with gender and the baseline test score variable and interactions with *Treated.* The interaction terms will test whether *Treated* depends on baseline ability or gender. Due to missing observations in the baseline test, a potential worry is a systematic relationship between missing observations and treatment across the ability distribution. We will examine the distribution of pre-test scores in the treatment and control group to examine this potential problem. If this analysis indicates that there is a relationship between missing observations and treatment across the ability distribution, we will re-weight the sample by applying entropy balancing (Hainmueller 2012) on baseline test scores. The matching will be done by cohort-year using the raw test score. Ideally, we balance on mean, standard deviation and skewness of the test score, but if doing so implies that we have to use extreme weights, we will instead coarsen the test-score that we balance on. We will always coarsen as little as possible to maximize achieved balance.

Multiple outcomes

The main objective is to evaluate the pooled ITT treatment effects for one primary outcome. We will not adjust the p-value for the test of this hypothesis. However, in addition we have secondary hypotheses where we test i) variation across the four cohorts, and ii) treatment heterogeneity on two variables (baseline test score and gender). When testing these secondary hypotheses, there is a risk that some hypotheses will be statistically significant by chance alone. In order to deal with the problem of multiple comparisons we impose pre-specified decision rules (Rosenblum and van der Laan 2011). We follow the recommendations of Fink, McConnell, and Vollmer (2014) and use Benjamini and Hochberg (1995) and Benjamini and Yekutieli's (2001) approach to minimize the false non-discovery rate (see also Almeida 2012). We conduct these adjustments separately for the two set of secondary analyses (cohort variation and treatment heterogeneity), but all results and p-values will be reported for researchers who want to adjust p-values across the two set of analyses.

The false discovery rate (FDR) method developed by Benjamini and Hochberg (1995) implies that the m p-values of the i hypotheses are ordered from low to high and that the adjusted p-value is found using the formula $p(i) = a \times i/m$. Here, a refers to the p-value threshold, for instance 0.05. For instance, assume we have four hypotheses with the following ordered p-values: $p(1) = 0.002$, $p(2) = 0.0126$, $p(3) = 0.040$, $p(4) = 0.051$. Without any correction for multiple testing, the first three null hypotheses would be rejected using the 5 percent significance threshold. With the FDR method the decision rule implies that two hypotheses would be rejected: $p(1) = 0.002<0.05 * 1/4$, $p(2) = 0.0126<0.05 * 2/4$, $p(3) = 0.04 >0.05 *3/4$, $p(4) = 0.051>0.05 * 4/4$. This approach is an effective way of limiting the risk of false discoveries, while retaining a higher level of statistical power compared to the Bonferroni-correction where all hypothesis face the same threshold of $p=a/m$.

Selective migration

Parents and students were informed about the randomization at the beginning of the school year 2016/2017. Thus, students enrolled in later school years knew the treatment status prior to enrollment. The admission system in most municipalities is based on a strict neighborhood rule (Opplæringsloven §8-1), meaning that the location of residence determines the school the student is enrolled in. There are, however, a few municipalities in our sample that allow school choice. We will therefore investigate if there are systematic differences over time in the extent of students residing outside the school district in treated and control schools. If there are no systematic differences over time in municipalities with school choice, there is no reason to expect selection in municipalities with neighborhood rules. If, however, we find systematic differences in enrollment in municipalities with school choice, we will separate cohorts as a robustness check to investigate if treatment effects are influenced by selection bias.

Attrition

By attrition we refer to the problem of selective exit from the population we study or missing outcome data which is non-random. We use register data to gather information on achievement levels. Since register data is comprehensive, missing data is less of a concern. As long as the students are still living in Norway, we will be able to gather information on them even if they moved to a different school district. A small percentage of students have no reported test score on the national test. As long as this type of attrition is random with respect to the treatment status, it will not bias our results and will only to a limited extent affect the power of the experiment. We will report the attrition level by treatment status to examine this type of bias. The analysis will follow the empirical specification we use to identify the treatment effect. If we find that the bias is correlated with treatment status we will calculate extreme bounds and trimming bounds for the treatment effect for the always-reporters (see Gerber and Green 2012: 226ff).

Power calculations

We used the clustersampsi command in Stata (Hemming and Marsh 2013) to calculate power. Based on the preliminary data from our own first year tests we set the ICC within strata to .066 and assume that the strata fixed effects and the pre-test predicts about 34 percent of the post-test. With a five percent significance level we have 80% power to detect a treatment effect of .12. Based on experience from analyses of national test data, we suspect that the pre-test and the strata fixed effects will be more predictive of the national test, which will improve power. In any case, we are well-powered compared to the treatment effect sizes in Fryer (2014), but slightly low powered compared to the effect sizes in Andersen et al . (2015).

Archive

The pre-analysis plan is archived at the EGAP (Evidence in Governance and Politics) registry: http://egap.org/content/registration.

References

Andersen, C., Beuchert, L., Skyt Nilsen, H., Thomsen, Kjærgaard, M. (2015), The Effect of Teacher Aides in the Classroom: Evidence from a Randomized Trial. Manuscript

Angrist, J. D., and V. Lavy (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. Quarterly Journal of Economics 114, 533–575.

Blatchford, P., Russell, A. and Webster, R. (2012) Reassessing the impact of teaching assistants: How research challenges practice and policy. Oxon: Routledge

Bloom, B.S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational Researcher, 13, 3-16.

Bonesrønning, H., J.M.V. Iversen, and I. Pettersen (2011) Kommunale skoleeiere: Nye styringssystemer og endringer I ressursbruk. SØF-report 05/11

Brophy, J.E. og T.L. Good (1986). Teacher behavior and student achievement. In M. Wittrock (red.), Handbook of research on teaching (3. utg. s. 328–375). New York: Macmillan.

Browning, M., and E. Heinesen (2007). Class size, teacher hours and educational attainment. Scandinavian Journal of Economics 109, 415-438.

Dobbie, W. and Fryer R.G. (2013) Getting Beneath the Veil of Effective Schools: Evidence from New York City. American Economic Journal: Applied Economics, 5(4):28-60.

Duflo, E., R. Glennester & M. Kremer, (2008): "Using Randomization in Development Economics Research: A Toolkit," Handbook of Development Economics, 4, 3896-3961.

Finn, J.D, and C.M. Achilles (1999) Tennessee's Class Size Study: Findings, Implications, and Misconceptions. Educational Evaluation and Policy Analysis, 21(2), 97-109

Fredriksson, P., and B. Öckert (2008). Resources and student achievement - evidence from a Swedish policy reform. Scandinavian Journal of Economics 110, 277-296.

Fryer, R.G. (2014) Injecting Charter School Best Practices into Traditional Public Schools: Evidence From Field Experiments. Quarterly Journal of Economics, 129(3):1355-1407.

Gerber, A.S. and D.P. Green (2012) Field Experiments. New York: W.W. Norton.

Grønmo, L.S., Onstad, T, Nilsen, T., Hole, A., Aslaksen, H., Borge, I.C. (2012). Framgang, men langt fram. Norske elevers prestasjoner i matematikk og naturfag i TIMSS 2011. Oslo: Akademika forlag.

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.

Hanushek, E.A., 2002. Publicly Provided Education'. In: Auerbach, A.J., Feldstein, M. (Eds.), Handbook of Public Economics. vol. 4. Elsevier, Amsterdam, pp. 2045–2141.

Hanushek, E.A.,Woessmann, L., 2008. The role of cognitive skills in economic development. Jornal of Economic Literature 46, 607–668.

Hanushek, E.A.,Woessmann, L., 2012. Do better schools lead to more growth? cognitive skills, economic outcomes, and causation. Jornal of Economic Growth 17, 267–321.

Hattie, J.A.C. (2006). The paradox of reducing class size and improved learning outcomes. International Journal of Education, 42, 387-425.

Hattie, J.A.C. (2008). Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement. NY: Routledge.

Hemming, K., and J. Marsh (2013). A menu-driven facility for sample-size calculations in cluster randomized controlled trials. Stata Journal 13(1), 114-135.

Hoxby, C. M. (2000). The effects of class size on student achievement: New evidence from population variation. Quarterly Journal of Economics 115, 1239–1285.

Imbens, G. (2011). Experimental Design for Unit and Cluster Randomized Trials. Manuscript.

Iversen, J. M. V., and H. Bonesrønning (2013). Disadvantage students in the early grades: Will smaller classes help them? Education Economics 21, 305-324.

Jepsen, C. and S. Rivkin (2009). Class Size Reduction and Student Achievement, Journal of Human Resources 44(1): 223-250.

Leuven, E., H. Oosterbeek and M. Rønning (2008). Quasi-experimental estimates of the effect of class size on achievement in Norway. Scandinavian Journal of Economics 110, 663–693.

Lin, W., D. P. Green, and A. Coppock (2016). Standard operating procedures for Don Green’s lab at Columbia. Version 1.05, June 8, 2016.

Mourshed, M., C. Chijioke & M. Barber (2010). How the World’s most improved schools systems keep getting better. London: McKinsey & Company.

Murnane, R., I. Sawhill and C. Snow (2012), Literacy Challenges for the Twenty-First Century: Introducing the Issue. Futures of Children 22(2) 3-15.

OECD (2005). Teachers Matter: Attracting, Developing and Retaining Effective Teachers. Paris: OECD.

Reynolds, D., and Muijs, 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

Schanzenbach, Diane W. “What Have Researchers Learned from Project STAR?” Brookings Papers on Education Policy (2006), 205–228.

Sharples, J., Webster, R., and Blatchford, P. (2015) Making Best Use of Teaching Assistants Guidance Report The Education Endowment Foundation

Slavin, R. E., Lake, C., Davis, S., & Madden, N. (2011). Effective programs for struggling readers: A best-evidence synthesis. Educational Research Review, 6: 1-26.

Webster, R., Blatchford, P., and Russell, A. (2013) Challenging and changing how schools use teaching assistants: findings from the Effective Deployment of Teaching Assistants project. School Leadership & Management, Vol. 33, No. 1, 78-96