I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12543

Instruction Time, Information, and Student Achievement: Evidence from a Field Experiment

Simon Calmar Andersen
Thorbjørn Sejr Guul
Maria Knoth Humlum

AUGUST 2019

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12543

Instruction Time, Information, and Student Achievement: Evidence from a Field Experiment

Simon Calmar Andersen
Aarhus University

Thorbjørn Sejr Guul
Aarhus University

Maria Knoth Humlum
Aarhus University and IZA

AUGUST 2019

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Instruction Time, Information, and Student Achievement: Evidence from a Field Experiment*

Prior research has shown that time spent in school does not close the achievement gap between students with low and high socioeconomic status (SES). We examine the effect of combining increased instruction time with information to teachers about their students' reading achievements by using a randomized controlled trial. We find that the teachers' baseline beliefs are more important for low-SES students' academic performance, that the intervention makes the teachers update these beliefs, and—not least—that the intervention improves the reading skills of low-SES students and thereby reduces the achievement gap between high- and low-SES students. The results are consistent with a model in which the teachers' beliefs about the students' reading skills are more important to low- than high-SES students, while at the same time, the teachers' beliefs are subject to information friction and Bayesian learning.

JEL Classification: I24, I28, D83
Keywords: information, learning, field experiment

Corresponding author:
Maria Knoth Humlum
Department of Economics and Business Economics
Aarhus University
Fuglesangs Allé 4
8210 Aarhus V
Denmark
E-mail: mhumlum@econ.au.dk

* We thank Michael Rosholm for supporting the setup of the project and Rambøll for data collection assistance and support. We thank seminar participants at Aarhus University and the National Centre for School Research along with participants at the IWAEE 2018, the SDU Applied Microeconomics workshop 2018, and the briq/IZA Workshop on Behavioral Economics of Education 2019. Especially, we want to thank James Heckman and Laura Justice for valuable comments to an earlier version of the paper. The implementation and evaluation of the randomized experiment was funded by the Danish Ministry of Education. The views expressed in the paper are those of the authors and do not necessarily reflect the views of the Danish Ministry of Education.

1 Introduction

If the difference in learning opportunities for students with low and high socioeconomic status (SES) is larger outside of school than within school, we expect that, as students spend more time in school, gaps in educational achievements would diminish. However, according to existing evidence, these achievement gaps do not close. If anything, the achievement gap between children of high- and low-income families widens from age 6 to age 12 in the U.S. (Carneiro and Heckman, 2003) and remains constant from grade 2 to grade 8 in Denmark (Nandrup and Beuchert-Pedersen, 2018; de Montgomery and Sievertsen, 2019). Even when comparing the achievement gaps in U.S. birth cohorts between 1954 and 2001, the SES achievement gaps are remarkably stable (Hanushek et al., 2019). Consistently, differential impacts of instruction time for high- and low-SES students were found in a study of a large German education reform that increased weekly instruction time (Huebener et al., 2017). An important question therefore becomes how time in school can create more equal opportunities for all students.

We study this question by using a randomized controlled trial, which we have designed in collaboration with the Ministry of Education in Denmark as a follow-up to an instruction time trial (Andersen et al., 2016). The new trial presented in this paper combines increased instruction time (a teaching program of two lessons per week in 16 weeks in fifth grade) with information to the teachers about their students' performance in monthly reading tests. Existing research shows that instruction is more effective if it is adapted to the students' current skill levels (Banerjee et al., 2007, 2017; Duflo et al., 2011). However, if the teachers have inaccurate beliefs about their students' skill levels, they cannot adapt their teaching adequately. Therefore, providing teachers with information about their students' progress may help them target their instruction to the students' skill levels. If low-SES students have poorer learning opportunities outside of school—if their parents are less able to compensate for inadequate instruction—accurate teacher beliefs may be more important for low- than high-SES students.

It is not clear from existing evidence whether it is correct that accurate teacher beliefs are more important for low-SES students. Some indications support this idea, though. A new study shows that conditioning on students' skills, teachers expect that black students are less likely to graduate from college than white students, and these lower teacher expectations causally reduce the students' chances of graduation—probably partly because of reduced educational achievement in school (Papageorge et al., 2018). It has also been demonstrated that schools with high shares of low-SES students are less willing to participate in a nation-wide student testing program, even though these students benefit the most from being tested and having their test results provided to their teachers (Andersen and Nielsen, 2016). These results from prior research suggest that providing teachers with updated information about their students' skills will reduce inaccurate beliefs and that such information in combination with more instruction time in school can improve the learning of especially low-SES students and thereby reduce the achievement gap. However, little if any research has examined this idea. A review by Damgaard and Nielsen (2018) finds twenty-seven studies of information interventions (such as providing information about student behavior, attendance of skills, school quality, or rates of return to education). All of these interventions are targeted at either students themselves or their parents—none of them are targeted at the teachers.

We show that the combined information and instruction time intervention improved the reading achievements of low-SES students at the end of the intervention period. The analysis of the potential mechanisms first shows that the intervention improved accucacy in teacher beliefs—particularly for low-SES students and that in the control group, inaccurate beliefs were more negatively correlated with reading achievements among low-SES than high-SES students. This result is consistent with the notion that high-SES students are less dependent on teacher beliefs. We also find that the effect of the intervention on reading achievements is larger for the low-SES students with a teacher belief test score gap at baseline. Furthermore, we find that the intervention reduces behavioral problems among the low-SES students. This indicates that the monthly tests do not stress the students. Finally, we examine which factors

explain the teachers' inaccurate beliefs and find that more experienced teachers have more accurate beliefs.

Our study contributes to a growing literature on information friction and educational investments. Accumulated evidence demonstrates that providing parents with information about their child's school performance reduces the gap between the test scores and the parents' beliefs in their child's skills and that this information makes the parents adjust their investment decisions—especially when it comes to low-SES parents (Dizon-Ross, 2019, see also Barrera-Osorio et al., 2018; Bergman and Chan, 2017; Bergman, 2015; Bergman and Rogers, 2017; Rogers and Feller, 2018). More generally, biased beliefs in the educational production function affect the parents' decisions about time invested in their child (Cunha et al., 2013). Older students at the university level similarly gain from receiving information on their past exam performance (Bandiera et al., 2015). The teachers in our study had access to the baseline test results. Nevertheless, their beliefs in the students' reading achievement diverge from an objective performance measure at baseline. This suggests that not only low-SES students' parents but also their teachers are prone to information friction.

We also contribute to the literature on performance information in schools. Taylor and Tyler (2012) find that providing teachers with feedback on their performance in class increases teacher productivity in the subsequent years—especially for teachers who performed poorly before being evaluated. Relatedly, Rockoff et al. (2012) randomly assign school principals to objective estimates of teacher performance. They find that the principals' perception of their teachers' performance becomes more accurate (relative to objective performance estimates) when the principals receive systematic performance information. They also find that the performance information exerts greater influence on the principals' perception after the intervention than when the principals have less precise beliefs prior to the intervention. The results are coherent with a Bayesian learning model in which the principals base their beliefs on prior as well as new information. Our results are coherent with a similar learning model: As teachers receive systematic information about their students, they update their

beliefs, which makes it possible to adapt the teaching to the students' different levels. This seems to be of particular importance to low-SES students, who may be more dependent on adequate instruction in school. High-stakes accountability performance information systems have been found to induce gaming and cheating (e.g., Jacob and Levitt, 2003, see review in Figlio and Loeb, 2011). The present study indicates that the key components of these systems, i.e., the testing of students and information provided to the teachers, improve the teachers' perception in a setup without any external accountability. In our study, no one else but the teachers had access to the test results. This corresponds to Andersen and Nielsen's (2016) findings that testing in itself may improve student learning in a low-stakes system.

Finally, the study contributes to the strand of education research on teacher quality. Many studies have emphasized and documented the importance of teacher quality for student achievement (e.g., Rivkin et al., 2005), and teacher experience has been shown to have a beneficial effect on the students' academic performance (Gerritsen et al., 2017; Staiger and Rockoff, 2010). Aside from this, the specific aspects of teacher quality or the specific characteristics of teachers who have an impact on student achievement are not well known (Hanushek, 2011). We demonstrate that older and thereby more experienced teachers have more accurate beliefs in their students' reading skills. These results suggest that the ability to correctly perceive the students' skill levels in the classroom may be one important component of what has often been termed "teacher quality", which has been a somewhat black box concept, not distinguished by much more than teacher experience.

In section 2, we briefly describe the institutional setting and the design of the field experiment. Section 3 introduces the data and the measures used in the empirical analysis. In section 4, the empirical strategy is described, and the results are presented. The section also presents analyses of the mechanisms as well as additional analyses demonstrating the robustness of the results. Finally, section 5 concludes.

2 Background and Experimental Design

2.1 Background

When the present study was conducted in 2014, the level of instruction time in Denmark was close to the OECD average. At the same time, Denmark was among the OECD countries that invested most in primary education. Yet, the correlation between the parents' socioeconomic status and the students' academic skills was stronger than in most other OECD countries (OECD, 2014, pp. 200, 207, 428). As a response to an OECD report arguing that the "evaluation culture" in public schools in Denmark was poor (OECD, 2004), a majority in the parliament decided in 2010 that all public schools should use ten national, standardized tests of student skills in different subjects. Most tests are in reading. The ten mandatory tests are spread out between second and eighth grade, and reading is tested in second, fourth, sixth, and eighth grades (for an extensive overview of the Danish national tests, see Nandrup and Beuchert-Pedersen, 2018). Besides the mandatory tests, the teachers can decide to use different tests of their choosing, and the municipalities can mandate specific tests. The municipalities govern schools in Denmark (comparable to school districts in the U.S.). Despite these efforts to improve the evaluation culture, regular testing continues to be much less prevalent in Denmark than in many other countries (OECD, 2016).

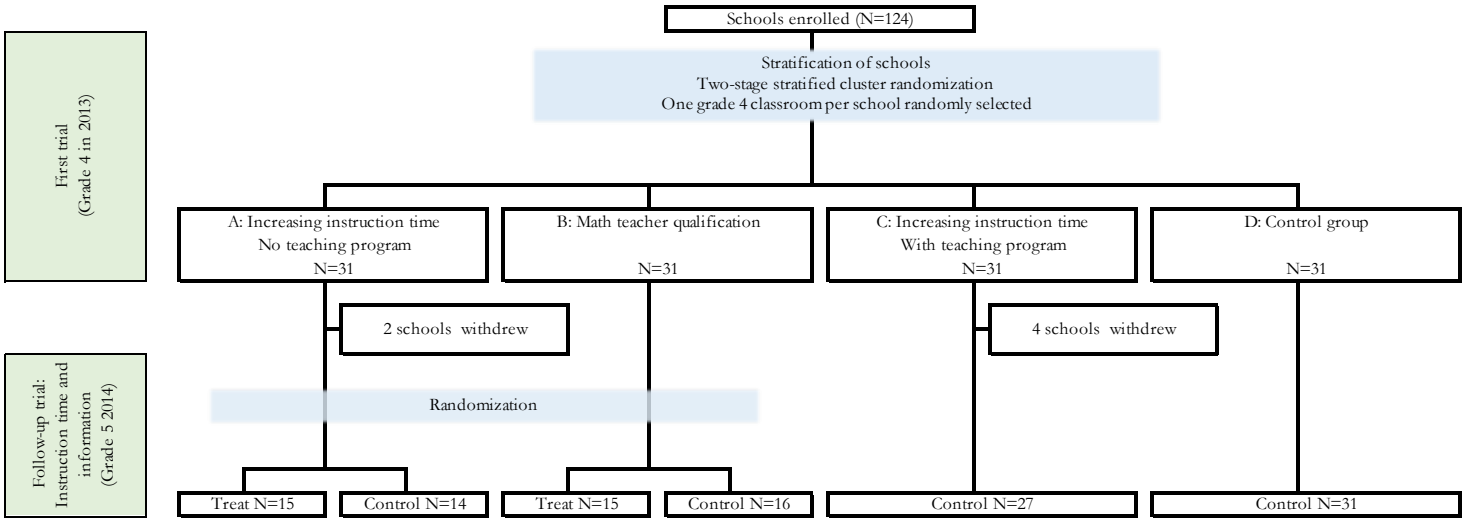
The public schools in Denmark enroll students with very different socioeconomic status. About 86 percent of all students attend public schools. The remaining 14 percent almost all attend private schools that are highly subsidized by the government. The private schools receive a voucher of about 75 percent of the average costs of a public school student. Even though public school students have slightly lower SES than private school students on average, the public school system caters to students of all levels of society (Ministry of Education, 2008; Andersen, 2008).

2.2 First Trial

In the fall of 2013, the Danish Ministry of Education funded the implementation of a large field experiment in fourth grade in the Danish primary schools. The results of this trial are reported by Andersen et al. (2016). The randomized trial that we report here was a follow-up to the first trial.[1] The participating schools were the same in the two experiments, and the design of the follow-up trial built on the first trial. A brief presentation of the design of the first trial therefore underpins the description of the follow-up trial. Figure 1 shows the overall design and relationship between the two rounds of experiments.

[1] Only one fourth grade classroom per school participated in the first round. In the second round, one fourth grade and one fifth grade classroom per school participated. Initial analyses documented large and skewed attrition along with noncompliance among the fourth grade classrooms (less than 60 percent of the classrooms complied with the assigned treatment). The analysis in this paper focuses on the fifth grade classrooms.

Figure 1: Diagram of the randomization of schools and classroom across the first and the follow-up trial



The Ministry of Education invited all Danish public schools to participate in the first trial. Since one of the main objectives was to improve the academic performance of bilingual students, who generally come from low-SES homes (approximately 50 percent of the bilingual students compared to 15 percent of the remaining students in our sample), the only inclusion criterion was that the schools should expect to have at least 10 percent bilingual students in fourth grade in the fall of 2013. Initially, 126 schools enrolled in the trial. Two schools with less than 10 percent non-Western students were randomly excluded from the stratified randomization due to resource constraints. The stratified randomization in the first-round experiment thereby included 124 schools. Within each school, one fourth grade classroom was randomly selected for participation.

The schools were allocated to either one of three treatment groups or the control group.[2] The treatments consisted of increased instruction time, increased instruction time with a teaching program, and an upgrade of the qualifications of the math teacher. Andersen et al. (2016) document positive and statistically significant effects on the reading test scores of the treatments that involved increased instruction time in Danish. However, whereas the overall treatment effects were large and positive, there was no indication that the increased instruction time had positive effects on the students of non-Western origin.[3] The point estimate for this group was statistically insignificant and close to zero. These results motivated us to adjust the treatments in the follow-up trial in an attempt to target the more disadvantaged students.

[2]Allocation to treatment and control groups was based on a two-stage stratified cluster randomization. The schools were stratified according to the share of non-Western students and second grade reading test scores. One fourth grade classroom per school was selected for participation based on simple randomization.

[3]We use both the terms "bilingual" and "non-Western" students. We consider these two groups of students to be overlapping to a large extent. In the administrative data, we do not have information about whether or not a student is bilingual—only whether the student is of non-Western origin. On the other hand, the actual requirements for participation were phrased in terms of "bilingual" students.

2.3 The Follow-Up Trial

The follow-up trial was implemented in the fall of 2014. The 124 schools in the first-round experiment were requested to participate in the second round as well. Only six schools withdrew from the follow-up trial prior to randomization (see Figure 1).

Certain restrictions on the design, which we explain below, meant that only groups A and B (the two left-hand treatment arms in Figure 1) were randomized to either a treatment or the control group in the follow-up trial. We run separate robustness analyses, including only these two groups. All main results are robust when only including these two fully randomized groups (see section 4.4.2). To increase statistical power, we include schools from groups C and D (the two right-hand arms in Figure 1) in the main analysis. Even though groups C and D were not randomized in the second round, their treatment status was determined only by the randomization of the first round and not by any self-selection.

The design restrictions were due to the Ministry of Education's initial promise that schools in the control group in the first-round experiment (group D in Figure 1) would receive the fourth grade treatment in the following cohort of fourth grade students. Therefore, the fifth grade students in these schools (group D) were maintained in the control group in the follow-up trial.[4] Schools in group C were all placed in the control group in the follow-up trial because the teaching program in the first-round randomization was similar to the program tested in the follow-up trial (see the next section for further information on the follow-up trial).

Schools in groups A and B were stratified based on the share of students of non-Western origin and the average performance on reading tests in second grade in 2013. As the schools in groups C and D were not stratified and randomized, we do not include strata fixed effects in our main models. However, when we only include the two fully randomized groups (see section 4.4.2) strata fixed effects do not change the results. The classrooms that participated

[4]The fourth grade classrooms were assigned to one of two treatment arms: (1) increased instruction time in Danish and (2) increased instruction time in Danish coupled with monthly tests.

in the first-round trial also participated in the follow-up trial.

The Danish Ministry of Education reimbursed the participating schools for all costs associated with the experiment. It was a requirement for reimbursement that they participated in the data collection and implemented the intervention they were assigned to. The control schools also received reimbursement for their costs of participating in the data collection.

2.4 The Intervention

The intervention in the follow-up trial had two basic components:

i. A teaching program including two extra lessons per week (additional instruction time).

ii. Regular testing including feedback to the teacher about student performance.

The first component, the teaching program, was called "General language comprehension" and included both texts and classroom exercises. National experts in language instruction specifically developed the teaching material for the experiments. The intention was to improve the learning outcomes of bilingual students and other students with low proficiency in Danish. The program included instructions on how to increase knowledge of words, weekly focused readings on specific topics, and weekly semantic topics, such as sayings. All elements were expected to upgrade the students reading skills by improving general language comprehension.

The teaching program lasted for sixteen weeks. Each week, the students in the treatment group were given (i) two extra lessons in which the students in the control group would get off from school, (ii) one lesson in which the other students would do homework or similar activities in school with the support of a teacher, and (iii) one lesson including a co-teacher in which the control group students would have just one teacher in the classroom. All lessons lasted for forty-five minutes.[5] A recent report shows that when 11-year-old students

[5]In the first trial, the interventions included four additional lessons in Danish. This was not feasible in the follow-up trial since a major school reform, which increased instruction time in all schools, was implemented in the beginning of the school year 2014/2015. The treatment and control group schools were all subject to the reform.

in Denmark are off from school, 36 percent attend a non-educational youth club, 80 percent attend spare time activities, such as sports clubs or gyms, and 76 percent play computer games for at least one hour a day (Ottosen et al., 2018). The additional instruction time was implemented without increasing the workload of the teachers since the teachers were released from other teaching duties.

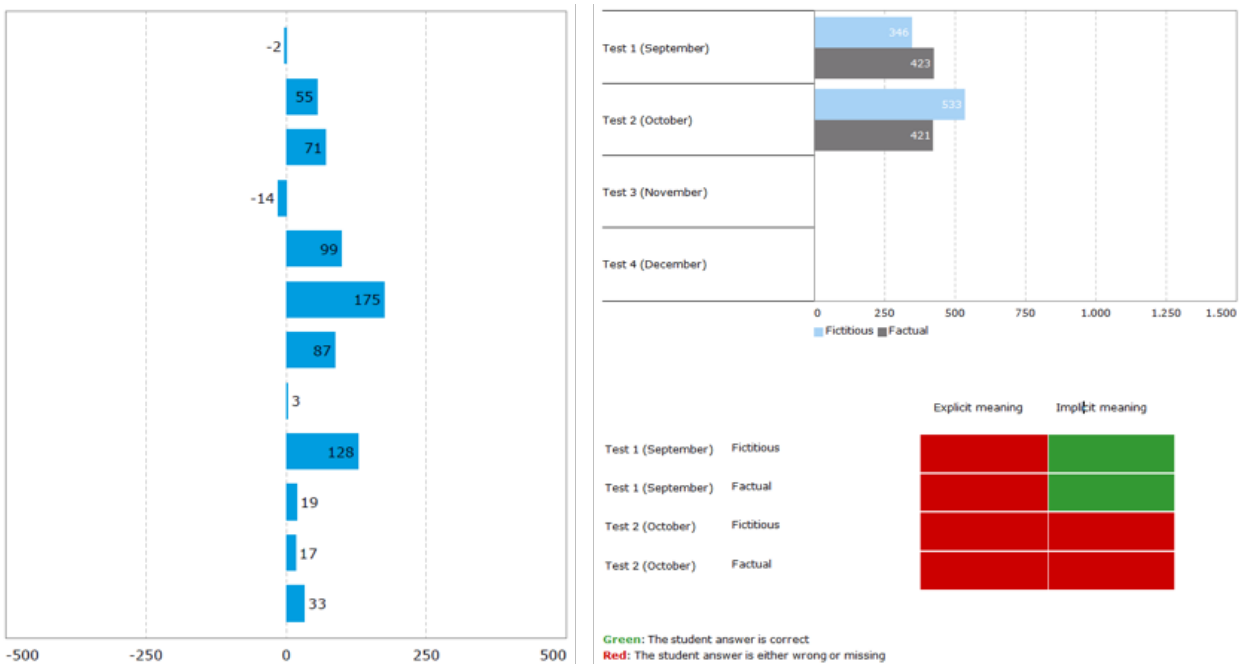
The second component was regular testing of the students. The first treatment component (the teaching program) was very similar to the treatments in the first trial. As mentioned, these treatments were not found to improve the academic performance of the students of non-Western origin (Andersen et al., 2016). As argued by Raudenbush (2008), the effect of increasing school resources is likely to depend on the instructional regime in the school, i.e., the set of rules for how to regulate the interplay between assessment and instruction (see also Banerjee et al., 2007, 2017; Duflo et al., 2011). In order to encourage the teachers to adapt the instruction to the level of more disadvantaged students, we added the regular testing component. The intention was that the availability of information about the students' performance would help the teachers target the instruction toward the more disadvantaged students.

Each month (four times in total), the students took a reading test consisting of one fictional and one factual text. The students should read each text for five minutes and mark the number of words they had finished reading when the time was up.[6] In addition, the students answered two questions about the explicit and implicit meaning of the text. The language teacher received a summary of the students' performance (for every test) and progression since the previous test (except, of course, for the first test). Furthermore, the teachers received a student report consisting of scorecards with individual results from each student. Figure 2 shows an example of a class overview (left-hand side) and a student scorecard (right-hand side). Though the test is rather simplistic, the test results explain 40

[6]We did a pilot test of the material to assess the difficulty level of the text material before initiating the intervention. We used these data to sort the test texts in order of difficulty level. To account for the text difficulty in the test scores, we standardized the reading scores from the second to the fourth test following the distribution of the test results from the first test.

percent of the variation in the participating students' national test scores in reading. The national reading tests are used to evaluate the effects of the intervention and should not be confused with the simple reading tests that are implemented as part of the intervention.

Figure 2: Example of class overview of progression (left) and student score card (right)



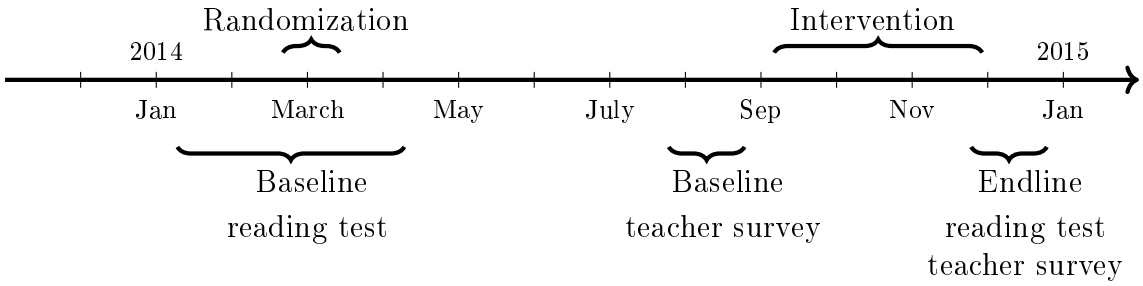
Note: The left-hand side of the figure shows the change in reading speed (words per minute) for each student within a class relative to the previous test. Positive scores indicate an increase in reading speed. Names of the individual students were placed to the left of the figure but have been deleted to protect anonymity. The figure on the right-hand side shows the test results of an individual student. The upper part shows reading speed for test 1 (September) and test 2 (October) for both a fictional (blue bars) and a factual text (gray bars). The lower part on the right-hand side shows the student's responses to two questions for each text. One question asks about part of the explicit meaning of the text (e.g., "Who took the cookie?"), and a second question asks about the implicit meaning of the text (e.g., "Why was he angry?"). Red squares indicate that the student answered the question incorrectly; green squares indicate correct answers.

2.5 Timeline of Experiment and Data Collection

The follow-up trial was conducted during the fall of 2014. The intervention period was sixteen weeks, and an extensive data collection was undertaken in connection with the experiment. To the extent possible, the experiment was designed so that the evaluation could be based on the administrative data. Figure 3 shows a timeline of the experiment and the associated

data collection.

Figure 3: Timeline of experiment and data collection



The randomization of the follow-up trial took place in March 2014, and schools were subsequently informed about the treatment status of their participating fifth grade classroom. The actual intervention was implemented from September 1 to December 13.[7] Surveys were conducted before and after the intervention with both students, teachers, and principals. Two implementation surveys with teachers were conducted during the intervention to assess fidelity.[8] The main outcome measure is academic performance as measured by the national reading test. In the period from January to April 2014, the participating students took the mandatory fourth grade reading test. We use this test as a baseline measure of academic performance. At the end of the intervention period, participating classrooms were instructed to take the national reading test again.

3 Data, Outcomes, and Balance

The surveys conducted in relation to the experiment are linked to administrative data hosted by Statistics Denmark and the National Agency for It and Learning, which both use unique individual identifiers. The detailed administrative data provide information about national test scores, gender, country of origin, socioeconomic background, and other relevant background variables. The student surveys were mainly used to evaluate the effect on student

[7] In Denmark, the school year begins in August.

[8] The implementation surveys were given to the Danish teachers and used to assess the degree to which teachers and schools complied with the intervention. All surveys were administered electronically.

behavioral problems and included the Strengths and Difficulties Questionnaire (SDQ) (Goodman and Goodman, 2009). We used the teacher surveys' pre- and post-intervention to extract information on the teachers' beliefs in their students' reading abilities.

3.1 Measurement of Key Variables

3.1.1 Baseline and Endline Reading Performance

Baseline reading performance is based on the students' scores on the fourth grade national reading test conducted in the period from January to April 2014 prior to the intervention (see Figure 3). The Danish national tests are standardized computer-scored tests.

At the end of the intervention period, the students took either the fourth grade or the sixth grade national reading test (see Figure 3).[9] They were encouraged to take the fourth grade reading test, but a smaller fraction (approximately 20 percent) of the students in our sample took the sixth grade reading test. Taking the sixth grade test does not correlate significantly with treatment status. Since we expect that the sixth grade test has a higher difficulty level, we account for this by including an indicator of whether a student took the fourth or the sixth grade reading test when we estimate the intervention effects on reading performance.

Each test consists of three subtest scores. We follow Nandrup and Beuchert-Pedersen (2018) and standardize our reading scores to zero mean and unit variance within each of the three domains based on the mean and variance of the population of students taking the corresponding mandatory national tests. We then take the average and standardize once again. The estimated intervention effects on reading achievement will thus be in terms of a standard deviation in the population.

[9] These tests were voluntary tests corresponding to the mandatory national tests in reading for fourth and sixth grades.

3.1.2 Gaps between Teacher Beliefs and Actual Reading Performance

For the purpose of investigating how teachers assess the reading skills of students with different SES, we need a measure of how the teachers assess their students' reading abilities along with an objective measure of the students' reading abilities. We elicited the teachers' beliefs in their students' reading abilities in the teacher surveys conducted both before and after the intervention. In the survey, the teachers were asked to rank their students' reading abilities according to five categories from "certainly below average" to "certainly above average."[10] In the ranking process, we allowed the teachers to decide how to distribute the students across the five categories. The teachers could therefore assess as many students as they liked into each of the five categories.[11]

The objective measure of the students' reading abilities (or the students' actual reading performance) is based on the students' scores on the mandatory fourth grade national reading test—our measure of baseline reading performance. The test score is based on a computer algorithm and is thereby blinded to the teachers' prior beliefs, the students' experimental condition in the trial, and other factors that might conflate the objective measure with the trial. When the students take the national tests, the teachers subsequently have access to the results. However, not all teachers may logon to the test score system to see their students' results, or they may not know how to interpret the results. In other words, there may be some level of information friction in this system.

Based on the baseline reading test scores, the students were ranked and allocated into five groups based on their rank. Within each classroom, we placed the same number of students in each group as the teacher had placed in each of the categories. This means that if, for instance, a teacher had placed 10 percent of the students in the category "certainly above

[10] While we do not know the teachers' reference points for certain, we know that they are probably not the same as the national norms since the sample on average had reading scores slightly below the national average, but the teachers assessed 17 percent of the students as "certainly above average," 23 percent "above average," 35 percent "about average," 17 percent "below average," and 7 percent "certainly below average." With more than 40 percent of the students in the top two categories, the distribution also shows that the teachers had a slightly positive bias in their beliefs.

[11] See appendix B for the exact wording of the question.

average," the best 10 percent of the students according to the reading test were also placed in that category.

In line with Dizon-Ross (2019), we measure inaccuracies in the teachers' beliefs as the absolute value of the gap between the teachers' rank of student i and the test score-based rank:[12]

$$Gap_i = |TestRank_i - TeacherRank_i| \tag{1}$$

One obvious question is whether the gaps reflect low reliability in the test score measure or low measurement validity in the sense that the teachers evaluate other dimensions of the students' reading skills than what the test scores capture. We have different indications that this is not the case. First, the tests explain approximately 50 percent of the variation in the corresponding ninth grade exit exam results in the same subject (Nandrup and Beuchert-Pedersen, 2018), which indicates a high degree of measurement validity. Second, if the experiment shows that teachers update their beliefs based on the performance information, and if they do so in the direction of the objective test scores, it suggests that they believe that their prior beliefs were inaccurate.

3.2 Socioeconomic Status

In order to measure whether the assessment varies with the socioeconomic status of the students, we use administrative data on the parents' highest level of education. In Denmark, large redistribution policies ensure low income inequality, however, educational inequality remains substantial (Landersø and Heckman, 2017). We define students as having low SES if neither of the parents has a college degree. If one of the parents was missing in the administrative data, we set the value to missing.[13] If at least one of the parents had a college degree (vocational or academic), we coded the student as a high-SES student. In our

[12] If we think of the teacher assessment as a forecast of true performance, the teachers' test divergence corresponds to the absolute forecast error. We focus on the absolute forecast error since we do not find any evidence that the teachers' forecasts are biased. Strictly speaking, the teachers' beliefs cannot be forecasts since teachers already know (or at least have access to) the true performance of the students.

[13] Using the information for students with one parent missing produces similar results (see section 4.4.2).

robustness analysis, we construct an alternative SES measure based on parental income.[14] All main results are similar when we base the SES measure on parental income rather than education. When we compare the intervention effects of students of Western origin with students of non-Western origin, we do not see the same differential intervention effects. SES and country of origin are correlated, but far from all students of non-Western origin have low SES. As mentioned, this is true for approximately 50 percent of the non-Western students in our sample. Our findings suggest that children's SES—rather than their country of origin—is important for the interplay between the teachers' beliefs, information, and instruction time.

3.3 Balance and Attrition

Table 1 shows that—at the outset—randomization successfully created comparable student groups on a number of covariates obtained from administrative registers. A negative average baseline reading score is a reflection of the fact that the participating schools are relatively disadvantaged.[15] Correspondingly, 26 percent of the participating students are of non-Western origin, which is substantially above the average of Danish public schools. About 20 percent of the students have low SES. The differences between the treatment and control groups are all small and statistically insignificant. Based on a joint F-test of the null of no differences between the treatment and control groups, we cannot reject the null.

Despite a little attrition, our estimation sample is also balanced between the treatment and control groups. Out of the 118 schools assigned to either the treatment or the control group, thirteen schools actively chose not to contribute to the survey data collection in the second round, primarily because they did not have the time to participate. The share of schools deciding not to contribute was equal in the treatment and control groups (10–11 percent). In addition, for nineteen schools, we did not observe the main outcome (reading

[14] We divide the students into quartiles based on the parent with the highest income and define low-SES students as students in the lowest quartile in the sample of fifth grade students. This is similar to the procedure used for the whole population in Nandrup and Beuchert-Pedersen (2018).

[15] Andersen et al. (2016) show that the distribution of reading scores for the participating schools is shifted to the left compared to the distribution for all public schools.

Table 1: Balance at baseline

	Control	Treatment	Difference
Baseline reading score	-0.08	-0.07	-0.01
Baseline reading score (m)	0.04	0.04	0.01
Low SES	0.21	0.20	0.00
Low SES (m)	0.07	0.07	0.01
Female	0.50	0.51	-0.01
Female (m)	0.03	0.03	0.00
Non-Western origin	0.26	0.26	0.00
Non-Western origin (m)	0.01	0.01	0.00
Baseline test score-belief gap	0.47	0.41	0.07
Baseline test score-belief gap (m)	0.31	0.20	0.11
Observations	1,927	619	

Note: The table reports means in the treatment and control groups and in the corresponding difference. Low SES, female, and non-Western origin are indicator variables. (m) denotes missing variable indicators. No differences are significant at the 10 percent level.

scores). This attrition was not evenly distributed across treatment (3.3 percent of schools) and control groups (18.2 percent of schools). However, section 4.4.1 shows that the differences between treatment and control groups are small and, except for one covariate, statistically insignificant at the 10 percent level in the estimation sample (the sample for which we observe the reading score). The main estimation sample consists of 1,518 students from eighty-six schools in total.

4 Results

Our analysis first examines the intervention effects for all students and for low- and high-SES students separately. Afterward, we explore the potential mechanism that drives the heterogeneous effects we find.

4.1 Intervention Effects

We use the following basic model to estimate the intervention effect on the students' outcomes:

$$Y_{ij} = \gamma_0 + \gamma_1 Treatment_j + \gamma_2 Prescore_i + \gamma_3 I[Grade6]_i + u_{ij}, \quad (2)$$

where Y_{ij} is the reading test score of student i in classroom j, $Treatment_j$ is an indicator of whether or not classroom j was in the treatment group, $Prescore$ is the fourth grade reading score (the baseline reading score), $I[Grade6]$ is an indicator of whether or not the student took the sixth grade (instead of the fourth grade) endline reading test, and u is the idiosyncratic error term.[16] We also include an indicator of missing baseline reading scores. We interpret γ_1 as the intention-to-treat effect of the intervention.[17] The baseline test score is included to improve precision in the estimation. In order to estimate the intervention effects moderated by SES, most of our specifications include an indicator of high-SES status and the interaction of this indicator with $Treatment$. In this case, the estimated coefficient on $Treatment$ will reflect the intervention effect for the group of low-SES students. To further assess the influence of any potential imbalance, we have also run regression models with and without covariates (not shown). In general, estimates are robust and, if anything, become more statistically significant when adding covariates as it would be expected based on the random assignment.

[16] It is only possible to include stratum fixed effects when we focus on the schools that were allocated to treatment or control groups based on the follow-up trial. Section 4.4.2 presents the results when the analysis is restricted to schools that were randomized in the follow-up trial with and without stratum fixed effects. The main results are robust to the inclusion of stratum fixed effects.

[17] In the treatment group, 66 percent of the teachers implemented all the reading tests and 80 percent implemented at least three out of four reading tests. This corresponds to all the schools that did not decline to contribute to the data collection or stated that they did not want to conduct the treatment upfront. Also, 80 percent of the schools implemented at least one additional lesson of instruction per week, and 63 percent of the schools conducted at least three lessons per week (two additional lessons and one placed either in the regular Danish instruction or scheduled in periods during the week when the students did not have formal classes).

4.1.1 Intervention Effects on Students' Reading Test Scores

Table 2 shows the intervention effects on the students' reading test scores. The average intervention effect is 0.05 (model 1) and somewhat smaller when including covariates in model 2, but in both cases it is statistically insignificant. When we allow for differential intervention effects by SES status, the estimated intervention effect is positive and substantially larger for low-SES students than for high-SES students (model 3). When baseline test scores are included to improve precision, the intervention effect for low-SES students is statistically significant at the 1 percent level, and so is the interaction term (model 4).

In other words, the intervention improved the reading performance of low-SES students. When controlling for baseline test scores, the effect size is almost as large as the control group gap between high- and low-SES students, which is reflected in the coefficient on high SES. Earlier studies suggest that increased instruction time may widen existing achievement gaps (Huebener et al., 2017; Andersen et al., 2016). Our results show that combining increased instruction time with regular information to teachers about student progress can improve the outcomes of disadvantaged students and narrow the gap in student achievement.

4.1.2 Intervention Effects on Students' Behavioral Problems

Improved learning should not come at the cost of increased behavioral problems or lower levels of student well-being. We measured the students' behavioral problems with the Strength and Difficulty Questionnaire (SDQ) (Goodman and Goodman, 2009). The SDQ measures the number of difficulties a child might experience. Thus, lower values indicate fewer behavioral problems. Table 3 shows that the intervention did not worsen the students' well-being on average. On the contrary, the treatment decreased the SDQ score for low-SES students and closed the control group gap in well-being between high- and low-SES students, whereas it remained more or less constant for high-SES students in the treatment group.

In sum, the intervention had positive effects on reading skills and led to fewer behavioral problems—and the effects were primarily driven by low-SES students. In the next section,

Table 2: Intervention effects on reading test scores

	(1)	(2)	(3)	(4)
Treatment	0.0457	0.0226	0.180	0.227**
	(0.0929)	(0.0435)	(0.143)	(0.0817)
High SES			0.700**	0.265**
			(0.0722)	(0.0447)
Treatment × High SES			-0.184	-0.263**
			(0.147)	(0.0788)
High SES (m)			-0.0520	0.0277
			(0.149)	(0.0957)
Treatment × High SES (m)			-0.300	-0.323
			(0.530)	(0.277)
Constant	0.0786	0.192**	-0.418**	-0.000931
	(0.0583)	(0.0338)	(0.0815)	(0.0570)
Observations	1,518	1,518	1,518	1,518
Sixth grade reading test	+	+	+	+
Baseline reading score	-	+	-	+
Adjusted R-squared	0.15	0.69	0.23	0.70

Note: Estimated coefficients based on OLS regressions with the endline reading score as the dependent variable. Specifications with the sixth grade reading test include an indicator for whether the endline test taken was the fourth grade or the sixth grade test. Specification with the baseline reading score includes the fourth grade reading score and an indicator for whether or not the score is missing. Standard errors clustered at school level in parentheses. + p<0.1, * p<0.05, ** p<0.01.

we examine what mechanisms may explain these results.

4.2 Exploring Mechanisms: The Importance of Information

4.2.1 Did the Intervention Improve the Accuracy of Teachers' Beliefs?

In order to explore the role of the information component of the combined instruction time and information intervention, we first examine the effect of the intervention on the accuracy of the teachers' beliefs. The results in Table 4 show that the intervention improved the accuracy of the teachers' beliefs measured as the absolute gap between teacher and test-based rankings of the students (model 1). This result is robust to the inclusion of the baseline

Table 3: Effect of treatment on SDQ scores. Interaction with SES

	(1)	(2)	(3)	(4)
Treatment	-0.317	-0.284	-1.470^{+}	-1.532*
	(0.352)	(0.327)	(0.747)	(0.678)
High SES			-1.821**	-1.079**
			(0.419)	(0.408)
Treatment × High SES			1.604^{+}	1.726*
			(0.824)	(0.759)
High SES (m)			0.600	0.236
			(0.901)	(0.883)
Treatment × High SES (m)			-1.008	-0.786
			(1.676)	(1.552)
Constant	9.711**	9.546**	11.04**	10.35**
	(0.208)	(0.198)	(0.372)	(0.366)
Observations	1,879	1,879	1,879	1,879
Baseline reading score	-	+	-	+
Adjusted R-squared	0.00	0.07	0.02	0.07

Note: Estimated coefficients based on OLS regressions with the endline SDQ score as the dependent variable. Specification with the baseline reading score includes the fourth grade reading score and an indicator for whether or not the score is missing. Standard errors clustered at the school level in parentheses. + p<0.1, * p<0.05, ** p<0.01.

teacher test gap (model 2).[18] In models 3 and 4, we allow the intervention effect to vary with SES status. For low-SES students, the intervention improved the accuracy of the teachers' beliefs, reducing the gap. This effect is statistically significant at the 5 percent level when the baseline teacher test gap is included. The interaction term is positive, but statistically insignificant. Thus, the intervention improved the accuracy of the teachers' beliefs, and the direction of the effects suggest that the intervention was more effective in improving the accuracy of the teachers' beliefs for low-SES students than for high-SES students.[19] As we will show later, at baseline, the teachers had more inaccurate beliefs about low- than high-SES students, albeit the difference is not statistically significant. This may be part of

[18] Missing data on baseline and endline teacher rankings imply that this sample differs slightly from the sample used in the reading score analysis. However, the treatment and control groups still balance well, see section 4.4.1.

[19] The measure of the teacher test gap is technically an ordered response variable. An ordered logit analysis produces substantially similar results, see section 4.4.2.

the explanation for why the information intervention was more effective for this group of students.

Table 4: Intervention effects on gap between teacher beliefs and reading test scores

	(1)	(2)	(3)	(4)
Treatment	-0.0865*	-0.0656+	-0.149+	-0.143*
	(0.0410)	(0.0381)	(0.0756)	(0.0689)
High SES			-0.0338	-0.0151
			(0.0376)	(0.0357)
Treatment × High SES			0.0732	0.0958
			(0.0737)	(0.0700)
High SES (m)			-0.0793	-0.0432
			(0.0993)	(0.101)
Treatment × High SES (m)			0.226	0.184
			(0.199)	(0.206)
Constant	0.504**	0.333**	0.532**	0.345**
	(0.0246)	(0.0265)	(0.0351)	(0.0344)
Observations	1,733	1,733	1,733	1,733
Baseline test score-belief gap	-	+	-	+
Adjusted R-squared	0.00	0.12	0.00	0.12

Note: Estimated coefficients based on OLS regressions with the endline test score-belief gap score as the dependent variable. Specification with the baseline test score-belief gap includes the baseline test score-belief gap and an indicator for whether or not the score is missing. Standard errors clustered at the school level in parentheses. + p<0.1, * p<0.05, ** p<0.01.

4.2.2 How are Teachers' Beliefs Related to Students' Reading Skills?

Assuming that high-SES students have better learning opportunities outside of school, low-SES students may be more dependent on having teachers that accurately perceive their level of competence and use instruction time to target their current level. Using only the control group, Table 5 shows correlational evidence that accurate teacher perception at baseline was more strongly correlated with the students' subsequent reading skills when controlling for reading skills at the baseline.

Whereas this cannot be interpreted as causal evidence—other factors may be correlated with both teacher beliefs and student outcomes—the results in Table 5 support the notion

that accurate teacher beliefs (and hence more appropriate teaching level) are more important to low-SES students. Interestingly, we do not find evidence that teachers systematically over- or underestimate students or groups of students (not shown). The beliefs are just more inaccurate for low-SES students.

Table 5: Reading test scores and accurateness of teacher's beliefs by SES. Control group only

	(1)	(2)
High SES	0.288**	0.411**
	(0.0445)	(0.0570)
No test score-belief gap	0.0711	0.265**
	(0.0451)	(0.0726)
High SES × No test score-belief gap		-0.259**
		(0.0685)
High SES (m)	0.0145	-0.00311
	(0.0940)	(0.0907)
No test score-belief gap (m)	0.0736	0.0863
	(0.0792)	(0.0778)
Constant	-0.0581	-0.151+
	(0.0778)	(0.0843)
Observations	1086	1086
Baseline reading score	+	+
Sixth grade reading test	+	+
Adjusted R-squared	0.69	0.69

Note: Estimated coefficients based on OLS regressions with the endline reading score as the dependent variable. Specifications with the sixth grade reading test include an indicator for whether the endline test taken was the fourth or the sixth test. Specification with the baseline reading score includes the fourth grade reading score and an indicator for whether or not the score is missing. Standard errors clustered at school level in parentheses. + p<0.1, * p<0.05, ** p<0.01.

4.2.3 Was the Intervention More Effective When Teachers Had Inaccurate Beliefs About Students at Baseline?

If the inaccuracy of the teachers' beliefs partly explains why low-SES students have lower reading skills, we would expect that low-SES students with high levels of baseline teacher test gap would benefit the most from the teachers receiving more information. To test this,

we examine whether the intervention effect is higher for students with inaccurate teacher baseline beliefs in their performance. Model 1 in Table 6 shows that the intervention tends to be more effective for students whose teachers had inaccurate beliefs about their reading skills at baseline, even though the difference is not statistically significant. In models 2 and 3 we divide the students into SES. Model 2 shows that the intervention effect is substantially larger for low-SES students for whom the teachers' beliefs were inaccurate. The effect is statistically significant at the 1 percent level. The interaction effect is correspondingly large, negative, and statistically significant at the 5 percent level, which suggests that the overall positive intervention effects for low-SES students are driven by the group of students whose teachers had inaccurate beliefs in their reading abilities at baseline. Model 3 confirms that for high-SES students, the teachers' inaccurate baseline beliefs do not seem to moderate the impact of the intervention.

Since these results indicate that the teachers' beliefs are especially important to low-SES students and that the intervention improved the accuracy of the teachers' beliefs and was most effective for low-SES students when the teachers had inaccurate beliefs at baseline, it becomes relevant to examine which factors predict the accuracy of the teachers' beliefs in their students' reading skills. We examine this question in the next section.

4.3 What Predicts the Accuracy of Teacher Beliefs?

Table 7 shows the result of regressing our accuracy measure, i.e., the gap between teacher beliefs and reading test results at baseline, on both student and teacher characteristics. We find that the teachers' beliefs in the reading skills are slightly less accurate for low-SES students than for high-SES students, but the difference is not statistically significant. Similarly, the beliefs are slightly less accurate for students of non-Western origin, but again, they are not significant.

Looking at teacher characteristics, Table 7 shows a negative relationship between the test score belief gap and teacher experience as proxied by the teacher's age. We find the same

Table 6: Intervention effects on reading test scores moderated by inaccuracy of teachers' beliefs

	All	Low SES	High SES
	(1)	(2)	(3)
Treatment	0.0564	0.430**	-0.0541
	(0.0646)	(0.127)	(0.0640)
No test score-belief gap	0.0660	0.278*	0.00233
	(0.0434)	(0.107)	(0.0428)
Treatment × No test score-belief gap	-0.0140	-0.340*	0.0809
	(0.0675)	(0.166)	(0.0685)
No test score-belief gap (m)	0.0479	0.120	0.0312
	(0.0856)	(0.151)	(0.0661)
Treatment × No test score-belief gap (m)	-0.437*	-0.327	-0.381*
	(0.184)	(0.289)	(0.154)
Constant	0.150**	-0.164	0.252**
	(0.0535)	(0.115)	(0.0436)
Observations	1518	340	1119
Sixth grade reading test	+	+	+
Baseline reading score	+	+	+
Adjusted R-squared	0.69	0.63	0.70

Note: Estimated coefficients based on OLS regressions with the endline reading score as the dependent variable. Specifications with sixth grade reading test include an indicator for whether or not the endline test taken was the fourth grade or the sixth grade test. Specification with the baseline reading score includes the grade 4 reading score and an indicator for whether or not the score is missing.
Standard errors clustered at the school level in parentheses.
+ p<0.1, * p<0.05, ** p<0.01.

overall tendency when using a survey-based and less fine-grained measure of experience. As mentioned, teacher experience remains one of the teacher characteristics related to teacher quality with the most empirical support (Gerritsen et al., 2017; Staiger and Rockoff, 2010). As expected, more experienced teachers have more accurate beliefs in students' reading skills. This suggests that inaccurate beliefs in student skills could be one of the channels through which teacher experience affects student performance. In line with this finding, Lavy and Megalokonomou (2019) show that better teachers exhibit lower gender bias in their grading.

Table 7: Student and teacher characteristics predicting gap between teacher beliefs and test scores

	(1)	(2)	(3)	(4)	(5)	(6)
Student high SES	-0.0433					
	(0.0406)					
Student high SES (m)	-0.0919					
	(0.104)					
Student female		-0.0413				
		(0.0316)				
Student female (m)		-0.0832				
		(0.128)				
Student non-Western origin			0.0460			
			(0.0385)			
Student non-Western origin (m)			n.a.			
Teacher female				0.0111		
				(0.0686)		
Teacher age					-0.00434*	
					(0.00209)	
Teacher (m)				0.112	-0.101	
				(0.108)	(0.137)	
Experience (ref. 0-5 years)						
-6-10 years						-0.0679
						(0.0709)
-11-20 years						-0.0433
						(0.0735)
-20+ years						-0.117+
						(0.0681)
Constant	0.488**	0.474**	0.440**	0.439**	0.652**	0.514**
	(0.0411)	(0.0278)	(0.0244)	(0.0637)	(0.106)	(0.0572)
Observations	1,314	1,314	1,314	1,314	1,314	1,314
Adjusted R-squared	-0.00	-0.00	0.00	-0.00	0.00	0.00

Note: Estimated coefficients based on OLS regressions with the baseline test score-belief gap as the dependent variable. A cell with less than six observations is omitted to comply with Statistics Denmark's data security policy.

Standard errors clustered at the school level in parentheses. + p<0.1, * p<0.05, ** p<0.01.

4.4 Robustness Analysis

4.4.1 Balance in Analysis Samples

Since we observe some attrition in the two outcome measures we use, i.e., reading test scores and teacher beliefs, we test how the treatment and control groups balance when we restrict the samples to those students whose outcomes we observe. In the appendices, Table A1 shows the balance in the sample with valid reading test scores at endline. We find only small differences, and only one out of ten variables differs at the 10 percent significance level. Table A2 shows the result of the sample with valid measures of teacher perception, and the result is similar. Table A3 shows the balance of the analysis sample with both reading test scores and teacher belief measures, and here we find no statistically significant differences.

4.4.2 Robustness of Effect Estimates

We conduct a number of robustness tests to ensure that our main findings in Table 2 are not sensitive to the specification of the model. In general, the results are robust across different specifications. First, the initial trial determined the treatment status of some classrooms. Neither the exclusion of the classrooms only randomized in the first trial (groups C and D in the two right-hand arms in Figure 1) nor the inclusion of strata fixed effects lead to substantial changes in the estimated effects on the reading scores (see Table A4). Therefore, results are not sensitive to the inclusion or exclusion of schools that were only randomized in the first round.

Second, to ensure that the results are not affected by attrition in the teacher belief endline survey, we run the primary model conditioning on valid measures of teacher beliefs. Table A5 shows that we find similar effects in this restricted sample.

Third, as a further robustness check, we conducted the analysis with alternative measures of socioeconomic background. We use parental income (based on the parent with the highest income) and define low SES as belonging to the lowest income quartile in the sample of fifth

grade students in the study. We also test that using observations with only one valid parent in the educational based measure of SES does not change the result. In the appendices, Table A6 models 2 and 3 show that the results are similar to the main results presented in Table 2 with these alternative measures of SES. Models 1 and 4 in Table A6 are identical to models 2 and 4 in Table 2 and only repeated in the appendices for ease of reference. The treatment significantly improved low-SES students' reading scores, and this effect is significantly lower for high-SES students. As mentioned, income inequality in Denmark is among the lowest in the world, but the same is not true for educational inequality (Landersø and Heckman, 2017). We therefore believe that the most interesting differences are those based on the educational measure of SES.

Finally, since the test score belief gap variable is ordinal, we estimate the effect of the treatment on this gap by using an ordered logit model as a supplementary analysis. Table A7 presents the results, using this model specification. Again, we find that the treatment reduces the test score beliefs gap and that this effect was especially pronounced for low-SES students.

5 Conclusion and Discussion

At baseline of this study, teachers had access to information about their students' test scores in computer-based, adaptive reading tests. Nevertheless, their beliefs in their students' reading skills differed somewhat from the objective test results. Consistent with a simple model of information frictions and Bayesian learning, the teachers in the treatment group updated their beliefs so that at the end of the intervention period, the gap between beliefs and test scores was smaller than in the control group.

Furthermore, we found correlational evidence that the accuracy of the teachers' beliefs is a more important factor in explaining the reading scores of low-SES students compared with the reading scores of high-SES students. We do not have data to test why that is, but

one natural explanation may be that learning outside of school is more effective for high-SES students and that parents may compensate more at home if the teachers do not target the instruction at an adequate level for the students.

The combination of teachers updating their beliefs and beliefs being more important for low-SES students may explain why the intervention of this study, which combined extra instruction time with regular testing and readily accessible information to the teachers, led to an improvement of the reading skills of low-SES students. The achievement gap between students from low- and high-SES families appears constant across grade levels and countries (Carneiro and Heckman, 2003; Nandrup and Beuchert-Pedersen, 2018), and reforms and interventions that increase instruction time tend to be less beneficial for disadvantaged students (Andersen et al., 2016; Huebener et al., 2017). Public schools thereby seem unable to compensate low-SES children for a poorer learning environment at home and thereby create equal opportunities for all students. Our estimates suggest that teachers' inattention to low-SES students' abilities and needs might partly account for this effect. Our results provide cause for optimism in relation to improving the learning possibilities for low-SES students. We show that combining additional instruction time with systematic information about the students' abilities improves the accuracy of the teacher's beliefs in low-SES students' performance and ultimately improves the reading skills of these students. We also find that the treatment reduces behavioral problems for low-SES students, thereby reducing concerns that more instruction time will come at the expense of student well-being in school.

As such, our results are very much in line with the findings of Dizon-Ross (2019), which are based on a field experiment in Malawi. She finds that parents' baseline beliefs are inaccurate and that providing the parents with performance information about their children leads to changes in important human capital decisions, such as school enrollment. She also finds that low-SES parents have more inaccurate beliefs.

The findings are also relevant for the discussion of school accountability systems. In this low-stakes test system, the teachers appear to learn quite a lot from these tests without

risking any detrimental effects of gaming and strategic behavior. Whereas we cannot know whether we can identify similar effects in different national contexts, this suggests that the hard incentives might not even be necessary to acquire gains from these systems.

Finally, our results add to a more general question about learning. As mentioned, the teachers had immediate access to the tests we use as an objective measure of student abilities, but they still held different beliefs in the students' abilities. We encourage future research to look into why exactly this difference emerges. Our findings suggest that when we continuously provide the teachers with systematic information about their students' skills, they update their beliefs and shift their attention to providing the instruction low-SES students need.

References

Andersen, S. C. (2008). Private Schools and the Parents that Choose Them: Empirical Evidence from the Danish School Voucher System. *Scandinavian Political Studies*, 31(1):44–68.

Andersen, S. C., Humlum, M. K., and Nandrup, A. B. (2016). Increasing instruction time in school does increase learning. *Proceedings of the National Academy of Sciences*, 113(27):7481–7484.

Andersen, S. C. and Nielsen, H. S. (2016). The Positive Effects of Nationwide Testing on Student Achievement in a Low-Stakes System. SSRN Scholarly Paper ID 2628809, Social Science Research Network, Rochester, NY.

Bandiera, O., Larcinese, V., and Rasul, I. (2015). Blissful ignorance? a natural experiment on the effect of feedback on students' performance. *Labour Economics*, 34:13 – 25. European Association of Labour Economists 26th Annual Conference.

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M., and

Walton, M. (2017). From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application. *The Journal of Economic Perspectives*, 31(4):73–102.

Banerjee, A. V., Cole, S., Duflo, E., and Linden, L. (2007). Remedying Education: Evidence from Two Randomized Experiments in India. *The Quarterly Journal of Economics*, 122(3):1235–1264.

Barrera-Osorio, F., Gonzalez, K., Lagos, F., and Deming, D. (2018). Effects, Timing and Heterogeneity of the Provision of Information in Education: An Experimental Evaluation in Colombia.

Bergman, P. (2015). Parent-Child Information Frictions and Human Capital Investment: Evidence from a Field Experiment. SSRN Scholarly Paper ID 2622034, Social Science Research Network, Rochester, NY.

Bergman, P. and Chan, E. W. (2017). Leveraging Parents: The Impact of High-Frequency Information on Student Achievement.

Bergman, P. and Rogers, T. (2017). The Impact of Defaults on Technology Adoption, and Its Underappreciation by Policymakers. SSRN Scholarly Paper ID 3098299, Social Science Research Network, Rochester, NY.

Carneiro, P. and Heckman, J. (2003). Human Capital Policy. SSRN Scholarly Paper ID 380480, Social Science Research Network, Rochester, NY.

Cunha, F., Elo, I., and Culhane, J. (2013). Eliciting Maternal Expectations about the Technology of Cognitive Skill Formation. Working Paper 19144, National Bureau of Economic Research.

Damgaard, M. T. and Nielsen, H. S. (2018). Nudging in education. *Economics of Education Review*, 64:313 – 342.

de Montgomery, C. J. and Sievertsen, H. H. (2019). The socio-economic gradient in children's test-scores - a comparison between the U.S. and denmark. *Danish Journal of Economics*.

Dizon-Ross, R. (2019). Parents' Beliefs About Their Children's Academic Ability: Implications for Educational Investments. *American Economic Review*.

Duflo, E., Dupas, P., and Kremer, M. (2011). Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. *American Economic Review*, 101(5):1739–1774.

Figlio, D. and Loeb, S. (2011). School Accountability. In Hanushek, E. A., Machin, S. J., and Woessmann, L., editors, *Handbooks in Economics*. Elsevier, The Netherlands: North-Holland. Google-Books-ID: SY3EJi30oCsC.

Gerritsen, S., Plug, E., and Webbink, D. (2017). Teacher Quality and Student Achievement: Evidence from a Sample of Dutch Twins. *Journal of Applied Econometrics*, 32(3):643–660.

Goodman, A. and Goodman, R. (2009). Strengths and Difficulties Questionnaire as a Dimensional Measure of Child Mental Health. *Journal of the American Academy of Child & Adolescent Psychiatry*, 48(4):400–403.

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review*, 30(3):466–479.

Hanushek, E. A., Peterson, P. E., Talpey, L. M., and Woessmann, L. (2019). The unwavering ses achievement gap: Trends in u.s. student performance. Working Paper 25648, National Bureau of Economic Research.

Huebener, M., Kuger, S., and Marcus, J. (2017). Increased instruction hours and the widening gap in student performance. *Labour Economics*, 47:15–34.

Jacob, B. A. and Levitt, S. D. (2003). Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. *The Quarterly Journal of Economics*, 118(3):843–877.

Landersø, R. and Heckman, J. J. (2017). The scandinavian fantasy: The sources of intergenerational mobility in denmark and the us. *The Scandinavian Journal of Economics*, 119(1):178–230.

Lavy, V. and Megalokonomou, R. (2019). Persistency in teachers' grading bias and effects on longer-term outcomes: University admissions exams and choice of field of study. Working Paper 26021, National Bureau of Economic Research.

Ministry of Education (2008). About Private Schools in Denmark.

Nandrup, A. B. and Beuchert-Pedersen, L. V. (2018). The Danish national tests at a glance. *Danish Journal of Economics*, 2018(1).

OECD (2004). *Reviews of National Policies for Education: Denmark 2004: Lessons from PISA 2000*. Reviews of National Policies for Education. OECD.

OECD (2014). *Education at a Glance 2014: OECD Indicators*. OECD Publishing, Paris. OCLC: 894171152.

OECD (2016). *PISA 2015 Results (Volume II): Policies and Practices for Successful Schools*. OECD Publishing, Paris.

Ottosen, M. H., Andreasen, A. G., Dahl, K. M., Hestbæk, A., Lausten, M., and Rayce, S. B. (2018). *Børn og Unge i Danmark. Velfærd og Trivsel*. VIVE - Det National Forsknings- og Analysecenter for Velfærd, Copenhagen, Denmark.

Papageorge, N. W., Gershenson, S., and Kang, K. M. (2018). Teacher Expectations Matter. Working Paper 25255, National Bureau of Economic Research.

Raudenbush, S. W. (2008). Advancing Educational Policy by Advancing Research on Instruction. *American Educational Research Journal*, 45(1):206–230.

Rivkin, S. G., Hanushek, E. A., and Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73(2):417–458.

Rockoff, J. E., Staiger, D. O., Kane, T. J., and Taylor, E. S. (2012). Information and employee evaluation: Evidence from a randomized intervention in public schools. *The American Economic Review*, pages 10–1257.

Rogers, T. and Feller, A. (2018). Reducing student absences at scale by targeting parents' misbeliefs. *Nature Human Behaviour*, 2(5):335.

Staiger, D. and Rockoff, J. E. (2010). Searching for Effective Teachers with Imperfect Information. *Journal of Economic Perspectives*, 24(3):97–118.

Taylor, E. S. and Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review*, 102(7):3628–51.

Appendices

A Supplementary analyses

Table A1: Balance in sample with valid reading test score

	Control	Treatment	Difference
Baseline reading score	-0.10	-0.09	-0.01
Baseline reading score (m)	0.04	0.02	0.02
Low SES	0.24	0.22	0.02
Low SES (m)	0.04	0.04	-0.00
Female	0.49	0.50	-0.01
Female (m)	0.03	0.02	0.01
Non-Western origin	0.28	0.27	0.01
Non-Western origin (m)	n.a.	n.a.	n.a.
Baseline test score-belief gap	0.47	0.42	0.05
Baseline test score-belief gap (m)	0.16	0.06	0.11+
Observations	1,086	432	

Note: The table reports means in the treatment and control groups and the corresponding difference. Low SES, female and non-Western origin are indicator variables. Cells with less than six observations are omitted to comply with Statistics Denmark's data security policy.
(m) denotes missing variable indicators.
N=1,518. +p<0.1 *p<0.05 **p<0.01.

Table A2: Balance in sample with valid inaccuracy measure

	Control	Treatment	Difference
Baseline reading score	-0.08	-0.08	0.00
Baseline reading score (m)	-	-	-
Low SES	0.23	0.23	0.00
Low SES (m)	0.03	0.04	-0.00
Female	0.49	0.51	-0.02
Female (m)	n.a.	n.a.	n.a.
Non-Western origin	0.27	0.27	-0.00
Non-Western origin (m)	n.a.	n.a.	n.a.
Baseline test score-belief Gap	0.46	0.40	0.06
Baseline test score-belief Gap (m)	0.13	0.04	0.09+
Observations	1,254	479	

Note: The table reports means in the treatment and control groups and the corresponding difference. Low SES, female and non-Western origin are indicator variables. Cells with less than six observations are omitted to comply with Statistics Denmark's data security policy.
(m) denotes missing variable indicators.
N=1,733. + p<0.1, * p<0.05, ** p<0.01.

Table A3: Balance in sample with valid reading test score and valid inaccuracy measure

	Control	Treatment	Difference
Baseline reading score	-0.10	-0.07	-0.03
Baseline reading score (m)	-	-	-
Low SES	0.25	0.23	0.03
Low SES (m)	0.03	0.04	-0.00
Female	0.49	0.51	-0.02
Female (m)	n.a.	n.a.	n.a.
Non-Western origin	0.29	0.28	0.01
Non-Western origin (m)	n.a.	n.a.	n.a.
Baseline test score-belief Gap	0.45	0.41	0.04
Baseline test score-belief Gap (m)	0.11	0.04	0.07
Observations	842	399	

Notes: The table reports means in the treatment and the control groups and the corresponding difference. Low SES, female and non-Western origin are indicator variables. Cells with less than six observations are omitted to comply with Statistics Denmark's data security policy.
(m) denotes missing variable indicators.
N=1,241. + p<0.1, * p<0.05, ** p<0.01.

Table A4: Effect of treatment on reading scores. Interaction with SES. Classrooms without second round randomization excluded

	(1)	(2)	(3)	(4)
Treatment	0.0584	0.270*	0.0854+	0.283*
	(0.0537)	(0.119)	(0.0494)	(0.130)
High SES		0.236**		0.244**
		(0.0872)		(0.0902)
Treatment × High SES		-0.248*		-0.240+
		(0.107)		(0.122)
High SES (m)		0.0982		0.0826
		(0.103)		(0.107)
Treatment × High SES (m)		-0.377		-0.290
		(0.255)		(0.266)
Constant	0.146**	-0.0476	0.164+	-0.0492
	(0.0472)	(0.107)	(0.0926)	(0.133)
Observations	829	829	787	787
Sixth grade reading test	+	+	+	+
Baseline reading scores	+	+	+	+
Strata fixed effects	-	-	+	+
Adjusted R-squared	0.72	0.73	0.74	0.75

Note: Estimated coefficients based on OLS regressions with the endline reading score as the dependent variable. Specifications with sixth grade reading test include an indicator for whether or not the endline test taken was the fourth grade or the sixth grade test.
Specification with the baseline reading score includes the fourth grade reading score and an indicator for whether or not the score is missing.
Specification with strata fixed effects include an indicator for each strata minus one.
Standard errors clustered at the school level in parentheses. + p<0.1, * p<0.05, ** p<0.01.

Table A5: Effect of treatment on reading scores. Only students with valid post inaccuracy measure

	(1)	(2)
Treatment	0.0280	0.204*
	(0.0475)	(0.0917)
High SES		0.235**
		(0.0543)
Treatment × High SES		-0.235*
		(0.0889)
High SES (m)		0.102
		(0.0785)
Treatment × High SES (m)		-0.165
		(0.186)
Constant	0.196**	0.0234
	(0.0366)	(0.0662)
Observations	1,241	1,241
Sixth grade reading test	+	+
Baseline reading scores	+	+
Missing post inaccuracy	Excluded	Excluded
Adjusted R-squared	0.71	0.71

Note: Estimated coefficients based on OLS regressions with the endline reading score as the dependent variable. Specifications with sixth grade reading test include an indicator for whether or not the endline test taken was the fourth grade or the sixth grade test. Specification with the baseline reading score includes the fourth grade reading score and an indicator for whether or not the score is missing. Standard errors clustered at the school level in parentheses. + p<0.1, * p<0.05, ** p<0.01.

Table A6: Effect of treatment on reading test scores. Interaction with different measures of SES.

	(1)	(2)	(3)	(4)
Treatment	0.0226	0.184*	0.221**	0.227**
	(0.0435)	(0.0814)	(0.0732)	(0.0817)
High SES		0.173**	0.249**	0.265**
		(0.0557)	(0.0424)	(0.0447)
Treatment × High SES		-0.200*	-0.255**	-0.263**
		(0.0790)	(0.0734)	(0.0788)
High SES (m)		-0.00608	-0.301+	0.0277
		(0.0866)	(0.177)	(0.0957)
Treatment × High SES (m)		-0.266	-0.770*	-0.323
		(0.172)	(0.305)	(0.277)
Constant	0.192**	0.0730	0.0156	-0.000931
	(0.0338)	(0.0638)	(0.0548)	(0.0570)
Observations	1,518	1,518	1,518	1,518
Sixth grade reading test	+	+	+	+
Baseline reading score	+	+	+	+
Measure of SES	-	Earning	One valid parent	Two valid parents (main measure)
Adjusted R-squared	0.69	0.69	0.70	0.70

Note: Estimated coefficients based on OLS regressions with the endline reading score as the dependent variable. Specifications with sixth grade reading test include an indicator for whether or not the endline test taken was the fourth grade or the sixth grade test.
Specification with the baseline reading score includes the fourth grade reading score and an indicator for whether or not the score is missing.
Standard errors clustered at the school level in parentheses.
+ p<0.1, * p<0.05, ** p<0.01.

Table A7: Ordered Logit Model. Intervention Effects on Gap between Teacher Beliefs and Test Scores

	(1)	(2)	(3)	(4)
Treatment	-0.308*	-0.255+	-0.550*	-0.614*
	(0.134)	(0.139)	(0.268)	(0.271)
High-SES			-0.0956	-0.0621
			(0.115)	(0.122)
Treatment × High-SES			0.296	0.447+
			(0.257)	(0.267)
High-SES (m)			-0.223	-0.195
			(0.319)	(0.362)
Treatment × High-SES (m)			0.608	0.706
			(0.650)	(0.792)
Cut 1	0.254**	0.839**	0.176+	0.790**
	(0.0754)	(0.103)	(0.105)	(0.127)
Cut 2	2.706**	3.532**	2.629**	3.487**
	(0.147)	(0.180)	(0.155)	(0.179)
Cut 3	5.179**	6.069**	5.102**	6.024**
	(0.394)	(0.416)	(0.409)	(0.425)
Observations	1,733	1,733	1,733	1,733
Baseline test score-belief gap	-	+	-	+

Note: Estimated coefficients based on ordered logit regressions with the endline test score-belief gapscore as the dependent variable. Specification with the baseline test score-belief gap includes the baseline test score-belief gap and an indicator for whether or not the score is missing. Standard errors clustered at the school level in parentheses. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$.

B Survey questions

Survey question for measuring the teacher's perceptions of the student's reading abilities

"As a way to follow the students' development through the trial, we ask you to evaluate the reading skills of each student in your classroom. You are therefore asked to indicate whether the student's reading abilities are (1) certainly below the mean, (2) below the mean, (3) about the mean, (4) above the mean, (5) certainly above the mean."

Response categories:

- Certainly below the mean
- Below the mean
- About the Mean
- Above the mean
- Certainly above the mean
- Student no longer in classroom
- Don't know