

# Instruction Time, Information, and Student Achievement. Evidence from a Field Experiment\*

Simon Calmar Andersen  
Department of Political Science  
[sca@ps.au.dk](mailto:sca@ps.au.dk)

Thorbjørn Sejr Guul  
Department of Political Science  
[tsg@ps.au.dk](mailto:tsg@ps.au.dk)

Maria Knoth Humlum  
Department of Economics and Business Economics  
[mhumlum@econ.au.dk](mailto:mhumlum@econ.au.dk)

TrygFonden's Centre for Child Research  
Aarhus University

## Abstract

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

*Keywords:* Information; Learning; Field experiment.

*JEL:* I24; I28; D83.

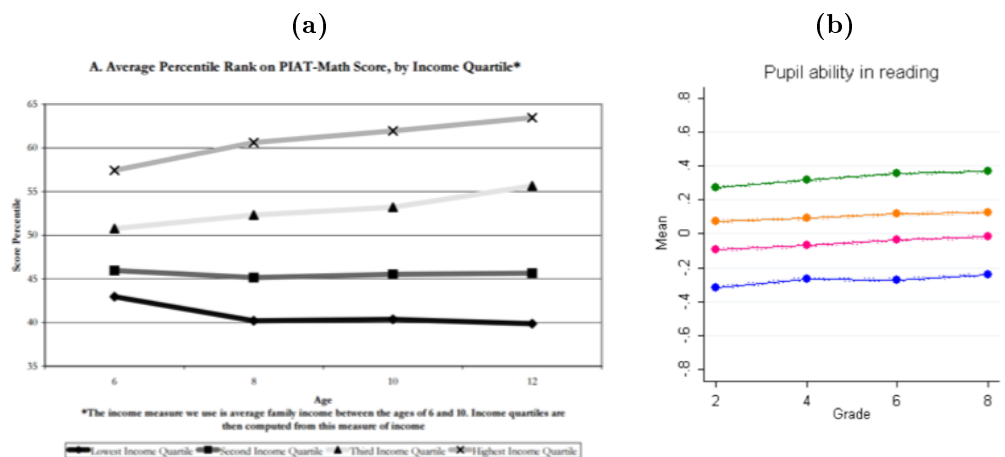
---

\*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 IWAAE 2018 and SDU Applied Microeconomics workshop. 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 socio-economic 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, as shown in Figure 1 these achievement gaps do not close—if anything they increase in both the US (Carneiro and Heckman, 2003) and in Denmark (Nandrup and Beuchert-Pedersen, 2018). Even when comparing the achievement gaps for U.S. birth cohorts 1954-2001, the SES achievement gaps are remarkably stable (Hanushek et al., 2019). Consistently, differential impacts of instruction time for high and low SES students was 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.

**Figure 1:** Test results by parental income quartile in the US (a) and Denmark (b)



Note: Figure 1.A from Carneiro and Heckman (2003, figure 8.A). Figure 1.B is from Nandrup and Beuchert-Pedersen (2018, figure 5.3 A).

We study this question using a randomized controlled trial that we designed in collaboration with the Ministry of Education in Denmark as a follow-up to a instruction time trial (Andersen et al., 2016). The new trial that we present in this paper combined increased instruction time (a teaching program of two lessons per week in 16 weeks in fifth grade) with

information to teachers about their students' performance in monthly reading tests. Existing research show that instruction is more effective, if it is adapted to students' current skill level (Banerjee et al., 2007, 2017; Duflo et al., 2011). However, if teachers have inaccurate beliefs about students' skill levels, the teachers cannot adapt their teaching adequately. Therefore, providing teachers with information about students' progress may help teachers to target instruction to students' skill level.

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. Whether this notion is correct, is not clear from existing evidence, but there are some indications. 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 nationwide 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 would suggest that providing teachers with updated information on the skills of their students 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.

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. Analyses of potential mechanisms first show that in the control group inaccurate beliefs were more strongly correlated with reading achievement among low-SES than high-SES students, which is consistent with the notion that high-SES students are less dependent on teacher beliefs.

We also find that effects on reading achievement were larger for the low-SES students with a teacher belief-test score gap at baseline. Furthermore, we find that the intervention reduced behavioral problems among the low-SES students. This indicates that the monthly tests did not stress the students. Finally, we examine what factors explain 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. Accumulating evidence demonstrates that providing parents with information about their children's school performance reduce the gap between test scores and parents' beliefs about their children's skills, and that this information makes parents adjust their investment decisions – especially 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 about the educational production function affect parents' decisions about time invested in their children (Cunha et al., 2013). Teachers in our study had access to the baseline test results. Nevertheless, their beliefs about students' reading achievement diverge from an objective performance measure at baseline. This suggests that not just low-SES students' parents but also their teachers are prone to information friction.

We also contribute to literature on performance information in schools. Rockoff and colleagues (2012) randomly assign school principals to objective estimates of teacher performance. They find that 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 when the principals have less precise beliefs prior to the intervention. The results are coherent with a Bayesian learning model where the principals base their beliefs on priors 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 level of the students. This seems to be of particular importance to low-SES students who may be more dependent on adequate instruction in schools. 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 – the testing of students and information provided to teachers – improve teachers’ perceptions 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](#)) finding that testing 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 students’ academic performance ([Gerritsen et al., 2017](#); [Staiger and Rockoff, 2010](#)). Aside from this, the specific aspects of teacher quality or specific characteristics of teachers that impact on student achievement are not well-known ([Hanushek, 2011](#)). We demonstrate that older and thereby more experienced teachers have more accurate beliefs about the reading skills of their students. These results suggest that the ability to correctly perceive the skill level of students in the classroom may be one important component in what has been called ‘teacher quality’, but 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. Finally, section [5](#) concludes.

## 2 Background and Experimental Design

### 2.1 Background

At the time 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 parents' socio-economic status and 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 students' skills in different subjects. Most tests were in reading. The ten mandatory tests were spread out between grade 2 and grade 8. Reading was tested in grades 2, 4, 6, and 8 (for an extensive overview of the Danish national tests, see Nandrup and Beuchert-Pedersen, 2018). Besides the mandatory tests, teachers could decide to use different tests of their choosing, and the municipalities could mandate the use of specific tests. Municipalities govern schools in Denmark (comparable to school districts in the US). Despite these efforts to improve the evaluation culture, use of 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 socio-economic 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 school receives a voucher of about 75 percent of the average costs of public school students. 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).

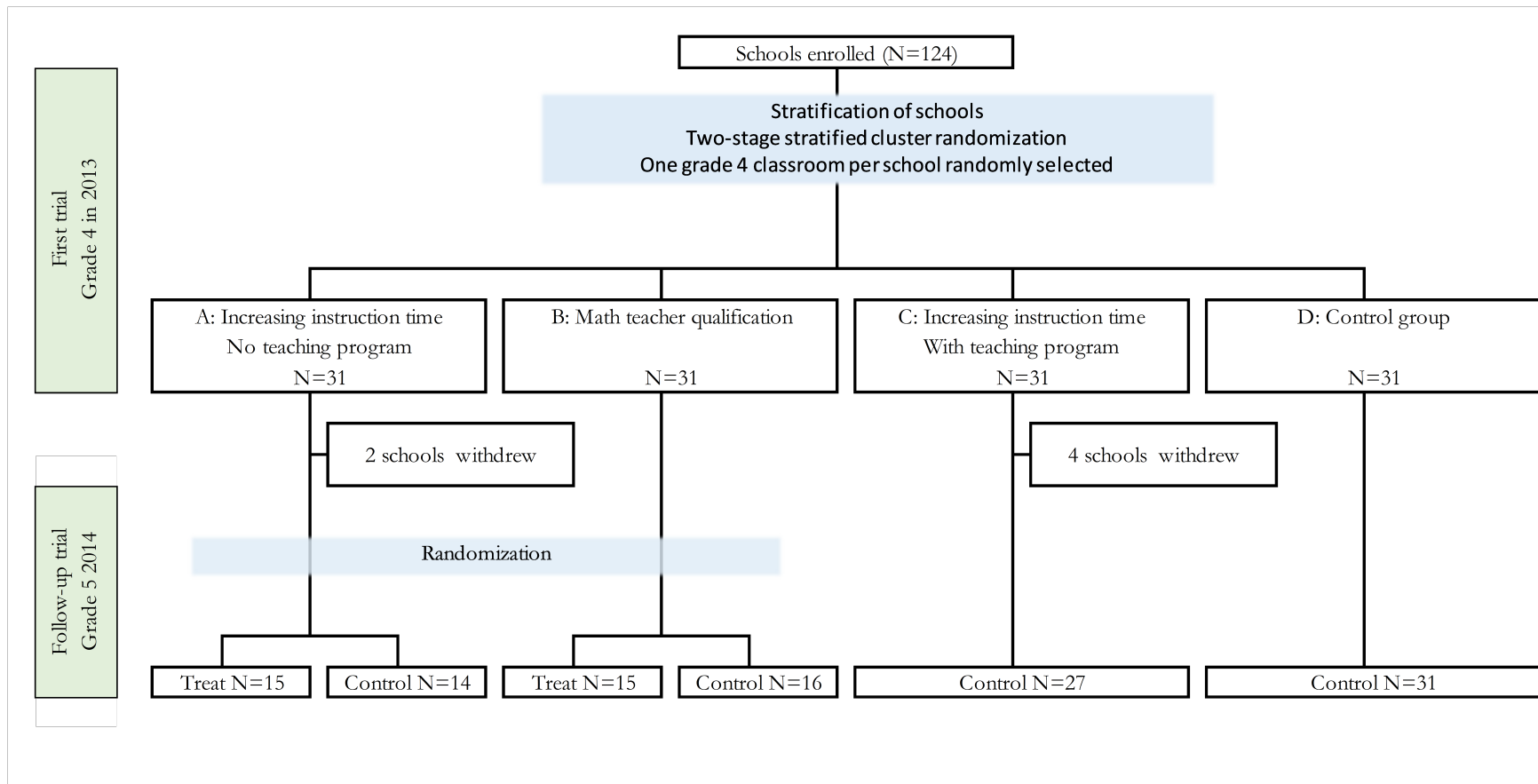
## 2.2 First Trial

In the autumn of 2013 the Danish Ministry of Education funded the implementation of a large field experiment in grade 4 in Danish primary schools. The results of this trial is reported by [Andersen et al. \(2016\)](#). The randomized trial that we report here was a follow-up to the first trial.<sup>1</sup> 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 2](#) shows the overall design and relationship between the two rounds of experiments.

---

<sup>1</sup>Only one grade 4 classroom per school participated in the first round. In the second round, one grade 4 classroom and one grade 5 classroom per school participated. Initial analyses documented large and skewed attrition along with noncompliance among the grade 4 classrooms (less than 60 pct. of the classrooms complied with the assigned treatment). The analyses in this paper focus on the grade 5 classrooms.

**Figure 2:** 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 that are generally of low SES (roughly 50 percent of the bilingual students compared to 15 pct. of the remaining students in our sample), the only inclusion criterion was that schools should expect to have at least 10 pct. bilingual students in grade 4 in the fall of 2013. 126 schools initially enrolled in the trial. Two schools with less than 10 pct. 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 grade 4 classroom was randomly selected for participation.

Schools were allocated to either one of three treatment groups or the control group.<sup>2</sup> 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 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 for students of non-Western origin.<sup>3</sup> The point estimates for this group were statistically insignificant and close to zero. These results motivated us to adjust the treatment in the follow-up trial in an attempt to target the more disadvantaged students.

## 2.3 The Follow-up Trial

The follow-up trial was implemented in the autumn of 2014. The 124 schools that participated in the first-round experiment were requested to participate in the second round as well. Only six schools withdrew from the second-round experiment prior to randomization

---

<sup>2</sup>Allocation to treatment and control groups was based on two-stage stratified cluster randomization. Schools were stratified according to the share of non-Western students and grade 2 reading test scores. One grade 4 classroom per school was selected for participation based on simple randomization.

<sup>3</sup>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.

(see Figure 2).

Certain restrictions on the design, which we explain below, meant that only groups A and B (the two left-hand treatment arms in Figure 2) were randomized to either treatment or 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. To increase statistical power, we include schools from groups C and D (the two right-hand arms in Figure 2) in the main analyses. 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 2) would receive the 4th grade treatment in the following cohort of 4th grade students. Therefore, the 5th grade students in these schools (group D), were maintained in the control group in the second-round experiment.<sup>4</sup> Schools in group C were all placed in the control group in the second-round experiment because the teaching program in the first-round randomization was similar to the program tested in the second-round trial (see the next section for further information on the second-round intervention)

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 grade 2 in 2013. One classroom per school had been randomly selected for participation in the first round trial and this classroom also participated in the follow-up trial.

The Danish Ministry of Education reimbursed the participating schools for all costs associated with participation in 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.

---

<sup>4</sup>The grade 4 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.

## 2.4 The Intervention

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

- i. A teaching program including two lessons per week (additional instruction time)
- ii. Regular testing including feedback to the teacher about student performance

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 teaching program was taught during two extra (45-minute) lessons per week.<sup>5</sup> The duration of the intervention was 16 weeks. In addition, teachers were required to implement one lesson per week as part of the regular Danish classes (with an additional teacher) and one lesson per week targeted at students with special needs and scheduled in periods during the week where the students do not have formal classes.

The first component of treatment (instruction time and treatment program) was very similar to the treatments in the first trial. As mentioned, these treatments were not found to improve the academic performance of 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 students’ performance would help teachers target the instruction towards the more disadvantaged students.

---

<sup>5</sup>In the first-round experiment, the interventions included four additional lessons in Danish. This was not feasible in the second-round experiment since a major school reform was implemented in the beginning of the school year 2014/15 that increased instruction time substantially. The treatment and control group schools were all subject to the reform.

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.<sup>6</sup> In addition, the students answered two questions about the explicit and implicit meaning of the text, respectively. The Danish teacher received a summary of his students' performance (for every test) and progression since the previous test (except for the first test). In addition, they received a student report consisting of scorecards with individual results for each student. Figure 3 shows an example of a class overview and a student scorecard. Though the test is rather simplistic, the test results explain 40 pct. of the variation in 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. For more details on the components of the treatment, see Appendix C.

## 2.5 Timeline of Experiment and Data Collection

The follow-up trial was conducted during the autumn of 2014. The intervention period was 16 weeks and an extensive data collection was undertaken in connection with the experiment. To the extent possible, the experiment was designed so that evaluation could be based on administrative data. Figure 4 shows a timeline of the experiment and the associated data collection.

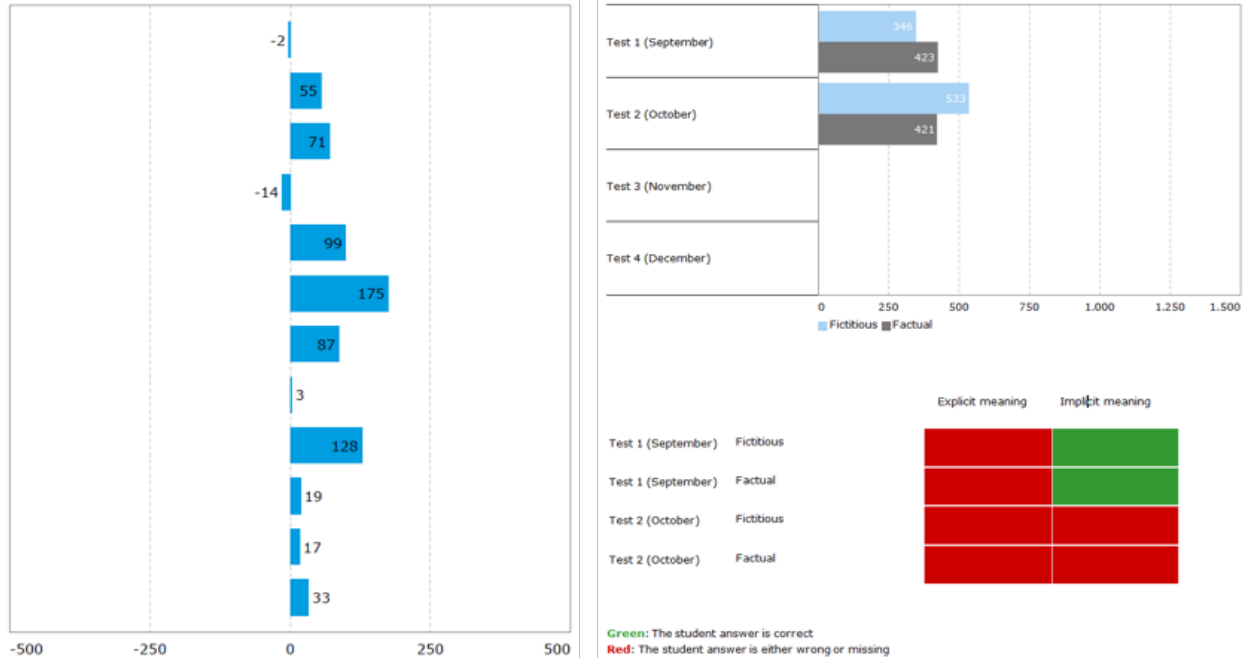
The randomization for the follow-up trial took place in March 2014 and schools were subsequently informed about the treatment status of their participating grade 5 classroom. The actual intervention was implemented from September 1st to December 13th<sup>7</sup> Surveys were conducted before and after the intervention with both students, teachers and principals.

---

<sup>6</sup>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 text difficulty in 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.

<sup>7</sup>In Denmark, the school year starts in August.

**Figure 3:** Example of class overview of progression and student score card



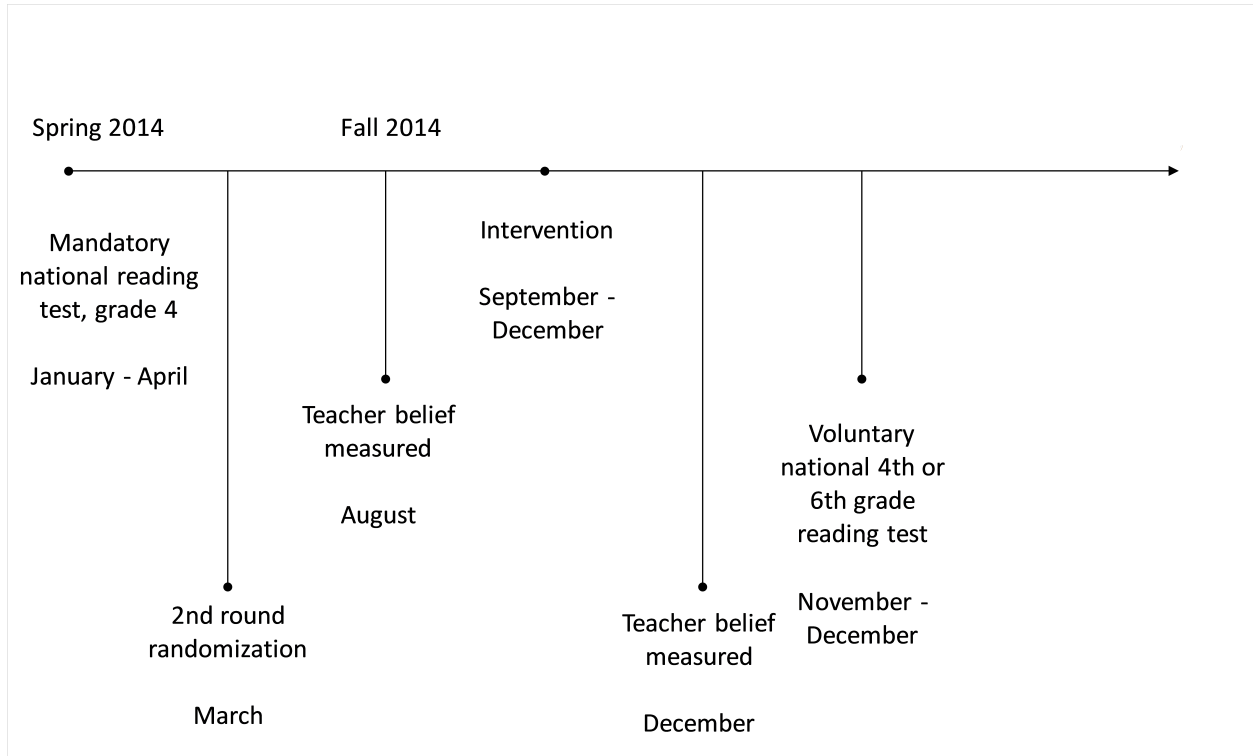
Two implementation surveys with teachers were conducted during the intervention to assess fidelity.<sup>8</sup> The main outcome measure is academic performance as measured by the national reading test. In the period January to April 2014, the participating students have taken the mandatory grade 4 reading 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 using unique individual identifiers. The detailed administrative data provide information about national test scores, gender, country of origin, socioeconomic background and other relevant background

<sup>8</sup>The implementation surveys were given to the Danish teachers and were used to assess the degree to which teachers and schools complied with the intervention. All surveys were administered electronically.

**Figure 4:** Timeline of experiment and data collection



variables. The student surveys were mainly used to evaluate the effect on student 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 about 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 grade 4 national reading test conducted in the period from January to April 2014 prior to the intervention (see Figure 4). The Danish national tests are standardized computer-scored tests.

At the end of the intervention period, students took either the grade 4 or the grade 6

national reading test (see Figure 4).<sup>9</sup> They were encouraged to take the grade 4 reading test, but a smaller fraction (roughly 20 percent) of the students in our sample took the grade 6 reading test. Taking the grade 6 test does not correlate significantly with treatment status. Since we expect the grade 6 test to have a higher level of difficulty, we account for this by including an indicator for whether a student took the grade 4 or the grade 6 reading test when we estimate intervention effects on reading performance.

Each test consists of three subtest scores. We follow [Nandrup and Beuchert-Pedersen \(2018\)](#) and standardize 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.

### 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 teachers' beliefs about their students' reading abilities in the teacher surveys conducted both before and after the intervention. In the survey, teachers were asked to rank their students' reading abilities according to five categories from 'certainly below average' to 'certainly above average'. 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 in each of the five categories.<sup>10</sup>

The objective measure of the students' reading abilities (or students' actual reading performance) is based on the students' scores on the mandatory grade 4 national reading

---

<sup>9</sup>These tests were voluntary tests corresponding to the mandatory national tests in reading for grade 4 and grade 6.

<sup>10</sup>See Appendix B for the exact wording of the question.

test—our measure of baseline reading performance. The test score is based on a computer algorithm and thereby blinded to teachers’ prior beliefs, students’ experimental condition in the trial and other factors that might conflate the objective measure with the trial. When students take the national tests, 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 result. In other words, there may be some level of information friction in this system.

Based on the baseline reading test scores, 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 pct. of the students in the category certainly above average, the best 10 pct. of the students according to the reading test were also placed in that category.

In line with [Dizon-Ross \(2019\)](#), we measure inaccuracies in teachers’ beliefs as the absolute value of the gap between the teachers’ rank of student  $i$  and the test score based rank:<sup>11</sup>

$$Gap_i = |TestRank_i - TeacherRank_i| \quad (1)$$

One obvious question is whether the gaps reflect low reliability in the test score measure or low measurement validity in the sense that teachers evaluate other dimensions of students’ reading skills than what the test scores capture. We have different indications that this is not the case. First, the tests explain roughly 50 pct. of the variation in the corresponding grade 9 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

---

<sup>11</sup>If we think of the teacher assessment as a forecast of the true performance, the teacher-test divergence corresponds to the absolute forecast error. We focus on the absolute forecast error, since we do not find any evidence that teachers’ forecasts are biased. Strictly speaking, the teachers’ beliefs cannot be forecasts since they already know (or have access to) the true performance of the students.



direction of the objective test scores, it suggests that they believe that their prior beliefs were inaccurate.

### 3.2 Socioeconomic Status (SES)

In order to measure whether the assessment varies with the socioeconomic status of the students, we use administrative data on parents' highest level of education. In Denmark large redistribution policies ensures 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 were missing in the administrative data, we set the value to missing.<sup>12</sup> If at least one of the parents had a college degree (vocational or academic), we coded the student as high-SES. In our robustness analysis, we construct an alternative SES measure based on parental income.<sup>13</sup> All main results are similar when we base the SES measure on parental income rather than education. When we compare intervention effects between students of Western and 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 roughly 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 teachers beliefs, information and instruction time.

### 3.3 Balance and Attrition

Table 1 shows that the randomization at the outset 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

---

<sup>12</sup>Using the information for students with one parent missing produces similar results (see Table A6 in Appendix A).

<sup>13</sup>We divide the students into quartiles based on the parent with the highest income and define low-SES as students in the lowest quartile in the sample of grade 5 students. This is similar to the procedure used for the whole population to produce Figure 1b in the introduction.

disadvantaged.<sup>14</sup> Correspondingly, 26 pct. of participating students are of non-Western origin which is substantially above the average in Danish public schools. About 20 pct. of students have low SES. The differences between the control and treatment group are all small and statistically insignificant. Based on a joint F-test of the null of no differences between the treatment and control group, we cannot reject the null.

**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 testscore-belief gap	0.47	0.41	0.07
Baseline testscore-belief gap (m)	0.31	0.20	0.11
Observations	1,927	619	

Notes: The table reports means in the control and treatment groups and the corresponding difference. Low SES, Female and Non-Western origin are indicator variables. (m) denotes missing variable indicators.

No differences significant at the 10%-level.

Despite a little attrition, our estimation sample is also balanced between treatment and control group. Out of the 118 schools assigned to either the treatment or control group, 13 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 the control group (10-11 pct.). In addition for 19 schools, we did not observe the main outcome (reading scores). This attrition was not evenly distributed across treatment (3.3 pct. of schools) and control groups (18.2 pct. of schools). However, Appendix Table A1 shows that differences between treatment and

<sup>14</sup> Andersen et al. (2016) show that the distribution of reading scores for participating schools is shifted to the left compared to the distribution for all public schools.

control group are small and, except for one covariate, statistically insignificant at the 10 pct. level in the estimation sample (the sample for which we observe the reading score). The main estimation sample consists of 1,518 students from 86 schools in total.

## 4 Results

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

### 4.1 Intervention Effects

We use the following basic model to estimate the intervention effect on 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 teacher  $j$ 's classroom,  $Treatment_j$  is an indicator for whether or not classroom  $j$  was in the treatment group,  $Prescore$  is the grade 4 reading score (the baseline reading score),  $I[Grade6]$  is an indicator for whether or not the student took the grade 6 (instead of grade 4) endline reading test, and  $u$  is the idiosyncratic error term.<sup>15</sup> We also include an indicator for missing baseline reading score. We interpret  $\gamma_1$  as the intention-to-treat effect of the intervention.<sup>16</sup> The baseline test score is included to improve precision in the estimation. In order to estimate the intervention effects

---

<sup>15</sup>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 second round randomization. Appendix Table A4 shows the main results when the sample is restricted to schools that were randomized in the second round and with and without stratum fixed effects. The main results are robust to the inclusion of stratum fixed effects.

<sup>16</sup>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 where the students do not have formal classes).

moderated by SES, most of our specifications include an indicator for 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 run regression models with and without covariates. In general, estimates are robust and, if anything, become more statistically significant when adding covariates, as would be expected based on the random assignment.

#### 4.1.1 Intervention Effects on Students' Reading Test scores

Table 2 shows the intervention effects on students' reading test scores. The average intervention effect is .06 (model 1) and a little smaller when including covariates in model 2, but in both cases 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 pct. 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.

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 as mentioned, the first trial determined the treatment status of some classrooms. Excluding the classrooms only randomized in the first trial (Groups C and D, i.e. the two right-hand arms in Figure 2) does not change the main results on reading

**Table 2:** Intervention effects on reading test scores

	(1)	(2)	(3)	(4)
Treatment	0.0457 (0.0929)	0.0226 (0.0435)	0.180 (0.143)	0.227** (0.0817)
High SES			0.700** (0.0722)	0.265** (0.0447)
Treatment $\times$ High SES			-0.184 (0.147)	-0.263** (0.0788)
High SES (m)			-0.0520 (0.149)	0.0277 (0.0957)
Treatment $\times$ High SES (m)			-0.300 (0.530)	-0.323 (0.277)
Constant	0.0786 (0.0583)	0.192** (0.0338)	-0.418** (0.0815)	-0.000931 (0.0570)
Observations	1,518	1,518	1,518	1,518
6th grade reading test	+	+	+	+
Baseline reading score	-	+	-	+
Adjusted R-squared	0.15	0.69	0.23	0.70

Notes: Estimated coefficients based on OLS regressions with the endline reading score as the dependent variable. Specifications with 6th grade reading test include an indicator for whether or not the endline test taken was the grade 4 or grade 6 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 school level in parentheses. +  $p < 0.1$ , \*  $p < 0.05$ , \*\*  $p < 0.01$ .

scores (see Appendix Table A4). This shows that the first trial does not account for these results.

Second, as a further robustness check we conducted the analysis with an alternative measure 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 grade 5 students in the study. Appendix Table A6 also shows results similar to the main results presented in Table 2 with this alternative measure of SES. 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 while 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.

Third, we exclude students that did not have a valid endline measure of the teacher-test gap. Appendix Table A3 shows that the treatment and control group are also well-balanced in this reduced sample. Appendix Table A5 shows that this does not change the main results. Again, this supports the robustness of our findings.

#### **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 wellbeing. 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' wellbeing on average. On the contrary, the treatment decreased the SDQ score for low-SES students – and closed the control group gap in wellbeing between high- and low-SES students – while it remained roughly 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 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?**

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 teachers' beliefs. The results in Table 4 show that the intervention improved the accuracy of 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 teacher-test gap

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

	(1)	(2)	(3)	(4)
Treatment	-0.317 (0.352)	-0.284 (0.327)	-1.470 <sup>+</sup> (0.747)	-1.532* (0.678)
High SES			-1.821** (0.419)	-1.079** (0.408)
Treatment × High SES			1.604 <sup>+</sup> (0.824)	1.726* (0.759)
High SES (m)			0.600 (0.901)	0.236 (0.883)
Treatment × High SES (m)			-1.008 (1.676)	-0.786 (1.552)
Constant	9.711** (0.208)	9.546** (0.198)	11.04** (0.372)	10.35** (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

Notes: Estimated coefficients based on OLS regressions with the endline SDQ score as the dependent variable. 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.

(Model 2).<sup>17</sup> In Models 3 and 4, we allow the intervention effect to vary with SES status. For low-SES students, the intervention effect is negative and statistically significant at the 5 pct. level when baseline teacher-test divergence is included. The interaction term is positive; but statistically insignificant. Thus, the intervention improved the accuracy of teachers' beliefs and the direction of the effects suggest that the intervention was more effective at improving the accuracy of teachers' beliefs for low-SES students than for high-SES students.<sup>18</sup> 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 the explanation for why the information intervention was more effective for this group of

<sup>17</sup>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 group still balance well, see Table A2.

<sup>18</sup>The measure of the teacher-test gap is technically an ordered response variable. An ordered logit analysis produces substantially similar results, cf. Appendix Table A7

students.

**Table 4:** Intervention Effects on Gap between Teacher Beliefs and Test Scores

	(1)	(2)	(3)	(4)
Treatment	-0.0865*	-0.0656 <sup>+</sup>	-0.149 <sup>+</sup>	-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 testscore-belief gap	-	+	-	+
Adjusted R-squared	0.00	0.12	0.00	0.12

Notes: Estimated coefficients based on OLS regressions with the endline testscore-belief gap score as the dependent variable. Specification with the baseline testscore-belief gap includes the baseline testscore-belief gap and an indicator for whether or not the score is missing. Standard errors clustered at the school level in parentheses. <sup>+</sup> 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 perceptions at baseline were more strongly correlated with 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 6 support the notion that accurate teacher beliefs (and hence more appropriate teaching level) is more important



to low-SES students, and that combining instruction time with regular student testing is especially important 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** (0.0445)	0.411** (0.0570)
No testscore-belief gap	0.0711 (0.0451)	0.265** (0.0726)
High SES × No testscore-belief gap		-0.259** (0.0685)
High SES (m)	0.0145 (0.0940)	-0.00311 (0.0907)
No testscore-belief gap (m)	0.0736 (0.0792)	0.0863 (0.0778)
Constant	-0.0581 (0.0778)	-0.151 <sup>+</sup> (0.0843)
Observations	1086	1086
Baseline reading score	+	+
6th grade reading test	+	+
Adjusted R-squared	0.69	0.69

Notes: Estimated coefficients based on OLS regressions with the endline reading score as the dependent variable. Specifications with 6th grade reading test include an indicator for whether or not the endline test taken was the grade 4 or grade 6 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.

### 4.2.3 Was the Intervention More Effective for Students that Were Misperceived by their Teacher at Baseline?

If the inaccuracy of teachers’ beliefs partly explains why low-SES students have lower reading skills, we would expect that the low-SES students with high levels of baseline teacher-test gap would benefit the most from teachers receiving more information. To test this, we examine whether the intervention effect is higher for students with inaccurate teacher baseline beliefs about their performance. Model 1 in Table 6 shows that the intervention tended to be

more effective for students whose teachers' had inaccurate beliefs about their reading skills at baseline, even though the difference was not statistically significant. In models 2-3 we split the students on SES. Model 2 shows that the intervention effect is substantially larger for low-SES students for whom teachers' beliefs were inaccurate. The effect is statistically significant at the 1 pct. level. The interaction effect is correspondingly large, negative and statistically significant at the 5 pct. level suggesting that the overall positive intervention effects for low-SES students are driven by the group of students for whom teachers had inaccurate beliefs about their reading abilities at baseline. Model 3 confirms that for high-SES students teachers' inaccurate baseline beliefs did not seem to moderate the impact of the intervention.

Since these results indicate that teacher beliefs are especially important to low-SES students and that the intervention improved the accuracy of teachers' beliefs and was most effective for low-SES students that were misperceived by their teacher at baseline, it becomes relevant to examine what factors predict the accuracy of teachers' beliefs about 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 teachers' beliefs about 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 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 overall tendency using a survey-based and less fine-grained measure of experience. As mentioned, teacher experience remains one of the teacher characteristics with most empirical

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

	All (1)	Low SES (2)	High SES (3)
Treatment	0.0564 (0.0646)	0.430** (0.127)	-0.0541 (0.0640)
No testscore-belief gap	0.0660 (0.0434)	0.278* (0.107)	0.00233 (0.0428)
Treatment $\times$ No testscore-belief gap	-0.0140 (0.0675)	-0.340* (0.166)	0.0809 (0.0685)
No testscore-belief gap (m)	0.0479 (0.0856)	0.120 (0.151)	0.0312 (0.0661)
Treatment $\times$ No testscore-belief gap (m)	-0.437* (0.184)	-0.327 (0.289)	-0.381* (0.154)
Constant	0.150** (0.0535)	-0.164 (0.115)	0.252** (0.0436)
Observations	1518	340	1119
6th grade reading test	+	+	+
Baseline reading score	+	+	+
Adjusted R-squared	0.69	0.63	0.70

Notes: Estimated coefficients based on OLS regressions with the endline reading score as the dependent variable. Specifications with 6th grade reading test include an indicator for whether or not the endline test taken was the grade 4 or grade 6 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.  
+  $p < 0.1$ , \*  $p < 0.05$ , \*\*  $p < 0.01$ .

support (Gerritsen et al., 2017; Staiger and Rockoff, 2010). As expected, more experienced teachers have more accurate beliefs about students' reading skills. This suggests that inaccurate beliefs about student performance could be one of the channels through which teacher experience affects student performance.

## 5 Conclusion and discussion

At baseline of this study teachers had access to information about students test scores in computer-based, adaptive reading tests. Nevertheless, their beliefs about the reading

**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 (0.108)	-0.101 (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** (0.0411)	0.474** (0.0278)	0.440** (0.0244)	0.439** (0.0637)	0.652** (0.106)	0.514** (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

Notes: A cell with less than four observations is omitted to comply with Statistics Denmark data security policy. Standard errors clustered at the school level in parentheses. +  $p < 0.1$ , \*  $p < 0.05$ , \*\*  $p < 0.01$ .

skills of their students differed somewhat from the objective test results. Consistent with a simple model of information frictions and Bayesian learning, teachers in the treatment group updated their beliefs so that at 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 teacher beliefs is a more important factor for explaining the reading scores of low-SES students than the reading scores of high-SES students. We do not have data to test why that is, but one natural explanation may be that student learning outside of school is more effective for high-SES students and that parents may compensate more at home, if 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 that combined extra instruction time with regular testing and readily accessible information to 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 teacher's beliefs about low-SES students' performance and ultimately improves the reading skills of low-SES students. We also find that the treatment reduces behavioral problems of low-SES students, thereby reducing concerns that more instruction time will come at the expense of student wellbeing in school.

As such, our results are very much in line with the findings of [Dizon-Ross \(2019\)](#) based on a field experiment in Malawi. She finds that parents' baseline beliefs are inaccurate and that providing 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, teachers appear to learn quite a lot from these tests, without risking any detrimental effects of gaming and strategic behaviour. 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 that we use as an objective measure of student abilities, but they still held different beliefs about 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 about student abilities 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.
- 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. page 61.
- 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. page 63.

- 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.
- 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 inter-generational mobility in denmark and the us. *The Scandinavian Journal of Economics*, 119(1):178–230.
- 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. *Nationaløkonomisk Tidsskrift*, 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.
- 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.

# Appendices

## A Robustness 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	

Notes: The table reports means in the control and treatment groups and the corresponding difference. Low SES, Female and Non-Western origin are indicator variables. A cell with less than four observations is omitted to comply with Statistics Denmark 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 misperception 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	

Notes: The table reports means in the control and treatment groups and the corresponding difference. Low SES, Female and Non-Western origin are indicator variables. A cell with less than four observations is omitted to comply with Statistics Denmark 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 misperception 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 control and treatment groups and the corresponding difference. Low SES, Female and Non-Western origin are indicator variables. A cell with less than four observations is omitted to comply with Statistics Denmark 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 (0.0537)	0.270* (0.119)	0.0854+ (0.0494)	0.283* (0.130)
High SES		0.236** (0.0872)		0.244** (0.0902)
Treatment $\times$ High SES		-0.248* (0.107)		-0.240+ (0.122)
High SES (m)		0.0982 (0.103)		0.0826 (0.107)
Treatment $\times$ High SES (m)		-0.377 (0.255)		-0.290 (0.266)
Constant	0.146** (0.0472)	-0.0476 (0.107)	0.164+ (0.0926)	-0.0492 (0.133)
Observations	829	829	787	787
6th grade reading test	+	+	+	+
Baseline reading scores	+	+	+	+
Strata fixed effects	-	-	+	+
Adjusted R-squared	0.72	0.73	0.74	0.75

Notes: Estimated coefficients based on OLS regressions with the endline reading score as the dependent variable. Specifications with 6th grade reading test include an indicator for whether or not the endline test taken was the grade 4 or grade 6 test.

Specification with the baseline reading score includes the grade 4 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 teacher-test gap

	(1)	(2)
Treatment	0.0280 (0.0475)	0.204* (0.0917)
High SES		0.235** (0.0543)
Treatment $\times$ High SES		-0.235* (0.0889)
High SES (m)		0.102 (0.0785)
Treatment $\times$ High SES (m)		-0.165 (0.186)
Constant	0.196** (0.0366)	0.0234 (0.0662)
Observations	1,241	1,241
6th grade reading test	+	+
Baseline reading scores	+	+
Missing post test score-belief gap	Excluded	Excluded
Adjusted R-squared	0.71	0.71

Notes: Estimated coefficients based on OLS regressions with the endline reading score as the dependent variable. Specifications with 6th grade reading test include an indicator for whether or not the endline test taken was the grade 4 or grade 6 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. +  $p < 0.1$ , \*  $p < 0.05$ , \*\*  $p < 0.01$ .

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

	(1)	(2)	(3)	(4)
Treatment	0.0226 (0.0435)	0.184* (0.0814)	0.221** (0.0732)	0.227** (0.0817)
High SES		0.173** (0.0557)	0.249** (0.0424)	0.265** (0.0447)
Treatment × High SES		-0.200* (0.0790)	-0.255** (0.0734)	-0.263** (0.0788)
High SES (m)		-0.00608 (0.0866)	-0.301+ (0.177)	0.0277 (0.0957)
Treatment × High SES (m)		-0.266 (0.172)	-0.770* (0.305)	-0.323 (0.277)
Constant	0.192** (0.0338)	0.0730 (0.0638)	0.0156 (0.0548)	-0.000931 (0.0570)
Observations	1,518	1,518	1,518	1,518
6th 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

Notes: Estimated coefficients based on OLS regressions with the endline reading score as the dependent variable. Specifications with 6th grade reading test include an indicator for whether or not the endline test taken was the grade 4 or grade 6 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.

**Table A7:** 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 $\times$ High-SES			0.296	0.447+
			(0.257)	(0.267)
High-SES (m)			-0.223	-0.195
			(0.319)	(0.362)
Treatment $\times$ High-SES (m)			0.608	0.706
			(0.650)	(0.792)
cut1	0.254**	0.839**	0.176+	0.790**
	(0.0754)	(0.103)	(0.105)	(0.127)
cut2	2.706**	3.532**	2.629**	3.487**
	(0.147)	(0.180)	(0.155)	(0.179)
cut3	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 testscore-belief gap	-	+	-	+

Notes: Estimated coefficients based on ordered logit regressions with the endline testscore-belief gapscore as the dependent variable. Specification with the baseline testscore-belief gap includes the baseline testscore-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
- Mean
- Above the mean
- Certainly above the mean
- Student no longer in classroom
- Don't know

### **Survey question for assessing the teacher's familiarity with performance information use**

"How often do you use academic tests of the students in your instruction? By academic tests, we mean tests with the purpose of evaluating the academic level of the students"

- 0-4 times per year
- 5-8 times per year
- 9-12 times per year
- More than 12 times per year

## C Language comprehension course description

The participating classroom received additional instruction in language and language comprehension focused on patterns similar across languages. The teaching was carried out in accordance with a teaching material that was aimed at improving all students' reading skills, though it was intended to be particularly helpful for bilingual students who already had a knowledge about several languages. The material was detailed included among other things specific readings and exercises for each week in the program. The goal was to improve linguistic awareness academically better students but it had also a more general educational aspect. Important parts of the material was text of the week, illustration, linguistic focus, focus word and word cards, language wall, dialogue and assignments on both Danish and other languages.