AARHUS UNIVERSITY

Coversheet

This is the accepted manuscript (post-print version) of the article.

Contentwise, the accepted manuscript version is identical to the final published version, but there may be differences in typography and layout.

How to cite this publication

Please cite the final published version:

Andersen, S. C., Beuchert-Pedersen, L. V., Nielsen, H. S., & Thomsen, M. K. The Effect of Teacher's Aides in the Classroom: Evidence from a Randomized Trial. *Journal of the European Economic Association*. <https://doi.org/10.1093/jeea/jvy048/5232116>

Publication metadata

Title:	The Effect of Teacher's Aides in the Classroom: Evidence from a Randomized Trial
Author(s):	Andersen, S. C., Beuchert-Pedersen, L. V., Nielsen, H. S., & Thomsen, M. K.
Journal:	Journal of the European Economic Association
DOI/Link:	10.1093/jeea/jvy048/5232116
Document version:	Accepted manuscript (post-print)
Document license:	

General Rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

If the document is published under a Creative Commons license, this applies instead of the general rights.

The Effect of Teacher's Aides in the Classroom: Evidence from a Randomized Trial*

Simon Calmar

Louise

Helena Skyt

Mette Kjærgaard

Andersen

Beuchert

Nielsen

Thomsen

sca@ps.au.dk

lobe@vive.dk

hnielsen@econ.au.dk

mtho@sam.sdu.dk

Abstract: Teacher's aides are used worldwide, in various school systems, although, there is no strong evidence of their impact on student outcomes. We use a randomized trial to challenge this state of evidence. We randomly allocate 105 schools to two types of treatment – aides with or without a teaching degree – compared to a control group. Both types of aides have positive impacts on test scores and the effects are persistent over time for disadvantaged students. Exploratory analyses of mechanisms suggest that a teacher's aide is not just a class-size reduction, but especially impactful when sharing instructional responsibility for the classroom.

JEL classification: I21, H52

Keywords: School Resources; Academic Achievement; Teacher Quality; Student–Teacher Ratio; Co-teaching.

Acknowledgements: We appreciate comments from the editor and three anonymous referees as well as Peter Blatchford, Paul Devereux, Maria Humlum, Anne Nandrup, and participants at the Family and Education Workshop Hawaii 2015; IWAE Catanzaro 2015; Workshop on Economics of Education Mainz 2015; CNESCO Beijing 2014; and seminar participants at the Paris School of Economics, Edinburgh University, Lund University, University of Uppsala, and Aarhus University. Financial support from the Danish Council for Independent Research (Sapere Aude Starting Grant, 10-079597) and from the Ministry of Education is gratefully acknowledged. The Ministry of Education has defined the objective of the intervention, but they are not responsible for the evaluation or conclusions. The usual disclaimers apply.

1. Introduction

On the one hand, research has consistently demonstrated that teacher quality is more important for student skill development than structural inputs.¹ But increasing teacher quality is easier said than done. On the other hand, structural, input-based education policies – such as reducing class size or increasing instruction time – are politically manageable, but often also inflexible and unfocused and therefore less effective. If, for instance, class size is reduced, it will be fixed for an extended period of time – which in some countries means many years – and potential benefits will be spread across all students in the classroom. Use of teacher's aide offers a middle ground, i.e. a politically manageable policy, but at the same time a much more flexible means of input adjustment. In recent years, teacher's aides have developed from practical assistants to participants in teams where two professionals (e.g. a certified teacher and a certified special education teacher) share instructional responsibility for a single classroom (Friend and Cook 2016). This paper investigates the impact of such teacher's aides on academic achievement.

Teacher's aides can target particular courses or be used in particular classes to reduce class size effectively for a period of time. Furthermore, different kinds of teacher's aides can be employed; more qualified, costly co-teachers with teaching degrees or less-qualified, less-costly teaching assistants that do not have a teaching degree, but may supplement educated teachers by handling other kinds of pedagogical tasks in the classroom. In that sense, teacher's aides can be used more flexibly than

¹See for instance Ladd and Sorensen (2017), Chetty et al. (2014), Aaronson et al. (2007) or Hanushek and Rivkin (2006).

class-size reductions to implement high-dosage tutoring, which seems to be an important component in improving learning among adolescents who are otherwise falling behind (Cook et al. 2015; Fryer 2014; see also Bloom 1984).

The use of teaching assistants has increased in many Western countries with high-spending school-systems. In England, the number of teaching assistants has tripled since 1997, and they comprised 24% of the mainstream school workforce in 2011 (Blatchford et al. 2012). In Iceland, which is also a high-spending country, the teacher–assistant ratio is 4:3, whereas in Italy, which is a lower spending country, it is 60:1 (OECD 2014).

Despite widespread use, intercountry variation in how much aides are used and how they are used, as well as the flexibility of teacher’s aides relative to class-size reductions, few studies have examined the impact of teacher’s aides and those that do fail to find any positive effects. Previous research largely indicates that teacher’s aides are primarily used in disadvantaged schools or classrooms (Cook et al. 2011; Murawski and Swanson 2001). A large-scale British study reported negative associations between the use of paraprofessional teaching assistants and student performance (Blatchford et al. 2012). In a quasi-experimental study, Leuven et al. (2007) exploit a Dutch allocation rule for extra teacher resources in schools with a high proportion of disadvantaged minority students. They find no evidence of extra teacher resources having any positive impact on student achievement. Yet there exists only one large-scale, resource-based intervention—the Student Teacher Achievement Ratio (STAR) field experiment—the evaluation of which reveals no significant effects of teacher’s aides, which may explain

why the STAR experiment has received most attention for another treatment arm, namely class-size reductions.²

In the teachers' aides treatment arm in STAR, teacher's aides were uneducated assistants who spent most of their time on practical tasks e.g. paper work, grading, lunch duty (Gerber and Finn 2001), which is quite far from today's aides who are professionals and paraprofessionals who share instructional responsibility (see Friend and Cook, 2016). Existing research suggests that teacher quality is important. Prior research on teacher value-added finds substantial short- and long-term impacts of being assigned a high value-added teacher for 1 year (Aaronson et al. 2007; Chetty et al. 2014), while research on teacher experience indicates that experience up to at least 5 years improves middle school achievement and behavior (Ladd and Sorensen, 2017).

While studies on the impact of teacher's aides are few, a large body of research is concerned with class-size effects. A teacher's aide is to some extent akin to class-size reduction, because both policies reduce student-teacher ratios, and because an aide *could* be used to effectively reduce class size. Evidence indicates that smaller classes benefit student achievement (Chetty et al. 2011; Fredriksson et al. 2013; Krueger 1999),³ in particular among students from disadvantaged backgrounds (Angrist and Lavy 1999; Schanzenbach 2006; Browning and Heinesen 2007; Fredriksson et al. 2013), and at early grade levels (Nandrup 2016). Teacher's aides is a more flexible

² It should be mentioned that STAR children were much younger than the students in our study. For a detailed description of the experiment, implementation, and main results of STAR, see e.g. Mosteller (1995), Gerber and Finn (2001) and Schanzenbach (2006).

³ This is not found in Norway, though (Leuven et al. 2008).

policy, though, because aides can be used to divide pedagogical task within the classroom and thereby target disadvantaged students.

We contribute to the scant research on teacher's aides by studying the impact of two variations of teacher's aides based on a large-scale, randomized experiment in sixth grade classrooms in Denmark. One treatment uses less-expensive teaching assistants without teaching degrees who, according to collective agreements, require less preparation time. For a fixed budget, this treatment maximizes the time spent with students in the classroom, as the assistants spend 14.5 lessons per week per class. The second treatment uses more expensive co-teachers with teaching degrees who, for the same budget, spend 10.5 lessons per week per class. As in many field experiments, the external validity and general equilibrium effects of the trial will depend on the supply and demand for teacher's aides if the intervention is scaled up. By comparing two randomly allocated types of teacher's aides, however, this study provides evidence on the trade-off between more qualified co-teachers and less qualified assistants with more time in the classroom. In this study, both co-teachers and assistants have some preparation time and mainly support academic achievement, only performing limited practical support. Most teaching assistants have some college (but not a teaching degree), and both types of aides engage in a team of (para)professionals who share instructional responsibility with mutual ownership, pooled resources and joint accountability (see Friend and Cook 2016).

Randomization is conducted at the school level to avoid spill-over effects between classrooms with different kinds of teacher's aides and is based on stratification by schools' predicted outcomes. Our sample includes 105 schools and 249 classrooms,

and the randomization procedure results in 35 schools (1,814 students) allocated to the co-teacher intervention, 35 schools (1,610 students) allocated to the teaching assistant intervention, and 35 schools (1,789 students) allocated to the control group. In order to examine the effect of the teacher aide intervention, we combine survey data with Danish register data, including student IT-based, self-scoring tests, and school enrollment statistics. This enables us to track students who change schools and still include their outcomes in the analysis as well as to make detailed attrition analyses. We survey students, teachers, teacher's aides, and school principals to monitor implementation during the intervention period.

We find statistically significant intention-to-treat effects on reading scores but insignificant effects on math scores. This reflects how the main emphasis in Danish compulsory education is on reading proficiency: more lessons are allocated to language arts compared to math. Our intervention scales up this gap as teacher's aides spend 50% more time in language arts than math classes. The effects tend to be greater when resources are spent on teaching assistants without teaching degrees who—for the same cost—can spend more time in class than co-teachers with teaching degrees, but effects are not significantly different. For students from disadvantaged backgrounds effect sizes are larger and, importantly, seem to persist at the 30-month follow-up. Exploratory analyses of mechanisms indicate that a teacher's aide is not merely akin to class-size reduction or practical assistant, but can support the teacher in classrooms with behavioral problems. We therefore suggest that any conclusion that the use of teacher's aides is ineffective is premature.

The paper unfolds as follows: Section 2 describes the institutional setting in which the teacher's aide intervention is implemented as well as the intervention itself. Section 3 presents the data used for the empirical analysis, including balancing tests between treatment and control groups as well as analysis of attrition. Next, Section 4 describes the estimation strategy, and Section 5 presents the results of the analysis of the impact of the teacher's aide interventions on student achievement. Finally, Section 6 discusses potential channels and mechanisms, while Section 7 concludes.

2. Institutional Setting and Experimental Design

2.1. Elementary School in Denmark

For the students in this study, education was compulsory from the calendar year in which the child turned seven until completing ninth grade. Prior to first grade there was an optional one-year preschool class with 83% enrollment (2005 figures; UNI-C 2012). After completing ninth grade, a student can enroll in either high school or vocational school. The preschool class, compulsory education, and post-compulsory education are free of charge at public schools, but private schools charge tuition fees. The experiment took place in the public school, which catered to 83% of the relevant sixth grade cohort (15% enrolled at a private school and 2% at a segregated special school).

Students are divided into classes when entering preschool and typically remain in the same class until grade nine if they do not change schools. The maximum class size, regulated by national law, is 28 students, whereas the average is 22 students, which is similar to other OECD countries (OECD 2016). In preschool, students are taught by a classroom teacher, whereas they are taught by subject-specific teachers from first to

ninth grade. The same subject-specific teacher typically follows the students through several grades. There are no legal regulations as to the organization of teachers across grades. However, most often teachers specialize in teaching a certain age group and many schools are informally divided into age groups: preschool to third grade, fourth to sixth grade, and seventh to ninth grade. Students in one age group are not taught in the same classroom but they share facilities such as playground, kitchen and toilettes. Most schools provide education until grade nine, but there are some small schools that only offer education until grade six or seven, with the students then finishing their compulsory schooling at another school. This institutional setting implies that class assignments persist unless students move or enroll at a private school. The reason behind this arrangement is a focus on stable peers, social relations and well-being. Most often class assignments persist for ten years meaning that class-size reductions could be extremely expensive.

Educators primarily consist of subject-specific teachers and social educators (who face similar labor market conditions). Teachers typically hold a 4-year college degree with qualification to teach three subjects at primary and lower-secondary level. Social educators hold a 4-year college degree with a focus on behavior and social interaction. Social education constitutes the largest degree program in Denmark. Social educators are an integral part of education in Denmark, as they are responsible for daycare, preschool, after-school care, social education and special needs education for children with behavioral problems. The availability of this large group of professionals means that the Danish setting is well-suited to study the time/quality trade-off when a teacher's aide is not just an unskilled assistant.

The school year consists of 200 school days from early August until late June. The minimum number of lessons taught in the humanities (338 hours/year on average across grades 4–6) is almost twice the minimum in the sciences (171 hours/year). The actual number of lessons offered is decided at the municipal level.⁴ In both public and private schools, students complete the compulsory centralized final exams at the end of grade nine (around age 16).

2.2. Intervention

The background for the intervention dates back to the general election in the autumn 2011. The Social Democrats promised to instate teacher's aides in all elementary school classes if they came into power. However, when they came into office they instead initiated two parallel experiments to evaluate the impact of teacher's aides. Based on the results of the evaluation, the local authorities (who define the overall budgets for the local schools) and the self-governing local schools were free to reallocate resources (see Andersen et al. 2014).

The intervention consisted of assigning a teacher's aide to the sixth grade classroom in all of the classes at a given school. The intervention was in place for roughly 85% of the school year, from October 1, 2012 to June 20, 2013. The intervention and control groups were announced on August 15, 2012, leaving the schools 1.5 months to search for and employ the respective teacher's aides. The timeline of initiation, announcement, and implementation of the experiment is found in Table

⁴ For further detail, see the Danish Public School Act.

A.1 in Appendix A. The objective of the intervention was to improve academic achievement as well as wellbeing and behavior.

Deciding what kind of teacher's aide to employ involves a trade-off between more qualified, experienced, and costly teacher's aides versus less-qualified, less experienced, and less costly aides. Our study examines the impact of two types of teacher's aides, varying in terms of qualifications and time intensity:⁵

- 1) A co-teacher with a teaching degree (minimum 10.5 lessons per week per class)*
- 2) A teaching assistant without a teaching degree (minimum 14.5 lessons per week per class)*

The treatments are equally costly (\$25,000 per class), and the impact assessment therefore indicates which treatment is most cost-effective. If high time intensity is more important than high qualifications, a teaching assistant will be more cost-effective than a co-teacher with a degree; and vice versa if qualifications are more important than time intensity.

The co-teacher intervention represents a reduction in the student–teacher ratio by 27%, while the teaching assistant intervention represents a reduction by 34%.⁶ In

⁵ Teacher's aides holding a teaching degree are denoted co-teachers, as they are expected to co-organize preparation time as well as classroom teaching (union-negotiated). Therefore, teachers are more expensive.

⁶ These numbers are based on an assumption of 28 lessons/week and a student–teacher ratio of 22:1 without the treatment (a lesson lasts 45 minutes).

budgetary terms, these changes correspond to class-size reductions from 22 to 16 students and from 22 to 14.5 students, respectively, which is almost identical to the changes applied in STAR.

The requirements in terms of how to make effective use of the teacher's aides were extremely flexible, allowing the school principal to take into account the particular group of students in each relevant sixth grade cohort. Table 1 reports the minimum requirements set by the experimental protocol along with the actual time use, tasks, and qualifications of the teacher's aides as reported on surveys collected from school principals, teachers, and the aides themselves. On average, the co-teachers spent 11.6 lessons per week per class ($SD = 3.4$) and the teaching assistants spent 20.7 lessons per week per class ($SD = 8.9$). On average, they spent more time per class than the minimum requirement because the hired personnel were less costly than predicted in terms of salary and centrally negotiated preparation time. Teacher's aides were present in most language arts and math lessons; on average, they were present in 4.5 language arts lessons and three math lessons per week per class.

In stark contrast to STAR, the teacher's aides in our study mainly supported academic achievement and did not supply much practical support.⁷ In the survey 1 month after the beginning of the intervention, teachers were asked to state the primary duty of the teacher's aide in four categories of support related to: i) academic achievement, ii) wellbeing and behavior, iii) special education needs, and iv) practical support. Table 1 shows the distribution of responses. About 80% of the teachers

⁷ Teacher's aides spent 30% of their time doing instructional tasks and 70% of their time on other tasks (e.g. paper work, grading, lunch duty) in STAR (Gerber and Finn 2001).

answered that the primary duty of the teacher's aides was to improve student achievement or wellbeing (avoiding conflict and the like), with student achievement being more dominant for co-teachers with a teaching degree than for teaching assistant without a degree. Only about 7-8% of the teachers answered that the primary duty of the teacher's aide was practical support.

Table 1 also describes how the teacher's aides actually worked in the class. Most co-teachers were used for ordinary classroom teaching as well as teaching groups of students inside the classroom. Only half of them taught students in groups outside the classrooms. These figures suggest that co-teachers were not just used for class-size reductions; they were used more flexibly for different tasks. Teaching assistants were also used for diverse tasks, though much less ordinary classroom teaching. As many as 70% of the aides reported spending time handling conflicts and other disruptions among students. However, there are some disagreements between aides' and teachers' reports as to whether the aides spent time handling conflicts and disruptive students. In Section 6, we return to the question of how activities of teacher's aides relate to improvements in student outcome.

TABLE 1. TEACHER'S AIDES TASKS AND QUALIFICATIONS

	Co-teacher w/degree	Teaching assistant w/o degree
Lessons per week per class		
Minimum requirement	10.5	14.5
Actual (principal-reported)	11.6 (3.4)	20.7 (8.9)
Actual (aide-reported, total)	9.4 (5.0)	17.9 (7.5)
Language arts lessons (aide-reported)	4.6 (5.6)	4.7 (3.6)
Math lessons (aide-reported)	3.2 (3.1)	3.1 (2.4)
Satisfy minimum requirement (principal-reported)	70 classes	71 classes
Main duty (aide-reported/teacher-reported)		
Primary duty: academic achievement	69% / 75%	59% / 51%
Primary duty: wellbeing and behavior	4% / 6%	15% / 29%
Primary duty: special education needs	14% / 12%	11% / 12%
Primary duty: practical support	13% / 7%	15% / 8%
Tasks in class (aide-reported/teacher-reported)		
Ordinary classroom teaching	61% / 55%	9% / 5%
Teaching groups of students outside the classroom	47% / 47%	48% / 31%
Teaching groups of students inside the classroom	71% / 72%	35% / 17%
Handling conflicts and disruptive students	72% / 51%	70% / 31%
Practical tasks and preparation for class	n.a. / 42%	n.a. / 7%
Educational background (aide-reported)		
Minimum requirement	Teaching degree	None
Actual education	100% teachers	29% social educator
	Here of	12% ass. social educator
	42% w/ language arts degree	16% master's degree
	27% w/ math degree	9% student teacher
	25% w/ SEN degree	13% HS graduate
	6% Other or unknown	20% Other or no
Experience (aide-reported)		
Teaching experience	6.7 (8.1)	3.4 (3.2)
Pedagogical experience	4.3 (3.0)	8.2 (5.7)
Teaching or pedagogical experience	6.8 (7.8)	7.0 (5.6)
At least 5 years of teaching experience	45%	24%
New employee	76%	79%
Age in years (N = 153) ^a	34.9 (9.2)	34.9 (10.3)
Educational achievement in compulsory and high school^a		
9th grade GPA (2002–10, N = 31)	0.64 (0.33)	0.38 (0.79)
9th grade, language arts written exam grade (2002–10, N=31)	0.73 (0.55)	0.45 (0.90)
9th grade, math written exam grade (2002–10, N = 31)	0.47 (0.53)	0.26 (1.20)
HS GPA (1978–2011, N = 100)	–0.02 (0.86)	–0.42 (1.01)
HS language arts written exam grade (2005–11, N = 35)	0.00 (0.69)	–0.52 (1.27)
HS math written exam grade (2005–11, N = 29)	0.38 (0.80)	–0.25 (0.77)
HS math study program (N = 100)	34%	25%

Note: Table is based on non-missing survey responses from school principals, teachers, and teacher's aides (1-month survey). In total, 153 aides were employed (80 co-teachers and 73 teaching assistants) of which 118 responded to the 1-month survey. Each aide participated on average in 2.5 classes. ^aBased on register data (in parentheses are reported the number of respondents matched with register data). N: Number of respondents. SEN: Special education needs. HS: High School. N.a.: Not available.

The co-teachers and teaching assistants have different levels of qualifications and experience, as seen at the bottom of Table 1, which most likely affects their approaches to interacting with the students. Co-teachers have on average 6.7 years of teaching experience, and teaching assistants 3.4 years. Co-teachers' main subjects of specialization are: language arts (42%), math (27%), or special educational needs (25%). Teaching assistants have no teaching degree, but instead various educational backgrounds (29% social educator, 16% master's degree, 12% assistant social educator, 9% student teacher, 13% high school graduates, and 20% others). Co-teachers have higher grades from high school and compulsory school than teaching assistants.

2.3. Randomization

The experiment was designed as a three-level cluster randomized controlled trial at the school level (see e.g. Spybrook et al. 2011). This implied that students were randomly assigned to interventions, not as individual units but as groups of units clustered in classrooms and nested within schools. The Ministry of Education initiated two parallel experiments to take place during the school year 2012–13 and invited all 98 municipalities in Denmark to participate. In total, 68 applied for participation. Twelve municipalities were then selected for the teacher's aide experiment, which is the focus of this paper. The municipalities were not randomly selected but chosen to reflect the variation in municipality size, school size, and geography.⁸ Baseline test scores of students at selected schools are close to the population mean and therefore do not indicate any substantial positive or negative selection.

⁸ Municipalities engaged in school consolidations and other experiments were disregarded.

Within each municipality, schools (the randomization unit) were allocated to treatment and control groups by stratified randomization to increase the power of the randomized controlled trial (RCT). The schools were stratified based on their predicted average achievement level for the relevant sixth grade cohort (our subjects of interest).⁹

The randomization resulted in 35 schools (1,814 students) being allocated a co-teacher with a degree, 35 schools (1,610 students) being allocated a teaching assistant without a degree, and 35 schools (1,789 students) being allocated to the control group. A total of 105 schools and 5,213 students were included in the experiment. Figure 1 illustrates the research design including student's participation flow from enrollment to analysis.

As described above, stratification and randomization were applied at the school level, meaning that if one school was treated, all of its sixth grade classes were treated. This approach was used to avoid spillover and other contamination effects at the school level (see e.g. discussion and power calculations by Schochet 2008). It is an advantage that all students at the same grade level at one school received the same amount of additional resources, as this is considered fair and thus reduces parent or teacher

⁹ The procedure consisted of the following steps: i) predicting school average test score (stratification variable) based on school and student characteristics available from register data, ii) stratifying schools into the smallest possible strata based on ranking on the stratification variable, and iii) allocating schools to treatment and control groups randomly within each strata. In municipalities where the number of schools was not divisible by three, excess schools were randomly drawn (prior to the randomization) and excluded from this analysis.

complaints that could otherwise compromise randomization, as observed in STAR.¹⁰ The disadvantage is that it reduces the statistical power of the study, since the analysis must take into account the dependencies of student responses and outcomes within clusters.

Some attrition among students was revealed at follow-up on the initial randomization. There were two main reasons for student attrition. First, less than 1.5% of the sample discontinued the intervention because they changed schools. These students are equally distributed across intervention and control schools and, hence, we do not expect student mobility to be driven by the experiment. Second, some students have missing post-measures because they did not complete the post-test, which is the case for less than 4%. Subtracting the number of students lost to follow-up from the number of students initially allocated to each experimental group resulted in the final number of students analyzed. Figure 1 shows this participation flow for each of the two main outcomes: reading and math. We return to define the specific response rates and balance in attrition across groups after the data description in Section 3.

¹⁰ In STAR, the randomization unit was the student (and teacher), resulting in schools having one classroom for each intervention. To respond to heavy parental and teacher complaints during the first year of the STAR intervention period, the schools re-randomized the students in the regular-sized classes with and without teaching aides after the first year. Furthermore, they reallocated incompatible children (6.5%) from small classes after the first year (see Mosteller 1995).

FIGURE 1. DESIGN AND PARTICIPATION FLOW

Note: Missing data on the pre-test measures are imputed with zero and, hence, do not affect the final sample-size analyzed. Reading and math scores are tracked through register data and are only missing if the student does not complete national testing, irrespective of student mobility. Students are clustered in schools and classrooms. From the allocation to analysis stage we only lose one complete cluster (classroom). All schools are represented in the final analysis sample for reading and math, respectively. The final cluster sizes range from 8–105 students per school and 3–39 students per classroom. For outcome-specific cluster size range, mean, and standard deviation see Table A.2 in Appendix A. Follow-up and analysis refer to the primary outcomes measured at the 8-month follow-up.

3. Data

This section describes outcomes, baseline data, response rates and attrition. In order to evaluate the impact of the teacher's aide intervention, we incorporate register data from Statistics Denmark on national test scores in reading and math, student ID, school identifiers, and various background characteristics about students and their parents. We are able to link the register data with survey data sampled from students, teachers, teacher's aides, and school principals through unique student IDs and school identifiers. All respondents were surveyed at the beginning and after the intervention period, typically in September 2012 and June 2013. A timeline of the implementation of the intervention and collection of surveys is shown in Table A.1 in Appendix A.

3.1. Student Achievement

Student achievement is measured using national test scores in reading and math, before and after the intervention period. The national tests are IT-based, self-scoring, adaptive tests.¹¹ Instead of giving all students the same questions and summing the number of correct answers, the software estimates an ability measure after each question answered and then finds a question with a difficulty level that matches the contemporary measure of the student's ability level. This test process continues until the student's ability is determined within a pre-specified standard error of measurement. Thus, the final ability estimate is not a function of the number of correct answers, but rather a

¹¹ For details on the national tests, see Beuchert and Nandrup (2018).

function of the difficulty level of the correctly answered questions. The final ability measures are distributed from -7 to 7 on a continuous logit scale.¹²

The test is designed to simultaneously estimate the student's ability in three cognitive areas of each subject, and the algorithm alternates question testing in each of these three cognitive areas.¹³ We standardize the ability measures in the population within year, subject, and cognitive area (mean 0, SD 1). Afterwards we calculate the average student ability score across the three cognitive areas in reading and math, respectively. Finally, these averages in reading and math are standardized in the population (mean 0, SD 1) to render regression coefficients comparable to the effect sizes of other studies.

The national tests are thought to have a pedagogical purpose rather than an accountability purpose. Thus, the main purpose of the tests is to give feedback to

¹² The psychometric model behind the adaptive algorithm is a Rasch model (Rasch 1960). The Rasch model orders the difficulty level of the test questions and the students' ability level on the same, continuous logit scale. For each subsequent test question that the student is given, the algorithm ensures that students are given questions that they have a 50% predicted probability of answering correctly and then re-estimates the student's ability level. The Rasch model may be regarded as a one-parameter version of Item Response Theory where the difficulty of the test questions (items) is the only item parameter considered; see Jacob and Rothstein (2016) who discuss modern student assessment systems in more details.

¹³ The three cognitive areas in reading are: 1) language comprehension, 2) decoding, and 3) reading comprehension, while in math they are: 1) numbers and algebra, 2) geometry, and 3) applied mathematics.

teachers, students, and parents on the individual child's ability level. In principle, the teacher can assist academically weak students or provide them with aids or breaks during tests. Unfortunately, information on assistance, aid, or other provisions made for these students is unavailable to researchers. In principle, the tests are compulsory for all students enrolled in public schools, but principals may exempt some students from the tests.

Students are tested ten times from grades two to eight, and the compulsory test is administered in the spring. The test may also be used on a voluntary basis in the autumn. Questions are drawn from a large item bank, which makes it possible for a student to complete a test, for example, in reading in sixth grade more than once. We define scores from the mandatory reading and math tests in sixth grade (spring 2013) as our primary post-test measures. These tests are characterized by very little and unsystematic attrition (see Section 3.3). The post-tests are completed about 8 months after the beginning of the intervention.¹⁴

As pre-test measures, we use test scores from reading in fourth grade (spring 2011) and math in third grade (spring 2010). At that time, the study was not yet planned and the test scores are not contaminated by the experiment. This is a great advantage

¹⁴ As additional follow-up tests, we take advantage of the above-described voluntary and mandatory test periods in the autumn and spring, respectively. We obtain follow-up tests at 1, 8, and 30 months after the beginning of the intervention. The 30-month test corresponds to the mandatory tests in eighth grade.

over other studies that only rely on a pre-ability measure obtained from a pre-test at the beginning of the intervention period.^{15,16}

3.2. Baseline Data and Balance

In this subsection, we document that the randomization was successful and that the estimation sample is balanced across treatment arms. Not surprisingly, the experimental groups balance on their predicted school average test scores, which were used for stratification.¹⁷ To further confirm that randomization was successful, we compare the distributions of predicted school average test score, because a potential concern could be whether schools from across the whole distribution are present in each of the experimental groups (i.e., that those schools allocated to specific interventions are neither positively nor negatively selected; see Figure 2). Taken at face value, Figure 2 suggests that the schools treated with a co-teacher and teaching assistant may be

¹⁵ Schochet (2010) formally discusses the pros and cons of using different pre-test scores. A related issue is potential contamination if the students at the control and intervention schools systematically differ in their test-taking patterns. Figure A.1 in Appendix A shows the distribution of test dates by test and experimental groups, which confirms that the timing of test-taking among control and treatment students is similar. Similar figures for the supplementary follow-up tests at 1 and 30 months are available on request.

¹⁶ Results on reading and math scores are robust when conditioning on students tested in May or June 2013.

¹⁷ The means and (standard deviations) are 0.13 (0.91), -0.11 (1.05), and -0.12 (1.03) for the schools assigned to the control group, the co-teacher treatment, and the teaching assistant treatment, respectively, and they are not significantly different (p-values > 0.3).

slightly negatively selected. However, two non-parametric tests, the two-sample Kolmogorov-Smirnov test and the Wilcoxon rank-sum test, both fail to reject the null hypothesis with p-values above 0.10 in all two-by-two experimental group comparisons, thus supporting the conclusion that the randomization was successful.¹⁸

FIGURE 2. DISTRIBUTION OF PREDICTED SCHOOL AVERAGE TEST SCORE

Note: Figure plots distribution of predicted school average test score by experimental groups: schools assigned to control group, co-teacher treatment, and teaching assistant treatment. Predicted school average test score is used as stratification variable. Curves are smoothed and based on 35 observations per group.

¹⁸ The Kolmogorov-Smirnov two-sample test tests the null hypothesis that the test score distributions of two groups of schools are equal. The resulting p-values are as follows: co-teacher = control (p-value 0.139); teaching assistant = control (p-value 0.595); and co-teacher = teaching assistants (p-value 0.595). The Wilcoxon rank-sum test tests the null hypothesis that the two groups of schools are random draws from a single population. The resulting p-values are as follows: co-teacher = control (p-value 0.348); teaching assistant = control (p-value 0.517); co-teacher = teaching assistant (p-value 0.652).

Since randomization was done at the school level, we may expect minor differences between the experimental groups at the student level. Table 2 shows baseline student background characteristics and tests the balance in baseline data across experimental groups. The balancing tests closely mirror the outcome estimation model. We regress each covariate on the full set of treatment indicators and randomization strata fixed effects and cluster standard errors at the school level. This means that the balancing tests take into account the stratification and clustering of the data in the exact same way as the regression models in the subsequent sections. Column 1 shows the control group's sample means and standard deviations (SD). Columns 2 and 3 show the coefficients from regressing each covariate on the two treatment indicators (i.e. reflecting the difference from mean of the control group). Column 4 reports the p-value from a joint test of the null that the coefficients to all treatment indicators equal zero.

Overall, the students in the control and treatment groups are very similar. In column 4, we reject the null (no differences across groups) for all but one out of 22 covariates (at 5% significance level) corresponding to what would be expected by chance.

TABLE 2. BASELINE DATA

	(1) Control		(2) Co-teacher		(3) Teaching assistant		(4) Test: All zero (2)=(3)=0 p-value
	Mean	SD	coef.	(s.e.)	coef.	(s.e.)	
<i>Pre-tests:</i>							
Reading 4th grade, spring 2011	-0.006	0.937	0.059	(0.046)	0.051	(0.043)	0.366
Math 3rd grade, spring 2010	-0.010	0.897	0.006	(0.045)	0.063	(0.050)	0.441
<i>Student characteristics:</i>							
Boy	0.514	0.500	-0.007	(0.014)	0.008	(0.017)	0.699
ADHD and similar diagnoses	0.030	0.171	0.000	(0.005)	0.001	(0.005)	0.949
Special education needs	0.104	0.305	0.023	(0.015)	0.052	(0.017) ***	0.010
Non-Western immigrants	0.019	0.137	-0.005	(0.004)	0.003	(0.005)	0.153
Non-Western descendants	0.107	0.309	-0.002	(0.016)	-0.010	(0.016)	0.808
Both parents, compulsory schooling only	0.105	0.307	-0.002	(0.012)	-0.019	(0.013)	0.197
Both parents, no college degree	0.593	0.491	-0.009	(0.020)	-0.030	(0.024)	0.454
<i>Mother characteristics:</i>							
Single mother	0.239	0.426	0.002	(0.014)	-0.021	(0.015)	0.204
Mother no college degree	0.619	0.486	-0.014	(0.019)	-0.018	(0.021)	0.653
Mother's age at birth (years)	29.021	6.756	0.132	(0.160)	-0.112	(0.197)	0.330
Mother's income (DKK 10,000)	24.856	17.770	0.376	(0.575)	-0.059	(0.595)	0.665
Mother employed	0.769	0.421	-0.009	(0.017)	0.002	(0.019)	0.772
Mother unemployed, insured	0.082	0.275	-0.011	(0.009)	-0.017	(0.009) *	0.169
Mother unemployed, uninsured	0.121	0.326	0.019	(0.013)	0.015	(0.016)	0.346
<i>Father characteristics:</i>							
Father no college degree	0.667	0.471	0.001	(0.019)	-0.010	(0.024)	0.861
Father's age at birth (years)	32.454	7.009	0.138	(0.239)	-0.006	(0.236)	0.767
Father's income (DKK 10,000)	33.147	28.267	1.448	(1.129)	-0.207	(1.080)	0.329
Father employed	0.795	0.404	0.003	(0.014)	0.010	(0.015)	0.795
Father unemployed, insured	0.052	0.222	0.003	(0.005)	-0.006	(0.005)	0.253
Father unemployed, uninsured	0.090	0.286	-0.003	(0.012)	0.003	(0.013)	0.840
No. students	1,789		1,814		1,610		
No. schools	35		35		35		

Note. *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$. The table reports the results from regressing each baseline covariate on the set of treatment indicators including randomization strata fixed effect and clustering at the school level. Column 1 reports the control group's sample means and standard deviations (SD), columns 2 and 3 report the coefficients (s.e.) on the treatment indicators for the co-teacher and teaching assistant treatment, respectively. Column 4 reports p-values from F-tests of the treatment indicators being jointly zero. School level is defined by the student's initial school allocation (August 2012). Baseline data are obtained from register data collected before the beginning for the intervention. Pre-test scores are standardized in the population with mean zero and a standard deviation of one within each year, grade, and subject. Missing data is imputed with the value zero. No. students=5,213. No. schools=105.

Although the randomization is successful, some concerns may be left due to small imbalances across treatment arms revealed in the final estimation samples. Tables A4 and A5 report balancing tests for the estimation samples for reading and math, and the results are almost identical to those of Table 2, because there is only little unsystematic attrition. Yet, to address the concern for pre-existing differences in test scores across treatments, we control linearly for the pre-test scores in all regressions.¹⁹

3.3. Response Rates and Attrition

Table 3 shows response rates across experimental groups for each of the reading and math tests completed pre and 1, 8, and 30 months post the intervention start. The response rates are generally very high, and the response rates on the primary test outcomes (i.e. the 8-month tests) are more than 95%.

Although the response rates are high, one concern might be that treatment effects suffer from attrition bias. As shown in Figure 1, only a few students drop out or change schools before (7 students) or during (72 students) the intervention period. Nevertheless, there may be other sources of systematic attrition due to missing pre- or post-measures. We have a unique opportunity to assess the importance of such attrition due to the availability of detailed register data for all students with a valid ID number. In Table A.3 in the Appendix, we analyze how post-test score attrition correlate with treatment assignment.

Reading and math scores are obtained from existing register data and a priori we would not expect attrition to be systematically related to the treatment. This is

¹⁹ As a robustness check, we also do propensity score matching conditioning on pre-test scores.

confirmed by Table A.3 showing no correlation between student attrition and either of the treatment indicators on the 8-month tests (primary outcomes). Table A.3 reveals that students with a non-valid ID, who drop out, repeat a grade, have much absenteeism, or have special education needs (SEN) are *more* likely to attrit, (columns 4 and 8). Importantly, interaction terms show that these do *not* systematically differ with treatment assignment (not shown).

The 1-month follow-up tests are supplementary to the mandatory national testing program. Even though the schools in the trial were required to complete these, we obtain lower response rates, and attrition analyses unfortunately reveal that response rates are correlated with treatment assignment and student background. Students assigned either of the teacher's aides are about 10 percentage points more likely to complete the 1-month supplementary tests (see Table 3),²⁰ and therefore, results based on this outcome should be interpreted with caution. Our main analyses are based on the 8-month tests, which are measured with the mandatory test program in the end of the intervention period and virtually unaffected by attrition. The 30-month tests are also part of the mandatory national test program and do not suffer from systematic attrition.²¹

²⁰ One observable variable that is significantly associated with this systematic attrition is whether the mother has a college degree or not. Students with low educated parents are an additional 3 percentage points more likely to take the 1-month test compared to students in the control group.

²¹ Students' completion of the 30-month tests are uncorrelated with treatment assignment after controlling for student pre-test score and randomization-strata. One exception, though, is that students assigned a teaching assistant are 2 percentages points more likely to complete at least one of the 30-month tests in the natural sciences. The attrition analyses are available on request.

TABLE 3. RESPONSE RATES

Outcomes	Data source	Test period	N	Response rate				Test: (2) =(3)=(4) P-value
				(1) Total	(2) Control group	(3) Co-teacher w/ degree	(4) Teaching assistant w/o degree	
Reading	National test							
	- pre-test ⁺	Feb.–April 2011	4,967	95%	94%	95%	96%	0.186
	- 1-month test	Oct.–Nov. 2012	4,859	93%	87%	96%	97%	0.046
	- 8-month test ⁺	Jan.–June 2013	5,018	96%	96%	96%	97%	0.136
	- 30-month test ⁺	Feb.–April 2015	4,601	88%	88%	88%	90%	0.300
Math	National test							
	- pre-test ⁺	Feb.–April 2010	4,494	86%	85%	87%	87%	0.765
	- 1-month test	Oct.–Nov. 2012	4,824	93%	86%	96%	96%	0.021
	- 8-month test ⁺	Jan.–June 2013	4,996	96%	96%	95%	96%	0.319
	- 30-month test ⁺ *	Feb.–April 2015	4,648	89%	88%	90%	91%	0.222

Note: Table shows response rates for national tests. Primary outcomes measuring student achievement are test scores measured in the 8-month tests. ⁺From the mandatory national test program. *Natural sciences.

An issue related to attrition is whether classroom composition is affected by treatment assignment. As mentioned in Section 2.1, students typically stay with the same classmates until grade 9, and therefore, class assignment is unlikely to change with treatment assignment. Indeed, data on students' classroom assignments pre- and post-intervention reveal that only 23 students (out of 5,018) swapped classes (within school) during the intervention period. In our analyses, we use the pre-intervention allocation meaning that we avoid endogenous change of treatment status. We estimate intention-to-treat effects, and thus reflect the effect of the treatment as well as the effect of students swapped between classrooms or moving schools. Results are robust to replicating the analysis on the ex-post sample of students.

4. Estimation Strategy

The randomization design provides our source of identification. With all students completing a pre- and a post-test, a simple evaluation would be to calculate the difference between the means of student test scores in the treatment and control groups. However, this approach does not take into account the features of the research design.

We estimate a regression model that regress the student's post-test score on two 0/1 treatment indicators (one for each of the two teacher's aide interventions: co-teacher with a teaching degree and teaching assistant without a teaching degree), the student's pre-test score, and (35-1) randomization-strata indicators. The regression equation is written as follows and estimated by ordinary least squares with clustering at the school level:

$$y_{1,isr} = \alpha + \beta_1 TC_s + \beta_2 TA_s + \partial_1 y_{0,isr} + \gamma_r + \varepsilon_{isr},$$

where y is the outcome variable for individual i , in school s , in strata r . The indicator variable TC is equal to one if the school, s , is randomly assigned to the co-teacher intervention, while TA is equal to one if the school, s , is randomly assigned to the intervention with a teaching assistant. y_0 is the outcome variable measured prior to the intervention, i.e. a measure of student i 's pre-test score. γ_r is a randomization-strata fixed effect and ε is the error term.

The effects of the interventions are identified by the random assignment of schools. The average treatment effect of being enrolled in a school with $TC_s = 1$ compared to being enrolled in a control school is measured by the parameter, β_1 . The least squares estimate of β_1 is an unbiased estimate of the average effect of a co-teacher

due to random assignment, which ensures that $E[\varepsilon_{is}|TC_s] = 0$. The same is true for the least squares estimates of β_2 .

The indicator variables TC and TA are defined based on initial treatment groups in August 2012. Therefore, the regression coefficient on each indicator variable is interpreted as the intention-to-treat effect of the respective intervention compared to students at the control schools.²² Controlling for students' pre-test scores reduces the variance in the outcome variables and lowers the number of schools required in the sample to obtain a given minimum detectable effect size. Including a strata-level fixed effect, γ_r , means that effects are identified within-randomization strata. Finally, standard errors are clustered at the school level to take into account that students are clustered in schools and that schools are the randomization unit. In Section 5.3, we check the robustness of the results in various model specifications expanding the conditioning set with baseline student and school characteristics and in alternative empirical approaches using a first-difference model and propensity score matching, respectively.

5. Results

This section presents average and heterogeneous effects of the intervention on student achievement. Furthermore, we discuss the robustness and persistence of the results.

²² Due to low student mobility (see above), the intention-to-treat is close to the treatment-on-the-treated if all schools implement the treatment exactly in accordance with the minimum requirements.

5.1. Main Results

Table 4 presents the average treatment effect on student achievement in reading and math. Since the reading and math scores are standardized with mean zero and an SD of one (in the entire student population), the treatment effects are directly interpretable as effect sizes.

Columns 1–3 show the results on reading and by gender. The effects of a co-teacher and teaching assistant intervention are significantly positive for reading. Students assigned to the intervention with a co-teacher with a teaching degree improved their reading scores by 8.6% of an SD compared to students in the control group. Students assigned to a teaching assistant with no teaching degree improved their reading scores by 13.1% of an SD. However, the effect of having a co-teacher compared to a teaching assistant is not significantly different (see bottom panel in Table 4). The point estimate on the co-teacher intervention is slightly larger for girls compared to boys, although the gender gap is not statistically significant. There is no gender difference in the point estimates on the teaching assistant. Our estimates translate to 7.6% and 11.5% of an SD per \$1,000 spending per child, which is arguably a large immediate effect.²³

²³ For individualized high-dosage tutoring, Cook et al. (2015) find a math score impact of 6% of an SD per \$1,000 spending per child. For Project STAR, Schanzenbach (2006) reports an effect of 15.2% of an SD on SAT-9 (combined reading and math) after 3rd grade, which corresponds to 0.8% of an SD per \$1,000 (total cost \$19,645 per student, 2013 dollars). Even though the immediate impact of class-size reduction appears small, it persists over time and benefits outweigh costs (e.g. Schanzenbach 2006; Chetty et al. 2011).

Columns 4–6 show the results on math. We find no significant improvements in math scores from any of the interventions, nor specifically for boys (column 5) or girls (column 6). Due to large standard errors on the estimates, however, we note that we cannot reject positive impacts on math. The high mean math score of the control group (see bottom panel) may also suggest a tendency to downward bias the estimated effect on math scores, although we do condition on pre-test scores.

The point estimates for math scores are only slightly smaller than those for reading scores and within the confidence interval of the reading score effect. However, there could be reasons to expect larger effects on reading than math. There tend to be a higher priority on language arts than math in school, which is scaled-up by the experiment. According to the survey, aides only participated in three math lessons per week compared to 4.5 lessons in language arts. A second argument may be that the math qualifications of the co-teachers and teaching assistants are relatively low. Table 1 shows that less than one third completed a mathematics study program in high school and only 27% of the co-teachers have a teaching degree in math. A third argument may be that middle school reading skills are improved through knowledge and conceptualizing, which can also be facilitated by others, e.g. parents or assistants who are not educated teachers (see discussion in Murnane, Sawhill, and Snow 2012). In contrast, math is considered more instruction-intensive and middle school math scores have proven particularly sensitive to teacher experience (see Ladd and Sorensen 2017).²⁴

²⁴ In line with this explanation, Ladd and Sorensen suggest that the skills needed to teach middle school English at a minimal level may be less than the skills needed to teach middle school math.

TABLE 4. AVERAGE EFFECT ON READING AND MATH AND BY STUDENT
GENDER

	Reading			Math		
	(1) All	(2) Boys	(3) Girls	(4) All	(5) Boys	(6) Girls
Co-teacher w/degree	0.086*** (0.030)	0.076** (0.034)	0.094** (0.038)	0.058 (0.050)	0.050 (0.062)	0.066 (0.052)
Teaching assistant w/o degree	0.131*** (0.038)	0.135*** (0.041)	0.132*** (0.049)	0.043 (0.054)	0.056 (0.062)	0.025 (0.062)
Pre-test reading (4th grade)	0.745*** (0.012)	0.725*** (0.017)	0.766*** (0.015)			
Pre-test math (3rd grade)				0.567*** (0.019)	0.579*** (0.021)	0.552*** (0.026)
Pre-test missing (0/1)	-0.604*** (0.082)	-0.715*** (0.111)	-0.465*** (0.109)	-0.148*** (0.050)	-0.197*** (0.068)	-0.095 (0.065)
<i>P-values from testing:</i>						
H0: Co-teacher = Teaching assistant	0.210	0.166	0.390	0.805	0.935	0.482
H0: Co-teacher equal across subgroups			0.648			0.760
H0: Teaching ass. equal across subgroups			0.950			0.570
Control group mean	-0.005	-0.081	0.075	0.062	0.073	0.050
Adjusted R-squared	0.602	0.597	0.606	0.350	0.354	0.341
No. students	5,018	2,568	2,450	4,996	2,562	2,434
No. schools	105	105	105	105	105	105

Note: *significant at the 10% level, **significant at the 5% level, ***significant at the 1% level. All models include a constant, randomization-strata fixed effect, and the pre-test. The pre-test is imputed with the value zero and a missing-indicator equal to one added if data on the pre-test is missing. Standard errors in parentheses are clustered at the school level.

Table 5 tests whether the results on student achievement in reading and math are heterogeneous across parental education: none of the parents hold a college degree or at least one parent holds a college degree. Columns 1–2 show the results on reading. Students of parents with no college degrees who are assigned to the teaching assistant intervention significantly improve their reading score by 18% of an SD. In comparison, the teaching assistant has no significant effect on the reading scores of students who have at least one parent with a college degree, and the gap is statistically significant (p-

value 0.001).²⁵ This implies that the teaching assistant closes almost a third of the achievement gap that initially exists between students of parents with and without a college degree (the achievement gap is -0.58 SD, see bottom of Table 5).²⁶ It is thus a relevant alternative to high-dosage tutoring (Cook et al. 2015), higher teacher quality (Chetty et al. 2014), lower class size (Fredriksson et al. 2013), or higher teacher experience (Ladd and Sorensen, 2017), all of which are known to be effective remediating policies for disadvantaged middle school children.

At the bottom of Table 5, p-values indicate that the difference between the treatment effects of co-teachers and teaching assistants on reading scores are borderline statistically significant (p-value 0.056) for students of parents without a college degree. This means that the teaching assistants without teaching degrees improve reading scores more than the educated co-teachers. We will later see that the positive impacts of the interventions actually last after the intervention period for such disadvantaged families although the gap between the two treatments disappears (effect for teaching assistants without a degree is only borderline significant, though, p-value 0.15)

²⁵ As an alternative, we split the sample in three equally sized groups by pre-test scores, see Table B.1. The results from this exercise mimic the results in Table 5. Students in the bottom third of the pre-test score distribution who are assigned a teaching assistant improve their test scores by 19% of an SD, whereas the effect on students in the top third is insignificant, and this gap is borderline statistically significant (p-value 0.053).

²⁶ This is *not* because teaching assistants are positively selected: Table 1 illustrates how high school GPA is 40% of an SD *lower* for teaching assistants compared to co-teachers with a degree (-0.42 (SD = 1.01) versus -0.02 (SD = 0.86)), and similarly for grades from compulsory school.

TABLE 5. EFFECT ON READING AND MATH BY PARENTAL EDUCATION

	Reading		Math	
	(1) No college degree	(2) At least one college degree	(3) No college degree	(4) At least one college degree
Co-teacher w/degree	0.092** (0.037)	0.073** (0.031)	0.050 (0.056)	0.065 (0.055)
Teaching assistant w/o degree	0.177*** (0.047)	0.048 (0.037)	0.043 (0.058)	0.010 (0.063)
Pre-test reading (4th grade)	0.706*** (0.015)	0.745*** (0.017)		
Pre-test math (3rd grade)			0.544*** (0.025)	0.552*** (0.023)
Pre-test missing (0/1)	-0.699*** (0.095)	-0.333** (0.133)	-0.258*** (0.061)	0.050 (0.066)
<i>P-values from testing:</i>				
H0: Co-teacher = Teaching assistant	0.056	0.491	0.908	0.424
H0: Co-teacher equal across subgroups		0.610		0.781
H0: Teaching ass. equal across subgroups		0.001		0.538
Control group mean	-0.245	0.335	-0.135	0.342
Adjusted R-squared	0.559	0.598	0.320	0.329
No. students	2,898	2,120	2,879	2,117
No. schools	105	105	105	105

Note: *significant at the 10% level, **significant at the 5% level, ***significant at the 1% level. All models include a constant, randomization-strata fixed effects, and the pre-test. The pre-test is imputed with the value zero and a missing indicator equal to one added if data on the pre-test is missing. Standard errors in parentheses are clustered at the school level.

Columns 3–4 in Table 5 show the results for math. We find that students of parents with no college degrees who are assigned to the teaching assistant tend to improve their math score more than students of parents with a college degree do. However, due to large standard errors on the estimates on math scores there are no significant or heterogeneous effects for math.²⁷

²⁷ When we split the sample in three equally sized groups by pre-test scores, we see that students in the bottom third of the pre-test score distribution who are assigned a teaching assistant improve their test scores by 13% of an SD. The estimate is significant at the 10% level, but it is neither significantly different from the small positive effect of a co-teacher or the small positive

5.2. Robustness

We have chosen to present the intention-to-treat effects while only conditioning on the students' pre-measures and the randomization-strata of the students' schools. Table 6 tests the robustness of the model specification. We present the unconditional differences in means for comparison and include student, parent and school characteristics to the conditioning set step-by-step. Consistent with previous randomization checks (see Section 2.2.), we find no substantial changes in the estimates when pre-tests are left out or when control variables are added (columns 1-4). The point estimates of the co-teacher and teaching assistant on reading are stable at around 0.086 and 0.131 SD, respectively. The point estimates on reading and math are slightly larger when data from school registers are included (column 5). As further robustness tests, we estimate the effect of the two treatments on reading and math scores from a first-difference model (column 6) and propensity score matching (column 7). Results are robust.

effect of a teaching assistant for students in the medium or top third of the distribution, see Table B.1.

TABLE 6. ROBUSTNESS OF THE AVERAGE EFFECTS, PRIMARY OUTCOMES

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	OLS	OLS	OLS	OLS	OLS	FD	PSM
Outcome: Reading							
Co-teacher w/degree	0.086*** (0.030)	0.111* (0.065)	0.095*** (0.030)	0.093*** (0.030)	0.130*** (0.030)	0.070*** (0.031)	0.088*** (0.024)
Teaching assistant w/o degree	0.131*** (0.038)	0.161** (0.068)	0.149*** (0.038)	0.148*** (0.038)	0.172*** (0.036)	0.119*** (0.042)	0.106*** (0.027)
Pre-test reading (4th grade)	0.745*** (0.012)		0.701*** (0.012)	0.671*** (0.011)	0.670*** (0.012)		
Pre-test, missing (0/1)	-0.604*** (0.082)		-0.567*** (0.078)	-0.516*** (0.076)	-0.504*** (0.077)		
No. students	5,018	5,018	5,018	5,018	5,018	5,018	4,967
No. schools	105	105	105	105	105	105	105
Outcome: Math							
Co-teacher w/degree	0.058 (0.050)	0.041 (0.078)	0.065 (0.051)	0.061 (0.050)	0.098* (0.052)	0.055 (0.053)	0.024 (0.034)
Teaching assistant w/o degree	0.043 (0.054)	0.089 (0.096)	0.059 (0.054)	0.059 (0.052)	0.072 (0.062)	0.014 (0.059)	0.008 (0.037)
Pre-test math (3rd grade)	0.567*** (0.019)		0.536*** (0.019)	0.502*** (0.018)	0.501*** (0.018)		
Pre-test, missing (0/1)	-0.148*** (0.050)		-0.141*** (0.048)	-0.110** (0.046)	-0.092** (0.046)		
No. students	4,996	4,996	4,996	4,996	4,996	4,996	4,494
No. schools	105	105	105	105	105	105	104
Includes:							
Randomization-strata fixed effects	X		X	X	X	X	
Pre-test and indicator if missing	X		X	X	X	X	X
Student register data ^a			X	X	X		
Parental register data ^a				X	X		
School register data ^b					X		

Note: *significant at the 10% level, **significant at the 5% level, ***significant at the 1% level. In columns 1-6, each model includes a constant and additional covariates as specified in the lower panel. Missing data is imputed with zero and an indicator for missing added. Standard errors in parentheses are clustered at the school level. Column 7 reports results from propensity score matching. The propensity scores condition on pre-test scores and are estimated using a probit model with 1 neighbor. Students with missing pre-test scores are left out of the PSM model. Standard errors in parentheses are estimated as Abadie and Imbens (2012) using Stata programme `teffects psmatch`. ^aStudent and parental register data includes the list of baseline characteristics from Table 2. ^bSchool register data includes number of students, classes, and special education and diagnosed students at the 6th grade cohort. OLS: ordinary least squares. FD: first-difference model. PSM: propensity score matching.

Raw plots of the post-test score distributions by experimental groups also confirms the overall results. For reading scores, the distribution is shifted to the right, while no shift is seen for the math score distribution, see Figure B.1.

5.3. Follow-up at 1, 8, and 30 Months

This section investigates effects on student test scores at the 1-, 8-, and 30-month follow-ups. Note the results presented so far are from the 8-month follow-up at the end of the intervention period. The students are required to complete supplementary sixth grade national tests in reading and math one month after the beginning of the intervention period (see Section 3.1.1). We employ the 1-month tests to capture the (intermediate) effect of the teacher's aide. The students also complete a set of mandatory national tests in reading and natural sciences at the end of grade eight, about 30 months after the intervention period began. At this time the intervention has already ended for quite some time and the 30-month tests will thus reveal whether or not the impact lasts. Throughout the follow-up period, we only include students who are in the baseline sample, and initial treatment assignments are fixed and do not change if the students change schools. Hence, new entrants in the school during the follow-up period are not included. All of the results are presented in Table B.2 in Appendix B. The point estimates at the 1-, 8-, and 30-month follow-ups are illustrated graphically in Figure 3 and overlaid with 95% confidence bands.

Figure 3 indicates an effect of the teacher's aides on student achievement already after 1 month. Since both treatment and control groups knew they were part of a trial and took part in data collections, we do not think this is the result of Hawthorne effects in the sense that they respond simply because of being observed. It may be that

the presence of a teacher's aide gives the teachers an initial motivational boost, but we see that as part of the treatment effect that should replicate outside the trial. Having said that, we also note that the response rate on the 1-month test was lower in the control group (see Table 3), so the 1-month results may be more uncertain.

FIGURE 3. EFFECT ON READING AND MATH AT 1, 8, AND 30 MONTHS

Note. Each figure plots the point estimates from three separate estimations. Point estimates are overlaid with 95% confidence bands. The 30-month tests are tests in reading and natural science (an average of scores in geography, biology, and physics/chemistry).

Figure 3 further shows that the positive effect of a co-teacher on reading is borderline significant at the 30-month follow-up and the effect of a teaching assistant seems to decrease over time after the intervention. However, this figure masks the heterogeneous effects of the education level of the students' parents that Table 5 showed.

Figures 4 and 5 illustrate the effects by subject and parental education. Figure 4 shows that both interventions seem to have lasting impacts on the reading scores of students with parents not holding college degrees (though only borderline significant for teaching assistants, p-value 0.15), and after 30 months the impacts are no longer different from each other. Figure 5 shows that none of the interventions seem to have an impact on the average natural science scores after the intervention period.

FIGURE 4. EFFECT ON READING AT 1, 8, AND 30 MONTHS

Note. Each figure plots the point estimates from three separate estimations. Point estimates are overlaid with 95% confidence bands.

FIGURE 5. EFFECT ON MATH AT 1, 8 MONTHS AND SCIENCE AT 30 MONTHS

Note. Each figure plots the point estimates from three separate estimations. Point estimates are overlaid with 95% confidence bands. The 30-month test is a test in natural sciences (an average of scores in geography, biology, and physics/chemistry).

6. Potential Channels and Mechanisms

In this section, we explore potential channels and mechanisms that may explain the impact of co-teachers and teaching assistants on reading scores, such as actual time use, aide's tasks, and team collaboration. We exploit surveys from school principals, teachers, and teacher's aides, which provide unique opportunities to monitor implementation of the intervention across schools.²⁸ It is important to note that responses are measured after the intervention started and may thus be endogenous. However, the purpose is to provide an exploratory examination of potential mechanisms

²⁸ Response rates on all surveys are more than 80%.

behind our findings. We conclude the section with a discussion of the extent to which the results may essentially be driven by class-size reduction.

6.1. Effect per Lesson

Our results reveal a positive average impact of both types of aides on reading scores confirming the hypothesis that high time intensity (i.e. number of lessons) is as important as high qualifications (i.e. teaching degrees). Below we explore this hypothesis further.

The treatment dose varies within the treatment groups. First, the actual number of lessons with an aide per week is higher than the minimum requirements for both treatments (see Section 2.2.). In Table 7 columns 1-3, we present 2SLS results instrumenting the actual number of lessons with the initial treatment allocation. In the first stage, the dependent variable is the number of lessons per week with an aide (co-teacher or teaching assistant). We regress the number of lessons with an aide on the two treatment indicators, the pre-test score, and the strata fixed effect. In the second stage, we regress students' reading scores on the predicted values obtained from the first stage, and the students' pre-test score and the strata fixed effect. The results suggest that the co-teacher improves reading scores by 0.7% of an SD per lesson per week, and the teaching assistants by 0.6% of an SD. Thus, teaching assistants appear to be almost as productive as co-teachers in terms of improving reading scores 8 months after the start of the intervention, and the difference is not significant (p-value 0.636).²⁹ This supports

²⁹ Inspired by Krueger (1999), this exercise implicitly assumes that teachers' aides do not change the education production function for the exclusion restriction to hold as well as linearity in the

findings in the existing literature that teacher quality is only weakly related to teachers' education (Hanushek and Rivkin 2006).

6.2. Minimum Number of Lessons

Secondly, not all schools uphold the minimum number of lessons per week per class.³⁰ This potentially downward biases the estimated effects and implies that the estimated intention-to-treat effect may be lower than the average treatment effect on the treated. In Table 7 columns 4-6, we present 2SLS results instrumenting the actual compliance with the initial treatment allocation.³¹ The results support a large, significant impact of co-teachers (0.104 SD) and teaching assistants (0.142 SD) after accounting for endogenous non-compliance to the lesson requirements.³²

number of lessons and production of learning. Both assumptions are debatable, and the results are only suggestive. The OLS equivalent of column 3 is reported in column 4 and the point estimates are only slightly smaller than the 2SLS estimates.

³⁰ The protocol required 10.5 or 14.5 lessons/week/class with a co-teacher or teaching assistant, respectively. According to the school principal reports, 14 co-teacher classrooms and 8 teaching assistant classrooms did not meet this requirement and three teachers respond that they never implemented the intended teacher's aide (due to sickness, leave, or lack of collaboration capability).

³¹ We measure actual compliance with an indicator equal to one if the minimum number of lessons per week per class were satisfied (10.5 or 14.5 lessons/week/class with a co-teacher or teaching assistant, respectively).

³² To address potential measurement errors in surveys, we replicate Table 7 using reports from several respondents. The correlation between measures of actual time spent in the classroom is

0.78. The estimated total effect on the treated is even greater when relying on the time spent in the classroom as reported by the aides themselves (not shown).

TABLE 7: 2SLS RESULTS ON READING: VARIATION IN TIME USE

Model	Effect per lesson				Minimum number of lessons			
	(1) 2SLS 1st stage	(2) 2SLS 1st stage	(3) 2SLS 2nd stage	(4) OLS Reading score	(5) 2SLS 1st stage	(6) 2SLS 1st stage	(7) 2SLS 2nd stage	(8) OLS Reading score
Dependent variable	No. of lessons w/ co-teacher	No. of lessons w/ teaching ass.	Reading score	Reading score	Comply with co-teacher requirement ^a	Comply with teaching ass. requirement ^a	Reading score	Reading score
<i>Initial treatment group:</i>								
Co-teacher w/ degree	11.759*** (0.435)	0.014 (0.735)			0.816*** (0.052)	0.007 (0.021)		
Teaching assistant w/o degree	0.016 (0.280)	21.344*** (1.269)			-0.015 (0.034)	0.932*** (0.034)		
<i>Interaction with actual implementation:</i>								
Co-teacher X No. of lessons per week/class			0.007*** (0.003)	0.004 (0.002)				
Teaching ass. X No. of lessons per week/class			0.006*** (0.002)	0.004** (0.001)				
Co-teacher X Comply with lesson requirement (0/1)							0.104*** (0.038)	0.061* (0.068)
Teaching ass. X Comply with lesson requirement (0/1)							0.142*** (0.042)	0.114 (0.040)
1st stage F-stat of excluded instruments	457.3	151.7			157.0	374.1		
2nd stage F-stat			185.3				179.9	
H0: Co-teacher = Teaching assistant (p-value)			0.636	0.023			0.354	0.016
Adjusted R-squared	0.929	0.856	0.600	0.606	0.841	0.933	0.601	0.605
No. students	5,018	5,018	5,018	5,018	5,018	5,018	5,018	5,018
No. schools	105	105	105	105	105	105	105	105

Note: *significant at the 10% level, **significant at the 5% level, ***significant at the 1% level. All models include a constant, randomization-strata fixed effects, and pre-test. Pre-test is imputed with the value zero and an indicator equal to one if pre-test is missing. Standard errors in parentheses are clustered at the school level. Information obtained from the principal survey collected 8 months after the beginning of the intervention. ^aDependent variable in the 1. stage is an indicator equal to one if the school complies with the minimum requirement. Protocol required 10.5 lessons/week/class with a co-teacher and 14.5 lessons/week/class with a teaching assistant.

6.3. Aides' Main Tasks

Table 1 in Section 2.2. suggests that co-teachers provide more instruction-intensive teaching, as co-teachers more often than teaching assistants undertake ordinary classroom teaching or teach a group of students inside or outside the classroom. In order to explore this as a potential channel explaining the impact of aides on reading scores, we include interaction effects between the treatment indicators and the aides' tasks in class (as described in Table 1). Table 8 shows the results of this exercise.

TABLE 8. RESULTS ON READING: INTERACTING TREATMENT WITH AIDES' TASKS IN CLASS

	Teacher survey What are the tasks of the aide?				
	(1) Ordinary classroom teaching	(2) Teaching groups outside class	(3) Teaching groups inside class	(4) Handling conflicts in class	(5) Practical tasks and preparation
Co-teacher	0.109** (0.052)	0.022 (0.043)	-0.038 (0.175)	-0.040 (0.060)	0.022 (0.056)
Co-teacher X Agree (0/1)	-0.030 (0.061)	0.089* (0.052)	0.128 (0.177)	0.148** (0.065)	0.086 (0.061)
Teaching assistant	0.149*** (0.041)	0.151** (0.064)	0.159*** (0.047)	0.187*** (0.062)	0.152*** (0.044)
Teaching ass. X Agree (0/1)	-0.151** (0.072)	-0.033 (0.071)	-0.095 (0.062)	-0.109 (0.068)	-0.123* (0.063)
No. students	5,018	5,018	5,018	5,018	5,018
No. schools	105	105	105	105	105

Note: *significant at the 10% level, **significant at the 5% level, ***significant at the 1% level. Each column is a separate regression. All models include a constant, randomization-strata fixed effects, and the pre-test, including indicator if pre-test is missing. Standard errors in parentheses are clustered at the school level. The interaction terms are equal to one if the teachers “agree” or “strongly agree” with the tasks stated in the column header. Aides' tasks are defined at the school level (which is the level of randomization) using non-missing responses from Danish teachers on the 1-month teacher survey, collected October-November 2012.

The results suggest that the positive effect of the co-teacher is driven by schools where the aide teaches smaller groups of students *outside* the classroom (main effect

small and insignificant; interaction effect 0.089 (0.052)), which essentially mimics a (flexibly implemented) class-size reduction. A similar pattern shows for schools where the co-teacher teaches smaller groups of students *inside* the classroom, but the interaction effect is not statistically significant (see columns 2-3).

Table 8 also reveals that the pattern of point estimates is strikingly different for the teaching assistant compared to the co-teacher: The main effect is stable across all specifications, and importantly, not driven by schools where the teaching assistant teaches smaller groups of students (see columns 2-3). However, the large positive main effect of the teaching assistant is counteracted by a significantly negative interaction effect when the teaching assistant is occupied by ordinary classroom teaching or practical tasks (see columns 1 and 5). These results suggest that teaching assistants do not function similar to reducing class sizes, as they take on other tasks than ordinary teaching.

6.4. Aides' Qualifications and Team Collaboration

Finally, we explore the characteristics of the team as a potential channel that may explain the impact of aides on reading scores. Table 9 includes interaction effects between the treatment indicators and indicators of the aides' experiences and other survey variables reflecting the collaboration of the teacher-aide pair. Table 9 shows that the co-teacher is more impactful when he or she has at least 5 years of teaching experience (column 2) and when the main teacher reckons that the co-teacher has the desired qualifications (column 6). On the other hand, the teaching assistant is more impactful in classrooms with behavioral problems such as noise and conflicts (column 1), and when they have talked about roles and responsibilities (column 5). Point

estimates also indicate that it may be beneficial when the teaching assistant has at least 5 years of pedagogical experience (column 3) and when the main teacher reckons that the teaching assistant has the desired qualifications (column 6), but these are not statistically significant.

TABLE 9. RESULTS ON READING: INTERACTING TREATMENT WITH AIDES' QUALIFICATIONS AND INDICATORS OF TEACHER-AIDE COLLABORATION

	Teacher assessment of class	Aides' teaching experience	Aides' pedagogical experience	Teacher assessment of teacher-aide collaboration		
	(1) Often noise and conflicts in class	(2) at least 5 years	(3) at least 5 years	(4) Not satisfied with team-work	(5) Talked about roles and responsi- bilities	(6) Aide has desired qualifi- cations
Co-teacher	0.087*** (0.030)	0.021 (0.041)	0.072 (0.043)	0.080** (0.032)	0.096*** (0.034)	0.017 (0.041)
Co-teacher X Agree (0/1)	-0.053 (0.084)	0.105** (0.053)	-0.021 (0.073)	-0.023 (0.051)	-0.062 (0.055)	0.127** (0.052)
Teaching assistant	0.129*** (0.040)	0.101*** (0.035)	0.060 (0.076)	0.118*** (0.040)	0.126*** (0.039)	0.113** (0.045)
Teaching ass. X Agree (0/1)	0.069* (0.037)	0.052 (0.098)	0.092 (0.087)	0.060 (0.097)	0.190*** (0.049)	0.074 (0.067)
No. students	5,018	4,394	3,844	4,989	4,989	4,989
No. schools	105	89	80	104	104	104

Note: *significant at the 10% level, **significant at the 5% level, ***significant at the 1% level. Each column is a separate regression. All models include a constant, randomization-strata fixed effects, and the pre-test, including indicator if pre-test is missing. Standard errors in parentheses are clustered at the school level. The interaction terms are equal to one if the teachers “agree” or “strongly agree” with the item measure stated in the column header. Measures are defined at the school level (which is the level of randomization) using non-missing responses from Danish teachers on the 1-month teacher survey, collected October-November 2012.

6.5. Do results reflect class-size reduction?

Based on the above (suggestive) analyses, we cautiously infer that the co-teacher with a degree to some extent may be regarded as a flexible sort of class-size reduction. In contrast, the teaching assistant without a degree does not seem to have an

effect through class-size reduction. Moreover, the assistant seems to be ineffective if he or she merely helps with practical tasks. However, as shown in Table 1 aides did not supply much practical tasks. This may explain why this study, in contrast to Project STAR, finds positive effects of teaching assistants. The teaching assistants in our study seem to be particularly effective in classrooms with behavioral problems and when they have talked about roles and responsibilities with the main teacher. We interpret this as an indication that complementarities explain why teaching assistants are effective at improving reading skills.

We believe our results support the notion that what we study here is not merely a flexible form of class-size reduction, but a positive way of using teaching aides that has not hitherto been documented in the literature.

7. Concluding Remarks

We study the impact of two types of teacher's aides on student achievement in a randomized trial. The experiment involved 105 schools and more than 5,200 sixth grade students. The two interventions varied in terms of the aides' teaching qualifications and time in the classroom. Overall, we find significant average effects on reading scores, which contradicts earlier studies that asserted no effects of teacher's aides.

To get some leverage regarding the importance of teacher qualifications relative to time intensity, we compare the effects of the two types of teacher's aides. Our results reveal a positive average impact of both types of aides on reading scores confirming the hypothesis that high time intensity (i.e. number of lessons) is as important as formal qualifications (i.e. teaching degrees). Furthermore, we compute the effect per lesson

spent in class on average, which does not differ significantly either. Thus, we cannot reject that the two interventions are equally cost-effective.

However, the potential mechanisms behind their effects appear to be different. The co-teacher with a degree seems to be particularly effective when used as a flexible sort of class-size reduction, whereas the teaching assistant without a degree seems to be particularly effective in classrooms with behavioral problems. Our results suggest that teacher's aides are potentially impactful, when they are part of a professional team who share responsibility for a classroom rather than help with practical tasks.

The effects on student test scores are measured eight months after the intervention period started. However, we also re-evaluate student achievement at the 30-month follow-up. The persistence of the effects drops slightly, but students with low-educated parents still perform better on average in reading tests if assigned one of the treatments (though the teaching assistant is only borderline significant at the follow-up, p-value 0.15).

Regarding the external validity, the included municipalities were selected based on geographical diversity, size of municipality and school size, and the included students' baseline test scores were very close to the population mean. We would therefore expect the results to generalize to (at least) the rest of the country. However, general equilibrium effects of experiments such as this are expected to depend on the supply and demand for teachers with different skills. Since the experiment provided only temporary financing of one year or less, the types of teaching graduates that can be hired for these jobs may be negatively selected compared to what would be the case if permanent jobs were offered. But it is also possible that scaling up would create a

shortage of highly qualified teacher's aides. While it is difficult to draw strong conclusions based on a single trial, the fact that we find similar effects of aides with and without formal qualifications may indicate that the results are not very sensitive to the teacher's formal qualifications.

This may explain why this study, contrary to Project STAR, finds positive effects of teaching assistants. The teaching assistants in our study seem to be particularly effective in classrooms with behavioral problems, and when they coordinate tasks with the main teacher. We interpret this as an indication that complementarities explain why teaching assistants are effective at improving reading skills.

In sum, the use of teacher's aides seems to be at least as efficient as class-size reductions if used as a universal instrument. However, it is a much more flexible intervention that can target specific groups of students for limited periods of time. The results of the randomized field experiment presented here provide guidance in terms of the most effective timing and type of teacher's aide.

Appendix A. Experimental Design (For online publication)

TABLE A.1. IMPLEMENTATION: TIMELINE AND DATA COLLECTION

Year	Date	Process
Before intervention period		
2012	March 16	Invite municipalities and schools to participate
2012	March 30	Application deadline
2012	May 24	Inform municipalities on participation or rejection (by letter)
2012	June 25–Aug. 31	Registration of students and teachers at schools (by principals)
2012	June 25	Registration of students with special education needs and inclusive education
2012	June–Aug.	Inform municipalities (by email) and schools (by letter) on their selection for participation in an experiment involving teacher aides
2012	Aug. 3	Collect register data for stratification and randomization
2012	Aug. 13	Randomization (sealed)
2012	Aug. 15	School year begins and students enroll in 6th grade
2012	Aug. 15	Inform schools on experimental group (co-teacher, teaching assistant, control) (by letter)
2012	Aug. 28–29	Information meetings for participating schools (in Copenhagen and Aarhus)
2012	Aug. 15–Oct. 1	Schools hire co-teachers and teaching assistants
During intervention period		
1 st half of the school year		
2012	Oct. 1	Intervention begins
2012	Oct. 22–Dec. 14	Reading and math test (1 month)
2012	Sep. 7–Oct. 17	Survey students (1 month)
2012	Oct.–Nov.	Survey principals, teachers, co-teachers, and teaching assistants (1 month)
2 nd half of the school year		
2013	Jan. 21–April 30	Reading and math test (8 months, primary outcomes)
2013	Apr. 1–Apr. 26	Lockout of teachers
2013	May 27–June 28	Reading and math test (postponed testing period due to the lockout)
2013	May	Post-registration of students, teachers, co-teachers, and teaching assistants at schools (by principals)
2013	May	Post-registration of students with special education needs and inclusive education (by principals)
2013	May–June	Survey principals, teachers, co-teachers, and teaching assistants (8 months)
2013	June 7–July 5	Survey students (8 months)
2013	June 28	School year ends
		Intervention ends
After intervention period		
2013	Aug. 12	School year begins and students enroll in 7th grade
2014	Aug. 11	School year begins and students enroll in 8th grade
2015	Jan. 19–Apr. 30	Reading and sciences test (30 months)

Note: The table illustrates the year and dates for the announcement of intervention and randomization results and collection of survey and register data before, during, and after the intervention period.

TABLE A.2. NUMBER OF CLUSTERS AND CLUSTER SIZE

	Total	Co-teacher w/degree	Teaching assistant w/o degree	Control group
Allocation stage				
School level	105 schools mean 49.6 (21.8) range 8–108	35 schools mean 51.8 (21) range 13–94	35 schools mean 46.0 (22.2) range 8–108	35 schools mean 51.1 (22.4) range 11–99
Classroom level	249 classes mean 20.9 (4.2) range 3–41	84 classes mean 21.6 (4.2) range 7–41	79 classes mean 20.4 (3.9) range 8–27	86 classes mean 20.8 (4.3) range 3–31
Analysis stage				
School level				
Reading	105 schools mean 47.8 (21.1) range 8–105	35 schools mean 49.6 (20.4) range 13–89	35 schools mean 44.7 (21.8) range 8–105	35 schools mean 49.1 (21.3) range 10–95
Math	105 schools mean 47.6 (21.1) range 8–94	35 schools mean 49.3 (20.3) range 13–89	35 schools mean 44.3 (21.7) range 8–106	35 schools mean 49.1 (21.4) range 9–94
Classroom level				
Reading	248 classes mean 20.2 (4.1) range 3–39	83 classes mean 20.9 (3.7) range 10–39	79 classes mean 19.8 (4) range 8–27	86 classes mean 20.0 (4.4) range 3–31
Math	248 classes mean 20.1 (4.1) range 3–38	83 classes mean 20.8 (3.8) range 10–38	79 classes mean 19.6 (4.1) range 8–27	86 classes mean 20.0 (4.4) range 3–31

Note: The table shows the mean (standard deviation) and range of the number of students per cluster unit. Cluster size statistics are shown for each cluster level (school and classroom level) and for each of the two main outcomes: reading and math. Allocation and analysis stage refers to the stages of the design and participation flow shown in Figure 1.

TABLE A.3. ATTRITION ANALYSIS: PRIMARY OUTCOMES (8 MONTHS TESTS)

	Post-test missing Reading				Post-test missing Math			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Co-teacher w/ degree	0.002 (0.008)	0.006 (0.006)	-0.000 (0.005)	0.001 (0.005)	0.010 (0.008)	0.014* (0.008)	0.007 (0.006)	0.009 (0.006)
Teaching assistant w/o degree	-0.012 (0.007)	-0.005 (0.006)	-0.005 (0.005)	-0.004 (0.005)	-0.002 (0.007)	0.002 (0.007)	0.003 (0.006)	0.004 (0.006)
<i>Pre-measures:</i>								
Reading, 4th grade (score)		-0.014*** (0.004)	-0.010*** (0.003)	-0.006* (0.003)				
Reading, 4th grade missing (0/1)		0.240*** (0.035)	0.115*** (0.026)	0.065*** (0.020)				
Math, 3rd grade (score)						-0.011*** (0.003)	-0.009*** (0.003)	-0.005* (0.003)
Math, 3rd grade missing (0/1)						0.095*** (0.017)	0.043*** (0.012)	0.025** (0.010)
<i>Student registration and mobility:</i>								
Non-valid student ID			0.877*** (0.028)	0.376*** (0.128)			0.945*** (0.013)	0.374*** (0.131)
Drop-out, before intervention			0.062 (0.271)	0.072 (0.273)			0.116 (0.261)	0.130 (0.260)
Drop-out, during intervention			0.280*** (0.063)	0.272*** (0.062)			0.237*** (0.059)	0.229*** (0.056)
Drop-out, to special school			0.581*** (0.213)	0.593*** (0.213)			0.610*** (0.227)	0.611*** (0.223)
Grade retention (0/1)			0.535*** (0.194)	0.568*** (0.201)			0.590*** (0.201)	0.593*** (0.205)
Absence, 5th grade (monthly avg.)			0.403*** (0.139)	0.423*** (0.139)			0.414** (0.158)	0.431*** (0.157)
<i>Register data information:</i>								
Boy				-0.002 (0.004)				-0.005 (0.005)
ADHD and similar diagnoses				0.009 (0.019)				0.013 (0.021)
Special education needs (SEN)				0.023** (0.010)				0.025** (0.011)
Non-Western immigrants				0.010 (0.013)				0.016 (0.012)
Single mother				0.004 (0.006)				0.002 (0.007)
Mother no college degree				0.000 (0.005)				-0.000 (0.005)
Mother's age at birth (years)				0.001 (0.001)				0.001 (0.001)
Mother's income (DKK 10,000)				-0.000 (0.000)				0.000 (0.000)
Mother unemployed, insured				0.010 (0.011)				0.004 (0.011)
Mother unemployed, uninsured				-0.000 (0.010)				0.014 (0.012)
Father no college degree				-0.002 (0.005)				0.007 (0.006)
Father's age at birth (years)				-0.000 (0.000)				-0.001 (0.000)
Father's income (DKK 10,000)				-0.000 (0.000)				-0.000 (0.000)
Father unemployed, insured				0.000 (0.012)				-0.004 (0.012)

Father unemployed, uninsured				0.004 (0.011)				-0.003 (0.012)
Includes:								
Constant	X	X	X	X	X	X	X	X
Randomization-strata fixed effect		X	X	X		X	X	X
Pre-test and indicator if missing		X	X	X		X	X	X
Student registration and mobility			X	X			X	X
Register data				X				X
Adjusted R-squared	0.001	0.083	0.240	0.265	0.000	0.033	0.199	0.229
No. students	5,213	5,213	5,213	5,213	5,213	5,213	5,213	5,213
No. schools	105	105	105	105	105	105	105	105

Note: *significant at the 10% level, **significant at the 5% level, ***significant at the 1% level. Each column is a separate OLS regression of the probability of missing a post-measure (0/1) on the set of covariates specified in the bottom panel. Robust standard errors in parentheses are clustered at the school level. Register data includes the list of student, mother, and father characteristics from Table 2. Missing data is imputed with value zero and an indicator included in the regression (not displayed).

TABLE A.4. BALANCING TEST: READING ESTIMATION SAMPLE BY SUBGROUPS

SAMPLE:	READING ESTIMATION SAMPLE (ALL)					BOYS		GIRLS		PARENTS: NO COLLEGE		PARENTS: AT LEAST ONE WITH COLLEGE	
	(1)		(2)	(3)	(4)	(1)	(4)	(1)	(4)	(1)	(4)	(1)	(4)
	Control		Co-teacher	Teaching assistant	Test: all zero	Control	Test: all zero	Control	Test: all zero	Control	Test: all zero	Control	Test: all zero
	Mean	SD	coef.(s.e.)	coef.(s.e.)	p-value	Mean	p-value	Mean	p-value	Mean	p-value	Mean	p-value
<i>Pre-tests:</i>													
Reading 4th grade, spring 2011	0.008	0.937	0.063(0.045)	0.049(0.044)	0.346	-0.059	0.445	0.084	0.237	-0.180	0.915	0.273	0.175
Math 3rd grade, spring 2010	-0.003	0.899	0.009(0.045)	0.059(0.049)	0.482	0.024	0.852	-0.032	0.069	-0.153	0.619	0.209	0.717
<i>Student characteristics:</i>													
Boy	0.517	0.500	-0.012(0.014)	0.004(0.017)	0.601					0.523	0.859	0.507	0.381
ADHD and similar diagnoses	0.030	0.170	0.001(0.005)	0.000(0.005)	0.993	0.043	0.683	0.016	0.336	0.030	0.897	0.030	0.736
Special education needs	0.100	0.300	0.022(0.014)	0.051(0.017)***	0.014	0.117	0.011	0.079	0.256	0.125	0.005	0.063	0.320
Non-Western immigrants	0.016	0.127	-0.002(0.004)	0.004(0.004)	0.317	0.012	0.415	0.021	0.069	0.023	0.341	0.007	0.334
Non-Western descendants	0.103	0.303	0.003(0.015)	-0.004(0.016)	0.895	0.092	0.397	0.114	0.645	0.133	0.450	0.059	0.082
Both parents, compulsory schooling only	0.093	0.291	0.004(0.012)	-0.012(0.012)	0.309	0.088	0.433	0.093	0.299	0.159	0.418		
Both parents, no college degree	0.586	0.493	-0.010(0.020)	-0.028(0.025)	0.534	0.594	0.419	0.576	0.778				
<i>Mother characteristics:</i>													
Single mother	0.227	0.419	0.008(0.015)	-0.016(0.015)	0.201	0.228	0.141	0.222	0.722	0.273	0.663	0.162	0.063
Mother no college degree	0.623	0.485	-0.016(0.020)	-0.016(0.023)	0.684	0.623	0.806	0.625	0.636			0.166	0.063
Mother's age at birth (years)	29.332	6.115	0.128(0.168)	-0.131(0.205)	0.324	29.301	0.160	29.543	0.730	28.282	0.438	30.820	0.299
Mother's income (DKK 10,000)	25.483	17.599	0.109(0.594)	-0.348(0.618)	0.681	25.820	0.440	25.275	0.931	20.642	0.741	32.348	0.117
Mother employed	0.787	0.409	-0.017(0.017)	-0.008(0.019)	0.577	0.805	0.207	0.773	0.894	0.715	0.409	0.890	0.668
Mother unemployed, insured	0.080	0.271	-0.010(0.009)	-0.013(0.009)	0.329	0.072	0.685	0.088	0.347	0.101	0.505	0.049	0.595
Mother unemployed, uninsured	0.115	0.319	0.027(0.013)**	0.021(0.015)	0.128	0.104	0.049	0.127	0.218	0.156	0.110	0.056	0.219
<i>Father characteristics:</i>													
Father no college degree	0.674	0.469	0.000(0.019)	-0.012(0.025)	0.835	0.702	0.590	0.648	0.827			0.373	0.816
Father's age at birth (years)	32.741	6.297	0.199(0.245)	0.000(0.242)	0.607	32.620	0.323	33.070	0.984	32.176	0.275	33.543	0.705
Father's income (DKK 10,000)	34.006	28.187	1.247(1.176)	-0.531(1.119)	0.355	33.186	0.647	35.094	0.190	26.510	0.457	44.638	0.339
Father employed	0.811	0.392	-0.001(0.014)	-0.001(0.014)	0.994	0.812	0.874	0.815	0.838	0.745	0.917	0.904	0.779
Father unemployed, insured	0.050	0.218	0.004(0.005)	-0.003(0.005)	0.508	0.057	0.172	0.042	0.907	0.065	0.872	0.030	0.812
Father unemployed, uninsured	0.087	0.282	-0.003(0.012)	0.005(0.014)	0.779	0.087	0.897	0.088	0.791	0.112	0.769	0.052	0.399
No. students					5,018		2,568		2,433		2,898		2,120
No. schools					105		105		105		105		105

Note. See Table 2. *significant at the 10% level, **significant at the 5% level, ***significant at the 1% level. Means and p-values in bold are significant at the 5% level.

TABLE A.5. BALANCING TEST: MATH ESTIMATION SAMPLE BY SUBGROUPS

SAMPLE:	MATH ESTIMATION SAMPLE (ALL)					BOYS		GIRLS		PARENTS: NO COLLEGE		PARENTS: AT LEAST ONE WITH COLLEGE	
	(1)	(2)	(3)	(4)	Test: all zero p-value	(1)	(4)	(1)	(4)	(1)	(4)	(1)	(4)
	<u>Control</u>	<u>Co-teacher</u>	<u>Teaching assistant</u>			<u>Control</u>	Test: all zero	<u>Control</u>	Test: all zero	<u>Control</u>	Test: all zero	<u>Control</u>	Test: all zero
	Mean	SD	coef.(s.e.)	coef.(s.e.)		Mean	p-value	Mean	p-value	Mean	p-value	Mean	p-value
<i>Pre-tests:</i>													
Reading 4th grade, spring 2011	0.006	0.936	0.069(0.046)	0.060(0.044)	0.259	-0.062	0.346	0.083	0.165	-0.184	0.847	0.274	0.169
Math 3rd grade, spring 2010	-0.001	0.903	0.007(0.044)	0.067(0.048)	0.373	0.020	0.699	-0.022	0.058	-0.154	0.533	0.216	0.674
<i>Student characteristics:</i>													
Boy	0.516	0.500	-0.008(0.014)	0.008(0.017)	0.662					0.524	0.921	0.504	0.434
ADHD and similar diagnoses	0.030	0.170	0.001(0.005)	0.000(0.005)	0.984	0.044	0.547	0.014	0.216	0.030	0.843	0.029	0.710
Special education needs	0.102	0.302	0.019(0.015)	0.047(0.017)***	0.025	0.122	0.025	0.079	0.309	0.128	0.009	0.065	0.352
Non-Western immigrants	0.016	0.127	-0.002(0.003)	0.003(0.004)	0.280	0.012	0.488	0.021	0.012	0.023	0.211	0.007	0.464
Non-Western descendants	0.104	0.305	0.002(0.016)	-0.005(0.016)	0.887	0.094	0.526	0.116	0.629	0.136	0.434	0.059	0.067
Both parents, compulsory schooling only	0.094	0.292	0.003(0.012)	-0.014(0.012)	0.219	0.089	0.455	0.094	0.215	0.161	0.303		
Both parents, no college degree	0.586	0.493	-0.010(0.020)	-0.030(0.025)	0.481	0.595	0.375	0.574	0.755				
<i>Mother characteristics:</i>													
Single mother	0.228	0.420	0.007(0.016)	-0.018(0.015)	0.189	0.231	0.208	0.221	0.599	0.273	0.544	0.166	0.141
Mother no college degree	0.623	0.485	-0.015(0.020)	-0.018(0.022)	0.657	0.626	0.748	0.624	0.651			0.169	0.063
Mother's age at birth (years)	29.325	6.128	0.107(0.165)	-0.114(0.199)	0.426	29.299	0.159	29.531	0.832	28.281	0.567	30.803	0.311
Mother's income (DKK 10,000)	25.415	17.559	0.116(0.599)	-0.280(0.618)	0.758	25.770	0.455	25.187	0.958	20.591	0.732	32.243	0.149
Mother employed	0.787	0.410	-0.019(0.017)	-0.006(0.019)	0.500	0.805	0.243	0.772	0.870	0.716	0.277	0.886	0.826
Mother unemployed, insured	0.081	0.273	-0.008(0.009)	-0.015(0.009)*	0.241	0.072	0.626	0.091	0.256	0.102	0.409	0.051	0.532
Mother unemployed, uninsured	0.115	0.319	0.026(0.014)*	0.022(0.016)	0.161	0.104	0.126	0.127	0.214	0.154	0.118	0.059	0.315
<i>Father characteristics:</i>													
Father no college degree	0.673	0.469	-0.002(0.019)	-0.014(0.025)	0.804	0.700	0.555	0.649	0.902			0.374	0.789
Father's age at birth (years)	32.756	6.310	0.150(0.236)	-0.023(0.235)	0.696	32.652	0.439	33.066	0.992	32.202	0.331	33.541	0.743
Father's income (DKK 10,000)	33.764	28.100	1.519(1.183)	-0.332(1.108)	0.307	32.930	0.378	34.861	0.238	26.269	0.414	44.375	0.274
Father employed	0.808	0.394	-0.001(0.014)	0.001(0.014)	0.991	0.811	0.877	0.809	0.952	0.740	0.955	0.903	0.745
Father unemployed, insured	0.051	0.219	0.004(0.005)	-0.004(0.005)	0.312	0.057	0.143	0.043	0.997	0.064	0.658	0.031	0.898
Father unemployed, uninsured	0.089	0.285	-0.004(0.012)	0.004(0.013)	0.798	0.087	0.881	0.092	0.866	0.116	0.826	0.051	0.415
No. students					4,996		2,562		2,418		2,879		2,117
No. schools					105		105		105		105		105

Note. See Table 2. *significant at the 10% level, **significant at the 5% level, ***significant at the 1% level. Means and p-values in bold are significant at the 5% level.

FIGURE A.1. DISTRIBUTION OF NATIONAL TEST DATES BY EXPERIMENTAL GROUP

Note: Figure plots distribution of test dates. Top: post-test reading (left) and math (right), 8 months after the beginning of the intervention. Bottom: pre-test reading (left) and math (right), 2 and 3 years earlier, respectively. Distributions shown by experimental groups: schools assigned to control group, co-teacher treatment, and teaching assistant treatment.

Appendix B. Additional Results (For online publication)

TABLE B.1: HETEROGENOUS EFFECTS BY STUDENTS' PRE-TEST SCORES

	Reading			Math		
	(1) Bottom	(2) Middle	(3) Top	(4) Bottom	(5) Middle	(6) Top
Co-teacher w/degree	0.101** (0.046)	0.083** (0.033)	0.056 (0.038)	0.037 (0.063)	0.069 (0.057)	0.043 (0.073)
Teaching assistant w/o degree	0.188*** (0.051)	0.121*** (0.039)	0.055 (0.047)	0.127* (0.067)	0.033 (0.066)	0.014 (0.072)
<i>P-values from testing</i>						
H0: Co-teacher = TA	0.075	0.326	0.977	0.218	0.589	0.756
H0: Co-teacher equal across subgroups			0.649			0.789
H0: TA equal across subgroups			0.053			0.287
Control group mean	-0.785	0.035	0.860	-0.517	0.076	0.736
Adjusted R-squared	0.300	0.130	0.306	0.110	0.108	0.195
No. students	1,594	1,616	1,628	1,433	1,456	1,476

Note: *significant at the 10% level, **significant at the 5% level, ***significant at the 1% level. All models include a constant, randomization-strata fixed effect, and the pre-test. The pre-test is imputed with the value zero and a missing-indicator equal to one added if data on the pre-test is missing. Standard errors in parentheses are clustered at the school level. Students' pre-test scores are divided into three equally sized groups: bottom, middle, and top third of the pre-test score distribution, respectively.

TABLE B.2. EFFECT ON TEST SCORES BY TIME AFTER INTERVENTION START

	(1)	(2)	(3)	(4)	(5)
				Parental education	
	All	Boys	Girls	No college degree	At least one with college degree
<i>1 MONTH (GRADE 6)</i>					
Outcome: Reading					
Co-teacher w/degree	0.044* (0.024)	0.026 (0.031)	0.056* (0.030)	0.051 (0.031)	0.035 (0.031)
Teaching assistant w/o degree	0.077*** (0.027)	0.097*** (0.037)	0.061** (0.030)	0.120*** (0.035)	0.007 (0.033)
No. students	4,859	2,487	2,372	2,793	2,066
No. schools	104	103	104	104	103
Outcome: Math					
Co-teacher w/degree	0.084** (0.042)	0.076 (0.049)	0.094* (0.048)	0.087* (0.049)	0.078* (0.045)
Teaching assistant w/o degree	0.056 (0.046)	0.065 (0.052)	0.050 (0.053)	0.055 (0.054)	0.034 (0.046)
No. students	4,824	2,466	2,358	2,789	2,035
No. schools	104	103	104	103	104
<i>8 MONTHS (GRADE 6)</i>					
Outcome: Reading					
Co-teacher w/degree	0.086*** (0.030)	0.076** (0.034)	0.094** (0.038)	0.092** (0.037)	0.073** (0.031)
Teaching assistant w/o degree	0.131*** (0.038)	0.135*** (0.041)	0.132*** (0.049)	0.177*** (0.047)	0.048 (0.037)
No. students	5,018	2,568	2,450	2,898	2,120
No. schools	105	105	105	105	105
Outcome: Math					
Co-teacher w/degree	0.058 (0.050)	0.050 (0.062)	0.066 (0.052)	0.050 (0.056)	0.065 (0.055)
Teaching assistant w/o degree	0.043 (0.054)	0.056 (0.062)	0.025 (0.062)	0.043 (0.058)	0.010 (0.063)
No. students	4,996	2,562	2,434	2,879	2,117
No. schools	105	105	105	105	105
<i>30 MONTHS (GRADE 8)</i>					
Outcome: Reading					
Co-teacher w/degree	0.051 (0.031)	0.080* (0.045)	0.013 (0.036)	0.068** (0.030)	0.018 (0.041)
Teaching assistant w/o degree	0.029 (0.037)	0.025 (0.053)	0.032 (0.037)	0.055 (0.038)	-0.026 (0.046)
No. students	4,601	2,372	2,229	2,667	1,934
No. schools	105	105	105	105	105
Outcome: Math ^a					
Co-teacher w/degree	0.025 (0.038)	0.021 (0.049)	0.025 (0.043)	0.021 (0.036)	0.015 (0.053)
Teaching assistant w/o degree	-0.007 (0.042)	-0.023 (0.056)	0.004 (0.044)	-0.006 (0.044)	-0.018 (0.051)
No. students	4,648	2,386	2,259	2,706	1,942
No. schools	105	105	105	105	105

Note: *significant at the 10% level, **significant at the 5% level, ***significant at the 1% level. All models include a constant, randomization-strata fixed effects, and the pre-test. Standard errors in parentheses are clustered at the school level. The pre-test is imputed with the value zero and a missing-indicator equal to one added, if data on the pre-test is missing. The 1-month test scores are standardized with the population parameters from the mandatory grade 6 national tests. ^aThe 30-months test is a test in natural science (an average of scores in geography, biology, and physics/chemistry).

FIGURE B.1. POST-TEST SCORE DISTRIBUTION BY EXPERIMENTAL GROUP

Note: Figures plot distribution of test scores. Left (right): reading (math), 8 months after the beginning of the intervention. Distributions shown by experimental groups: schools assigned to control group, co-teacher treatment, and teaching assistant treatment.

References

Aaronson, Daniel, Lisa Barrow and William Sander (2007). “Teachers and Student Achievement in the Chicago Public High Schools.” *Journal of Labor Economics*, 24, 95–135.

Andersen, Simon C., Louise V. Beuchert, Helena S. Nielsen and Mette K. Thomsen (2014). *2L Rapport – Undersøgelse af effekten af tolærerordninger* [Evaluating the effect of co-teaching]. TrygFonden’s Centre for Child Research, Aarhus University.

Angrist, Joshua D. and Victor Lavy (1999). “Using Maimonides’ Rule to Estimate the Effect of Class Size on Scholastic Achievement.” *Quarterly Journal of Economics*, 114, 533–75.

Beuchert, Louise V. and Anne B. Nandrup (2018). “The Danish National Tests at a Glance.” *Nationaløkonomisk Tidsskrift* [Open access journal of the Danish Economic Society], 1(2), 1-37.

Blatchford, Peter, Anthony Russell and Rob Webster (2012). *Reassessing the Impact of Teaching Assistants: How Research Challenges Practice and Policy*. Oxon: Routledge.

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(6), 4–16.

Browning, Martin and Eskil Heinesen (2007). “Class Size, Teacher Hours and Educational Attainment.” *Scandinavian Journal of Economics*, 109, 415–38.

Chetty, Raj, John N. Freidman, Nathaniel Hilger, Emmanuel Saez, Diane W. Schanzenbach and Danny Yagan (2011). “How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR.” *Quarterly Journal of Economics*, 126, 1593–60.

Chetty, Raj, John N. Friedman and Jonah E. Rockoff (2014). “Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood.” *American Economic Review*, 104, 2633–79.

Cook, Bryan G., Kimberly A. McDuffie-Landrum, Linda Oshita, and Sara C. Cook (2011). “Co-teaching for Students with Disabilities: A Critical Analysis of the Empirical Literature.” In James M. Kauffman and Daniel Hallahan (eds.) *Handbook of Special Education*, pp. 147–59. New York: Routledge.

Cook, Philip J., Kenneth Dodge, George Farkas, Roland G. Fryer Jr., Jonathan Guryan, Jens Ludwig, Susan Mayer, Harold Pollack, and Laurence Steinberg (2015). “Not Too Late: Improving Academic Outcomes for Disadvantaged Youth.” IPR WP-15-01, Northwestern University.

Fredriksson, Peter, Björn Öckert and Hessel Oosterbeek (2013). “Long-Term Effects of Class Size.” *Quarterly Journal of Economics*, 128, 249–85.

Friend, Marilyn P. and Lynne Cook (2016). *Interactions: Collaboration Skills for School Professionals*. Pearson, 8th ed.

Fryer, Roland G. Jr. (2014). “Injecting Charter School Best Practices into Traditional Public Schools: Evidence from Field Experiments.” *Quarterly Journal of Economics*, 129, 1355–407.

Gerber, Susan B., Jeremy D. Finn, Charles M. Achilles and Jayne Boyd-Zacharias (2001). “Teacher Aides and Students’ Academic Achievement.” *Educational Evaluation and Policy Analysis Summer*, 23, 123–43.

Hanushek, Eric A. and Steven G. Rivkin (2006). “Teacher Quality.” In Eric A. Hanushek and Finis Welch (eds.) *Handbook the Economics of Education*, vol. 2, pp. 1052-1078. Amsterdam: North Holland.

Jacob, Brian and Jesse Rothstein (2016). “The Measurement of Student Ability in Modern Assessment Systems.” *Journal of Economic Perspectives*, 30(3), 85-107.

Krueger, Alan B. (1999). “Experimental Estimates of Education Production Functions.” *Quarterly Journal of Economics*, 114, 497–532.

Ladd, Helen F. and Lucy C. Sorensen (2017). “Returns to Teacher Experience: Student Achievement and Motivation in Middle School.” *Education Finance and Policy*, 12, 241-279.

Leuven, Edwin, Mikael Lindahl, Hessel Oosterbeek and Dinand Webbink (2007). “The Effect of Extra Funding for Disadvantaged Pupils on Achievement.” *Review of Economics and Statistics*, 89, 721–36.

Leuven, Edwin, Hessel Oosterbeek and Marte Rønning (2008). “Quasi-experimental Estimates of the Effect of Class Size on Achievement in Norway.” *Scandinavian Journal of Economics*, 110, 663–93.

Mosteller, Frank (1995). “The Tennessee Study of Class Size in the Early School Grades.” *Future of Children*, 5(2), 113–27.

Murawski, Wendy W. and H. Lee Swanson (2001). “A Meta-analysis of Co-teaching Research: Where are the Data?” *Remedial and Special Education*, 22, 258–67.

Murnane, Richard, Isabel Sawhill and Cathrine Snow (2012). “Literacy Challenges for the Twenty-first Century: Introducing the Issue.” *Future of Children*, 22(2), 3–15.

Nandrup, Anne B. (2016). “Do Class Size Effects Differ across Grades?” *Education Economics*, 24, 83–95.

OECD (2014). *TALIS 2013 Results: An International Perspective on Teaching and Learning*, TALIS, OECD Publishing.

OECD (2016). *Education at a Glance 2016*, OECD indicators. OECD Publishing.

Rasch, Georg (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Pædagogiske Institut.

Schanzenbach, Diane W. (2006). “What Have Researchers Learned from Project STAR?” *Brookings Papers on Education Policy*, 9, 205–28.

Schochet, Peter Z. (2008). “Statistical Power for Random Assignment Evaluations of Education Programs.” *Journal of Educational and Behavioral Statistics*, 33, 62–87.

Schochet, Peter Z. (2010). “The Late Pretest Problem in Randomized Control Trials of Education Interventions.” *Journal of Educational and Behavioral Statistics*, 35, 379–406.

Spybrook, Jessica, Howard Bloom, Richard Congdon, Carolyn Hill, Andres Martinez and Stephen W. Raudenbush (2011). *Optimal Design Plus Empirical Evidence: Documentation for the “Optimal Design” Software*. Kalamazoo: Western Michigan University.

UNI-C (2012). *Analyse af Nystartede Elever og Omgængere i Grundskolens Børnehaveklasse* [Analysis of Newly Enrolled and Retained Pre-School Pupils]. UNI-C.