Teachers' Intentions to Engage Their Students and Achievement – a Within-Student Between-Subject Approach Using TIMSS and PIRLS Data

Stephan Sievert[1], 17/02/2017

Abstract

I provide evidence on whether teachers' intentions to engage their students in learning have an effect on elementary school student achievement in Germany. I make use of a unique dataset that allows me to observe every student in math, science, and reading. My identification strategy relies on within-student between-subject estimation and a comprehensive set of teacher control variables. I find that teachers' intentions to engage their students has no effect on achievement in the full sample. However, it positively affects children from low socio-economic backgrounds. This finding is robust to a number of sensitivity checks and implies that equality of opportunity could be strengthened in Germany by more emphasis on classroom actions that engage students with what is being taught.

It is well-established that teachers are important inputs in educational production (see e.g. Hanushek & Rivkin, 2010; Hattie, 2009; Woessmann, 2003). However, economists have not been very successful at identifying what exactly makes teachers effective in conferring skills upon their students. Socio-economic characteristics such as gender, teaching experience, age, and education of the teachers cannot adequately account for the huge achievement differences attributable to different instructors (see e.g. Lavy, 2011). Much more important seems to be what teachers do and how they interact with their students (Hattie, 2009).

In recent years, the advent and expansion of large-scale assessment studies such as PISA, TIMSS, and PIRLS has led to a considerable increase in the economic literature on teacher effectiveness in general and teaching practices in particular. However, most of the studies on teaching practices have limited themselves to the dichotomy of traditional versus modern teaching (see e.g. Lavy, 2011; Bietenbeck, 2014). This paper is an attempt to go further than this and assess the impact of specific teaching practices on student achievement, namely the effectiveness of teachers' intentions to engage their students in learning. Student engagement is a well-known concept in the educational sciences, which can best be described as capturing the "in-the-moment cognitive interaction" of the student with what is being taught (McLaughlin et al., 2005). Whenever such interaction occurs, learning should happen at a faster rate than otherwise. A meta-analysis of teaching effectiveness studies by Seidel & Shavelson (2007) illustrates the relevance of the concept: In a seven-component model of learning, active student engagement falls under the heading "Basic information processing",[2] which is defined as the attempt to "cognitively engage students in the learning content" and "stimulating students' in-depth elaboration and organization of learning content (p. 470, Seidel & Shavelson, 2007). Importantly, it is focused on generalizable, not subject-specific factors. In the 29

[1] Freie Universität Berlin, Kaiserswerther Str. 16-18, 14195 Berlin.; Berlin Institute for Population and Development, Schillerstr. 59, 10627 Berlin, Germany; Email: stephan.sievert@fu-berlin.de

[2] Basic information processing itself is a dimension of "Execution of learning activities."

studies on basic information processing, which form part of the meta-analysis, the concept is found to be weakly positively related to cognitive and modestly so to non-cognitive achievement measures. Other literature reviews find additional evidence that more engaged students achieve better learning results (see e.g. Fredricks et al., 2004). High educational achievement in turn is one central predictor of labour market outcomes later in life (Woessmann, 2016). Next to this, higher engagement can also lead to lower incidence of delinquency, aggression, and early school dropout (see e.g. Fredricks et al., 2004; Hill & Werner, 2006).[3] Against this background, knowing how to engage students in classrooms would be of great social value. Such knowledge could be incorporated into teacher training and, if successfully applied in classrooms, could yield substantial monetary and non-monetary rewards for students themselves and for society as a whole. However, a necessary precondition for this is that student engagement can be altered by outside influences. It is reassuring that researchers have found engagement to be relatively malleable, among others by school and classroom factors (Fredricks et al., 2004).

In this article, I use data on teachers' intentions to engage their students in learning from the 2011 waves of the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS). TIMSS and PIRLS are internationally comparative assessments of student achievement in math, science and reading of fourth- and eighth-graders. TIMSS deals with the two subjects of mathematics and science, while PIRLS is dedicated to reading. Both studies are administered by the International Association for the Evaluation of Educational Achievement (IEA), an organization that has been carrying out international assessments of student achievement since 1959. Via the two studies, it aims at enabling participating countries to improve their educational policy. Next to achievement data, both studies collect extensive background information on student, family, and institutional factors. Choosing the 2011 waves of the two studies is beneficial in two ways: First, 2011 marked the first occasion that the two studies comprised the so-called "Engaging Students in Learning Scale" (ESL scale), which serves as my main independent variable. Second, 2011 is the only year so far in which TIMSS and PIRLS were sampled together – at least those in fourth grade. That provides the unique opportunity of observing elementary school students in three different subjects, namely math, science, and reading. In total, 34 countries and 3 benchmarking entities participated in the joint sampling of TIMSS and PIRLS.[4] The full TIMSS and PIRLS 2011 database contains information for 185,475 students, 171,098 parents, 14,258 teachers, and 6,469 school principals (Foy, 2013). For a variety of reasons, however, I solely focus on Germany, the largest European country, in the present paper. First of all, there are only very few countries in which a significant share of elementary school students is taught by different teachers in different subjects. This, however, is a pre-condition for my identification strategy. Secondly, these countries are predominantly Arab countries that may be very different from Western countries in terms of learning approaches and learning content. Since learning is context sensitive, I also refrain from pooling the available data from other countries, as educational systems differ and teaching practices that are beneficial in one country do not need to be equally beneficial in another country.

[3] In the educational literature, engagement is often defined as multifaceted construct encompassing several, intertwined dimensions. These include the affective sphere (do children like school?), a behavioural component (participating in schooling activities, doing homework) as well as a cognitive part (e.g. motivation) (Perdue et al., 2009). In the present analysis, I will use the somewhat narrower definition of engagement given by McLaughlin et al. (2003), which is also the basis for the indicator as constructed the TIMSS and PIRLS.
[4] Benchmarking entities are regional jurisdictions within countries that participate separately in the assessments.

The fact that I observe test scores in multiple subjects for each student allows me to use within-student estimation for identification of the effect of teachers' intentions to engage their students on achievement outcomes. By doing so, I am able to control for individual time-invariant characteristics that also affect achievement outcomes such as underlying ability, lagged achievement, parental background, and school resources. A major advantage for the identification of the desired effect is the fact that I am using elementary school students as units of analysis. Most of these students have no long history of different teachers in different subjects, which makes it more likely that any observed teacher effect can be attributed to the current instructor. What is more, selective admission and within-school sorting by ability should be less of a problem in elementary school than in secondary school. A caveat of the chosen approach is that I assume that the effect of the ESL scale on student achievement is the same in all three subjects. This is a common assumption in empirical research on the economics of education and in line with the theoretical background outlined above – namely that student engagement is a subject-independent input into educational production. Yet, in this paper I can check how restrictive this assumption is by comparing estimates obtained by pooling only two subjects at the same time.

This article adds to the literature by being the first to empirically estimate the causal effect of teachers' intentions to engage their students on achievement outcomes. It thereby expands the relatively scarce economic literature on the effectiveness of specific teaching practices. The results I present in this paper indicate that teachers' intentions to engage their students as measured by the ESL scale have no significant effect on student achievement. However, I do find a significant positive effect for students from low socio-economic backgrounds. The size of this effect is small to modest. A one standard deviation increase on the ESL scale raises test scores in math, science and reading by 0.05 standard deviations of the test score distribution. The latter is equal to about three points on the achievement test. This finding suggest that engaging students in learning can yield societal gains in terms of greater equality of opportunity. This is important in the context of Germany, where intergenerational educational mobility is generally found to be low (see e.g. Heineck & Riphahn, 2009). The finding is also in line with some previous research on teaching practices that has found differential results by subgroup under study (see e.g. Lavy, 2011). What is more, there seems to be some negative sorting of students to teachers in Germany. This is reflected by the fact that the "naïve" OLS results tend to be more negative than the fixed effects estimates. Such sorting could for instance happen if parents of comparably worse students send their children to schools with teachers who put particular emphasis on engaging students. It could also be the result of teachers adjusting their behavior to a class of low performing students in the sense that they try to more actively encourage their students to get involved in the subject matter. The estimations also suggest that subject differences in the effect of the ESL scale on achievement are not a major cause of concern. Overall, the results are robust to a variety of sensitivity checks. Generally, it should be noted that the within-student between-subject approach used in this paper does not allow me to entirely rule out bias stemming from unobserved teacher characteristics. This weakens the interpretation of my results as reflecting causal mechanisms.

The remainder of this article is structured as follows: Section 1 provides an overview over the economic literature on teacher effectiveness. Section 2 outlines the identification strategy. Section 3 introduces the data, the analysis sample and the central variables used in the empirical estimations. Section 4 presents the results of the OLS and student fixed effects models as well as several robustness checks. Finally, section 5 concludes.

1. Related Literature

The economic literature on teacher effectiveness[5] is relatively young. This is especially true for the focus on what happens in classrooms as compared to what observable teacher characteristics can predict student achievement. Nevertheless, two main strands can be identified in this literature: one that is trying to measure effectiveness in terms of a teacher's value added to student achievement and one that is relying on the direct measurement of objective and subjective teacher characteristics. Both approaches have its merits and weaknesses.

The first strand, value-added modeling, was heavily relied upon by early papers on teacher effectiveness. The idea behind this approach is to compare students' achievement gains (value-added) in a certain period of time and ascribe them to different teachers while conditioning on a set of covariates. This literature has consistently found huge differences between teachers (for overviews see Koedel et al., 2015; Hanushek & Rivkin, 2010). Chetty et al. (2014) and Hanushek (2011) estimate that the monetary gains from improving the quality of teaching based on value-added evidence would be quite substantial. Personnel decisions based on past track record could for instance be efficient in the case of mandatory layoffs or other retention-or-removal situations. However, while value-added measures tell us *which* teachers are effective, they do not tell us *why* teachers are effective. Thus, success cannot be replicated. This is not satisfactory from an efficiency standpoint, as teacher training cannot be improved. Instead, every teacher would have to be tested and approved or removed, which would be very costly as one would have to train more teachers than are actually needed.

The second main strand in the literature aims at measuring teacher characteristics directly. This is relatively straightforward in the case of objective traits such as age, gender, education, and experience. The main result from studies dealing with these characteristics is that teacher experience has a positive effect on student achievement (see e.g. Goldhaber & Anthony, 2007; Clotfelter et al., 2006; Rivkin et al., 2005). However, the effect appears to be non-linear, leveling off after around five years (Rivkin et al., 2005). Most other characteristics are generally found to have either negligible or no effects on student achievement (Aaronson et al., 2007; Rivkin et al., 2005). Significant effects can mostly be found for certain subgroups of students or specific student-teacher pairings on certain characteristics. For instance, Paredes (2014) finds that teacher-student gender matching can have positive effects on performance via role model effects.

Since readily observable teacher characteristics can only explain a small fraction of the variation in student achievement, many researchers have tried to go beyond objective characteristics and attempted to assess what happens in classrooms. The data for this are typically gathered in one of three ways: They are either based on (1) classroom observations by trained experts, (2) student reports, or (3) teacher self-reports. Prominent examples of the first group of studies are provided by Tyler et al. (2010) and Kane et al. (2011) who use data collected by the Cincinnati Public Schools' Teacher Evaluation System (TES). They find that observational quality measures are clearly related to achievement outcomes. In an analysis of different components of the overall TES score, Tyler et al.

[5] Instead of teacher effectiveness, some researchers prefer to use the term teacher quality. Essentially, both terms express the same concept, namely the teachers' effectiveness in conferring skills upon their students, and can be used interchangeably.

(2010) find that teachers who place more emphasis on the classroom environment instead of focusing on specific teaching practices can reap particularly large achievement gains among their students. Similarly and particularly relevant for this research, teachers who engage their students in questions and discussions are more effective than teachers who routinely focus on additional content. This result, however, is only valid for reading, not for mathematics. The result of questioning and discussion being particularly effective in reading is corroborated by Kane et al. (2011).

Another recent contribution based on observational classroom data is a paper by Blazar (2015). Lamenting that the literature has not yet coalesced around certain teaching practices he uses two different observational instruments, the Mathematical Quality of Instruction (MQI) and the Classroom Assessment Scoring System (CLASS) to investigate learning among fourth- and fifth-graders. Especially the latter concept is interesting for this research as it revolves around general classroom quality instead of focusing purely on mathematics as the MQI does. The author uses two different dimensions of the CLASS indicator, namely classroom emotional support, which focuses on the classroom climate and teachers' interactions with students, and classroom organization, which among others comprises specific learning formats. In both cases, the author finds no significant effects on student achievement.

Studies based on student or teacher reports generally make use of large-scale assessment data. Most of these papers deal with the question whether "traditional" teacher-centered teaching or "modern" student-centered teaching is more effective in conferring skills upon students. The former is characterized by heavy reliance on lecturing and direct instruction, while the latter shifts the emphasis onto group work and discussions among peers. Using teacher self-reports, Schwerdt & Wuppermann (2011) find tentative evidence that traditional lecture-style teaching is superior to modern teaching. However, they admit that their results may be influenced by selection bias and conclude that traditional teaching is at least not worse than modern teaching. Van Klaveren (2011) finds no significant effect of lecture-style teaching, while other studies provide some evidence of explicitly negative effects of some elements of modern teaching (Murnane & Phillips, 1981; Goldhaber & Brewer, 1997).

Distancing themselves from the notion that modern teaching always comes at the expense of traditional teaching (and vice versa) and postulating that the two can co-exist alongside each other, two papers find positive effects for both teaching approaches. Lavy (2011) finds large payoffs for both traditional and some facets of modern teaching, which do, however, differ by subgroup under study. While girls and students from low socio-economic backgrounds seem to benefit most from teacher-centric education, students from higher socio-economic backgrounds can be especially well targeted by modern teaching methods. Using TIMSS data for US eighth-graders, Bietenbeck (2014) demonstrates that traditional and modern teaching methods promote different skills in children. While traditional teaching is particularly useful to increase students' factual knowledge, modern teaching improves reasoning skills. He goes on to argue that standardized achievement tests do not measure reasoning skills well, which can explain the often negligible or even null effects found in connection with student-centered learning. Both Lavy (2011) and Bietenbeck (2014) rely on student reports to measure teaching styles.

There are very few economics papers based on large-scale survey data that go beyond the simple dichotomy of modern and traditional teaching and try to shed light on the nexus between specific

teaching practices and student achievement.[6] A notable exception is provided by Aslam & Kingdon (2011), who use teacher self-reports in Pakistan to find that certain teaching practices are indeed significant predictors of student outcomes. This is especially true for the use of quizzing and questioning in class as well as planning of the lesson at home. They go on to demonstrate that the omission of variables capturing teaching practices biases coefficients of more routinely researched variables such as job experience, tenure, and education. However, their results are mostly confined to private schools and originate from a developing country.

Similar to Aslam & Kingdon (2011), I focus on specific teaching practices and use teacher self-reported data, thereby adding to the second strand of literature.

2. Estimation Strategy

When estimating the effect of teaching practices on achievement, endogeneity bias may arise for different reasons. The results as indicated by the coefficient of interest could be confounded by biases due to systematic self-selection and sorting of students and teachers to each other and/or to specific schools. For instance, if students with particular unobserved characteristics such as particularly high ability systematically select into schools with a large share of teachers who employ potentially engaging teaching practices, any "naïve" OLS estimate of the effect of teacher intentions to engage their students on student achievement would be upward biased.

One way of circumventing the endogeneity problem stemming from unobserved individual factors such as ability, lagged achievement, family background, and motivation is to estimate within-student between-subject models. Recently, this has been done by a number of economists (see e.g. Dee, 2005; Clotfelter et al., 2010; Schwerdt & Wuppermann, 2011; Bietenbeck, 2014; Lavy, 2015). This procedure rules out bias due to unobserved individual factors, because all the variation in these models stems from performance differences of the same individual in different subjects and their (systematic) association with differential input factors in these subjects. Based on this approach, I examine whether differences in achievement are systematically related to differences in teachers' intentions to engage students in math, science, and reading. This is made possible by the fact that many students face different teachers in some or all of the three subjects. The fact that I am using information on three different subjects provides me with extra variation as compared to using just two subjects (mostly math and science) as is usually done in empirical research on education.[7] The basic idea for identification is that student, teacher and school characteristics are constant across subjects except for differences in teachers' intentions to engage their students and in all control variables.

[6] There is a rather large literature on the effects of computer use in classrooms, which, however, is not directly relevant for this article. Recent contributions from this strand of literature suggest that ICT use in classrooms is not per se good or bad, but depends on how computers are used and for what tasks (see e.g. Falck et al., 2015; Lorena Comi et al., 2017).

[7] Lavy (2015) also uses three different subjects. He examines the effect of instruction time on achievement among 15-year-olds.

Based on this identification strategy, I consider an education production function of the following form:

$$A_{ijsk} = \alpha_i + \beta ESL_{js} + \gamma X_{is} + \delta T_{js} + \eta S_s + \tau_j + \xi_s + \upsilon_k + \varepsilon_{ijsk}, \quad (1)$$

where A_{ijsk} is the achievement of student *i* with teacher *j* in school *s* and subject *k*, ESL_{jsk} is teacher *j*'s score on the ESL scale in school *s*, X_{ics} is a vector of control variables pertaining to the personal background of student *i* in school *s*, T_{jsk} is a vector of covariates related to the personal background and teaching characteristics of teacher *j* in school *s*, and S_s is a vector of characteristics of school s.[8] The coefficient β is the main parameter of interest. τ_j, ξ_s and υ_k represent unobserved characteristics of the teachers, the schools, and the subjects, while ε_{ijsk} is an idiosyncratic error term. Importantly, α_i is a student fixed effect that drops out of the within-student models. This fixed effect captures the effects of a student's family background, his or her prior educational career, innate ability, motivation, and other constant personality-related factors. Due to the student fixed effect, all general individual background factors that are observed in the data, denoted by X_{is}, also leave the model. Note that by controlling for a student fixed effect, I also control for school-level factors, as every student is only observed in one school. For that reason, the terms ηS_s and ξ_s drop out of the equation, too. Thus, the within-student models allow me to control for a wide range of student and school characteristics and their interactions that may cause bias in the estimations. Such bias could arise if there is a correlation between (unobserved) general school quality and the use of potentially engaging teaching practices by teachers employed at this school. If teachers who spend a lot of time trying to engage their students systematically select into "good" schools, upward bias will be introduced. The bias would be even stronger, if high ability students would also self-select into these schools. However, a negative bias would also be imaginable, for instance if some teachers who frequently use engaging teaching practices are at the same time keen on helping disadvantaged children and, as a result, sort into more "problematic" schools. Finally, it is imaginable that teachers adjust their behaviour according to the group of students they are facing. For example, teachers may more frequently resort to encouraging their students and getting them involved in the subject matter, if they are dealing with a group of less motivated students. This would also be a cause of downward bias in the OLS estimates. It is a priori unclear what kind of bias (upward or downward) should be expected. In any event, the student fixed effect effectively ensures that none of the above is a problem in the present study.

However, there are some issues in connection with my identification strategy that warrant mention. First of all, my approach hinges upon the assumption that β is the same in all three subjects. This is in line with the theory laid out at the beginning of this article that student engagement should be a

[8] Note that I rule out within-school variation in class size, as classroom composition usually does not differ across subjects in elementary school in Germany. This claim can be backed up by the data: Pairwise correlations between teacher-reported class sizes show coefficients of more than .98 in all three cases. Therefore, it seems safe to assume that the remaining variation is due to measurement error and, more generally, not relevant for the sake of this estimation. Practically, I am using teacher-reported class size in science as a proxy for all class sizes, as there are the fewest missing values in this variable.

subject-independent input in educational production. I will, however, provide further evidence that this assumption holds true for my sample.

Second, the effect captured by the coefficient β is "net" of any spillovers from one subject to another (i.e. if a student "imports" his or her higher engagement triggered by teacher actions in subject A to subject B).

Third, a threat to my identification strategy could be student sorting to schools and teachers by subject-specific ability. Positive bias could result if students with high ability in math systematically chose schools, in which the math teachers apply more engaging teaching practices. For this to happen, however, there would have to be clear differences in subject-specific ability between students. Yet, it is unclear to what extent this is true. Clotfelter et al. (2010) provide some evidence that academic ability is in fact highly correlated across subjects. Even if significant subject-specific ability differences existed, a number of additional preconditions would have to be fulfilled so that my identification strategy would be threatened. First, parents would have to have prior knowledge about the specific strengths and weaknesses of their offspring. Second, teaching practices would have to systematically differ between subjects within the same school. And third, parents would have to have information about how teaching practices differ within schools. While the first condition may hold, it seems unlikely that all three conditions are met for a significant share of students. This notwithstanding, I can partially take care of the problem with the available data. Practically, I use a control variable in the empirical estimations that indicates whether or not a school suffers from a shortage of teaching materials in each of the three subjects. The idea behind this approach is that systematic differences in teaching practices within schools most likely occur in schools that specialize in certain subjects. Such schools, in turn, should be less likely to suffer from shortages of teaching materials in that subject.

Fourth, a further concern would be systematic within-school sorting of students to teachers. This is less of a problem for elementary school students than for secondary school students, however, as there are very few electives in elementary schools. Furthermore, for such sorting to happen, the criteria outlined in the previous paragraph would have to be met, too, i.e. knowledge about subject-specific ability, subject-differences in teaching practices within schools, and information about the latter. One instance, in which information about this could exist, is after the children have started school, i.e. after the first, second, or third grade. If they then switch to a different class, sorting could theoretically take place. In reality, this does not seem to happen frequently, though. Ammermueller & Pischke (2009) provide evidence that systematic ability grouping does not happen in German elementary schools in grade 4. Furthermore, I can deal with this problem by stratifying my sample according to good proxies of whether or not sorting is likely in a school. For instance, I know the total number of students in grade 4 in every school. By splitting the sample into smaller schools, which in many cases have only one class per grade, and schools with more classes, I can see whether any effects are concentrated among the larger schools that offer more room for within-school tracking. I also observe how much emphasis is given by schools to academic success. I assume that sorting into special ability classes is more likely in these schools than in others. Overall, the results I obtain from these stratifications are very similar to the baseline results. This suggests that within-school sorting by ability is not a likely cause of bias.

Fifth, while it is true that my estimates are stripped of any unobserved individual and school-level heterogeneity, they could still be contaminated by non-random sorting of teachers into certain

teaching practices. This challenge is faced by virtually all studies that deal with teaching practices and are not based on randomized controlled trials. Any bias introduced due to teacher sorting into specific teaching practices would be captured by the term τ_{js} in equation 1. In practice, such sorting could arise if teachers with more favourable unobserved characteristics such as motivation or pedagogical skills use more engaging teaching practices. In that case, any positive effect of *ESL* would be over-estimated due to unobserved teacher traits. In order to minimise this risk, I have included a large set of teacher characteristics and teacher behaviour variables as controls. In this respect, the expansion of teacher- and teaching-related information that has come with the 2011 wave of the TIMSS and PIRLS studies is of great value to me. It is further reassuring that Kane et al. (2011) find empirical evidence in teacher-fixed-effects estimations (with fewer teacher controls) that unobserved sorting of teachers into teaching practices is likely not a big issue in a similar setup. However, a closer look at the data is certainly warranted. Departing from the well-established idea that the amount of selection on observables provides some guidance to the magnitude of selection on unobservables, in Table A.1 in the annex I provide estimates of the correlations between observable teacher characteristics and teachers' scores on the ESL scale. 10 out of 21 teacher controls turn out to be significantly related to the intention to engage students. So there does seem to be some systematic relation between teacher characteristics and certain teaching practices. To get an idea on how this affects my estimation outcomes, I will estimate different models with and without teacher- and teaching-related controls.[9] It is reassuring that the results do not change much depending whether the set of covariates is included or not.

3. Data

3.1. *The TIMSS and PIRLS Studies*

TIMSS and PIRLS are large-scale international assessment studies dealing with the educational achievement of fourth- and eighth-graders. TIMSS tests the knowledge and skills of students in math and science, while PIRLS is dedicated to the subject of reading. TIMSS is the "older" study among the two. The first wave of testing was conducted in 1995. Subsequently, the study has been carried out every four years, i.e. in 1999, 2003, 2007, 2011, and 2015. PIRLS was first established in 2001 and has since been conducted every five years, i.e. in 2006, 2011, and 2016. Thus, so far the year 2011 marks the only occasion on which the two studies have coincided. As a result of this special timing, several countries decided to sample TIMSS and PIRLS together – at least among fourth-graders. The resulting dataset comprises information on 34 different countries and 3 benchmarking entities and allows researchers to analyse achievement of elementary school students in three different subjects. In this study, I focus on country information on Germany. Here, a total of 4,067 students that are representative for the population of fourth-graders in the country were sampled (Bos et al., 2012a, 2012b). Due to the sampling design of TIMSS and PIRLS not all students were sampled with the same probability. For this reason, I apply probability weights throughout the analysis.

Both TIMSS and PIRLS are administered by the *International Association for the Evaluation of Educational Achievement* (IEA), an organization first established in 1958 that consequently disposes over vast experience in monitoring educational processes and outcomes. TIMSS and PIRLS apply a

[9] Falck et al. (2015) follow the same procedure in an analysis of the effects of computer use in classrooms on achievement and find that the results do not change significantly.

two-stage stratified sampling design. In the first stage, participating schools are chosen and in the second stage, classes within these schools are selected. Stratification in TIMSS and PIRLS takes into account regional differences, school-type differences, level of urbanisation, socio-economic indicators, and school performance on national examinations (Joncas & Foy, 2013). Testing in 2011 was carried out on two consecutive days; in half of the schools, students started with the TIMSS questionnaire on the first day and in the other half, students answered the PIRLS questionnaire first. The TIMSS assessment framework was organized around two different dimensions in 2011: content and cognition. The content section focused rather closely on what students should have learned in their curricula. In mathematics, this section contained questions related to numbers, geometric shapes, and measures, while in science, it comprised life science, physical science, and earth science. The cognitive section put bigger emphasis on applying knowledge and reasoning. Generally, questions were split about evenly into multiple-choice and open-response. In total, the TIMSS questionnaire encompassed 175 items in math and 217 in science. The PIRLS assessment framework also focused on two different sections: reading for literary purposes and reading to acquire and use information. Within each of these sections, four comprehension processes were assessed: retrieving, inferencing, integrating, and evaluating. The text passages encompassed around 800 words with 13 to 16 questions underneath. PIRLS 2011 comprised a total of ten passages (five for each section), resulting in 135 questions (Martin & Mullis, 2013). To obtain as much information about the students' learning environment as possible, in addition to the actual tests, background questionnaires were administered to students, their parents, teachers, and school principals (Bos et al., 2012a, 2012b).

For my main variable of interest that measures teachers' intentions to engage their students in learning I use information from the teacher questionnaires. In 2011, TIMSS and PIRLS divided their teacher questionnaires into general questions that were answered by all teachers independent of the subject as well as subject-specific questions. The former are most useful for the sake of my three-subject comparison. Among others, the so-called Engaging Students in Learning Scale was introduced in this section of the questionnaires (Mullis et al., 2012a). The scale is inspired by work done by McLaughlin et al. (2005), who introduced the concept of student content engagement (Martin et al., 2012; Mullis et al., 2012a, 2012b). It is based on teacher self-reports on specific classroom actions. Such teacher data have previously been used by Hidalgo-Cabrillana & López-Mayan (2015), Schwerdt & Wuppermann (2011) as well as Aslam & Kingdon (2011). The ESL scale is based on a six-item instrument. Specifically, teachers were asked how often they (1) summarize what students should have learned from the lesson, (2) relate the lesson to students' daily lives, (3) use questioning to elicit reasons and explanations, (4) encourage all students to improve their performance, (5) praise students for good effort, and (6) bring interesting materials to class. All questions could be answered on a four-point scale ranging from "every or almost every lesson" over "about half the lessons" and "some lessons" to "never or almost never" (IEA, 2011a, 2011b). Using item response theory, the raw data were transformed into the ESL scale by the IEA. The scale is constructed such that the mean value of all participating countries is 10 and the standard deviation 2 (for more detailed information, see Martin et al., 2012). Note that scores on the ESL scale are constant within teachers, as teachers do not make statements on the use of potentially engaging teaching practices by subject.

3.2. *Sample Selection and Descriptive Statistics*

My full sample comprises 4,067 students in 205 classes and 197 schools. However, in order to estimate the desired effect, I have to apply certain restrictions. First, I have to limit the analysis to students who have no more than one teacher per subject. That way, every student can be uniquely linked to exactly one teacher in math, one teacher in science, and one teacher in reading. I thereby lose 135 students. Second, I consider only those students, whose teachers have valid information on the ESL scale, which means that a further 106 students are excluded. And finally, I only use those students, who participated in the achievement tests in all three subjects, which eliminates 413 students.[10] My final sample then consists of 3,413 students in 171 classes in 170 schools. This translates into 10,239 student-subject observations. Out of the 3,413 individual students, 1,684 students are taught by the same teacher in all three subjects, 1,024 students have the same teacher in science and reading but not in math, 434 students have the same teacher in math and reading but not in science, 190 students have the same teacher in math and science but not in reading, and 81 have different teachers in all three subjects. I leave the 1,684 students who have the same teacher in all subjects in the sample because of the valuable information on control variables that they provide. The large number of students with the same teacher is partly a result of the fact that in many cases math and science are taught as a single subject. In the robustness section, I demonstrate that including or excluding these students does not substantially alter my results.

Table 1 provides summary statistics of the ESL scale by subject. The teacher values on the scale range from 2.95 to 13.27, have a mean of 8.74 and a standard deviation of 1.56. This indicates that German teachers make on average less use of potentially engaging teaching practices than teachers in other countries. Generally, very few teachers state that they use the techniques in question never or almost never. Table A.2 in the annex shows the means of the TIMSS and PIRLS achievement scores and the ESL scale as well as standard deviations both within students and between students. The mean test score for students in Germany is 533.8, well above the international centerpoint of 500 that was set as the mean achievement value in the first TIMSS and PIRLS studies. The standard deviation of test scores between students is 64.7, while the within-student standard deviation is about half as large at 31.5. That means, there is substantial variation within students that can be explained in the regressions. The ESL scale has a between-student standard deviation of 1.36 and a within-student standard deviation of 0.56. To make the data more comparable across subjects and facilitate the interpretation of the results, I standardise both the achievement scores in math, science, and reading to have mean 0 and standard deviation 1 and the teacher scores on the ESL scale.[11] For detailed information on the remainder of the variables used in the empirical estimations including definitions and summary statistics by subject consider Table A.3 in the annex.

[10] In order not to lose too large a number of observations, I impute missing values on control variables by setting them to the respective mean and adding a dummy variable taking on the value 1 if the value was generated that way. In the case of dichotomous controls, I simply add a category for missing and use two dummies in the estimations with missing as the reference.

[11] In the analysis, I use the first plausible value for all subjects. Each participant in TIMSS and PIRLS gets a total a five plausible values describing his or her performance. Plausible values are used to correct for different degrees of difficulty in the exercises, as not all students answer the exact same questions.

Table 1

Summary statistics of ESL scale by subject

	Mean	Std. dev.	Min.	Max.	Observations
Overall	8.74	1.56	2.95	13.27	10,239
Math	8.73	1.61	2.95	13.17	3,413
Science	8.74	1.57	4.57	13.27	3,413
Reading	8.75	1.51	4.57	13.27	3,413

Data source: TIMSS/PIRLS 2011.

4. Results

4.1. *Estimates of the Effect of Teachers' Intentions to Engage Students on Achievement*

Table 2 reports the estimated coefficients of the effect of teachers' intentions to engage their students on individual achievement. All regressions contain subject fixed effects and are weighted by probability weights as supplied in the TIMSS dataset. Columns (1) and (3) present results of pooled OLS models, while columns (2) and (4) report estimates based on student fixed effects specifications. The coefficients in the OLS models are negative and borderline significant. While the estimate reported in column 1 from a model containing only personal and school background control variables reaches statistical significance at the 90% level, the coefficient from the full model that includes comprehensive information on teachers, classrooms and teaching practices barely fails to reach significance (see column 3). These estimates, which suggest that teachers' intentions to engage their students may have a negative effect on achievement, are potentially biased by all sorts of student and teacher self-selection into schools and classrooms. In fact, when considering the student fixed effect models and especially my preferred specification in column 4, statistical significance disappears and the point estimates are almost equal to zero. In these models, student self-selection should play no role. Against the background of potential sorting of teachers into different teaching practices, it is reassuring that the inclusion of teacher- and teaching-related control variables in column (4) as opposed to column (2) does not significantly alter the results. The basic conclusion to be drawn from these analyses is therefore, that teachers' intentions to engage their students do not directly affect achievement. The difference in the results between the OLS models and the student fixed effects models suggests, that there may be some negative sorting of either students to teachers or teachers to students and/or schools in Germany. For instance, it may be that parents of less motivated or low-ability children intentionally send their offspring to schools that are known for their engaging teaching practices. The negative coefficients in the OLS models could also be the result of reverse causality, i.e. if teachers of worse students more often resort to potentially engaging teaching practices than other teachers. This makes intuitive sense, as especially low-performing students may need and receive additional teacher support. For high-performing students, it is imaginable that teachers substitute potentially engaging teaching practices for other classroom actions, since students follow the course content in any event. While there is no natural counterpart to potentially engaging teaching practices, the most likely alternative would be giving additional

exercises, as instructional strategies to raise engagement mostly focus on repetition, summarizing, questioning, encouraging and praising. Thus, engaging students may be related to more intense study of certain material at the cost of additional material.

Since it is likely that a mix of different classroom actions produces the highest achievement, I report estimates for models that allow for non-linearity in the effect of potentially engaging teaching practices on achievement in the lower panel of Table 2. Practically, I add a squared term of the standardized teacher score on the ESL scale to the models. However, no estimates from these models turn out significant. This may have something to do with the way the questions are framed in the questionnaires. Teachers are asked whether they apply the potentially engaging teaching practices in "every or almost every lesson", "about half the lessons", "some lessons", or "never". This is different from asking whether teachers try to engage their students during the whole lesson, half the lesson or less often, since it does not offer information on the actual time spent on engaging students *during* those lessons. In the latter case, a non-linear effect would intuitively be more likely to appear.

Table 2

Estimated Effect of Teachers' Intentions to Engage Students on Student Achievement

	OLS (1)	Student FE (2)	OLS (3)	Student FE (4)
(a)				
ESL	-.041*	-.004	-.030	.000
	(0.02)	(0.01)	(0.02)	(0.01)
(b)				
ESL	-.040	-.010	.037	.036
	(0.13)	(0.08)	(0.14)	(.11)
ESL-squared	-.000	.000	-.002	-.001
	(0.00)	(0.00)	(0.00)	(0.00)
Subject dummies	yes	yes	yes	yes
Personal and school characteristics	yes		yes	
Teacher and teaching characteristics			yes	yes
Number of observations	10,239	10,239	10,239	10,239

Notes. * p < 0.10, ** p < 0.05, *** p < 0.01. Table shows regressions of students' z-standardized achievement scores on teachers' z-standardised values on the ESL scale. Fixed effects are at the student level. Each regression also contains subject fixed effects. In the upper panel, only the level of teachers' values on the ESL scale are considered. In the lower panel, the levels and squared terms of values on the ESL scales enter the models. All regressions are weighted by the students' sampling probability. Standard errors in parentheses are clustered at the classroom level. Control variables are listed in Table A.3. Data source: TIMSS/PIRLS 2011.

So far, I have assumed that trying to engage students in learning is equally effective (or ineffective) in math, science, and reading. In reality, this need not be the case. For that reason, I estimate models based on the three possible samples that include only two of the three subjects. The results of this exercise are presented in Table A.4. Again, the estimated effect of ESL on achievement does not turn out significant in any of the models and the point estimates do not vary much across the different models. This is no definitive proof for the hypothesis that engaging students has the same effect in all subjects. However, the mean value of the three estimates is -.003 and thus almost identical to the full sample estimate.

4.2. *Heterogeneous Effects*

No significant effects in the full sample do not preclude the possibility that certain subgroups of students may still be affected by their teachers' intentions to engage them. For instance, in an article on modern and traditional teaching practices, Lavy (2011) finds that children from different socio-economic backgrounds are quite differently affected by teachers' classroom actions. In other words, a specific teaching practice may be good for some students but bad for others, simply because different groups of students have different needs. Heterogeneous effects could also be the result of differences in the "baseline" engagement of different groups of students. If there are many students in a particular group who are engaged during the whole lesson in any case, further time spent on engaging them should not be advantageous. In all likelihood they would even be negatively affected because their teachers' attempts to further engage them crowd out alternative actions such as giving additional exercises. While this example is an extreme case, differences in the average "baseline" engagement could be expected between students from different socio-economic backgrounds and between boys and girls. For instance, boys are often found to be less engaged in schooling matters than girls already in elementary school (see e.g. McCoy et al., 2012). Socio-economic status could matter if students from higher socio-economic backgrounds have learned a more "pro-education" attitude from their parents. However, empirical evidence on this is not conclusive (for an overview see Shernoff, 2013). Finally, one could expect differences between children who speak German with their parents and those who do not, as the latter may need to be addressed differently in class.

Table 3 presents the results of the subgroup-specific estimations. Again, the upper panel deals with linear analyses, while the estimates in the lower panel stem from specifications that allow for non-linearities. As can be seen in column 2, I find a positive effect for children from low socio-economic backgrounds.[12] The magnitude of the effect is small to modest: A one standard deviation increase on the ESL scale raises test scores in math, science and reading by 0.05 standard deviations of the test score distribution or approximately three points on the achievement tests. This finding suggests that engaging students in learning can yield societal gains in terms of greater equality of opportunity. This is important in the context of Germany, where intergenerational educational mobility is generally found to be low (see e.g. Heineck & Riphahn, 2009). Still, the lower panel of column 2 suggests that the observed effect may not be linear along the distribution of values on the ESL scale. While the level coefficient is positive, significant and very large, the coefficient estimate of the squared term is significantly negative. However, the calculated turning point of 16.63 is well outside the data range

[12] Children are defined as having a low socio-economic status if neither of their parents has a post-secondary degree. All others are classified as having a high socio-economic background.

Table 3

Estimated Effect of Teachers' Intentions to Engage Students on Standardised Test Scores for Different Subgroups, Student Fixed Effects Models

	Socio-economic background		Gender		Language mostly spoken at home	
	High	Low	Boys	Girls	German	not German
	(1)	(2)	(3)	(4)	(5)	(6)
(a)						
ESL	-.013	.047*	.004	.007	.008	-.014
	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.03)
(b)						
ESL	-.099	.399**	.135	-.043	.132	.035
	(0.10)	(0.20)	(0.16)	(0.11)	(0.11)	(0.32)
ESL-squared	.003	-.012*	-.005	.002	-.004	-.002
	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.01)
Subject dummies	yes	yes	yes	yes	yes	yes
Teacher and teaching characteristics	yes	yes	yes	yes	yes	yes
Number of observations	4,314	3,297	5,124	5,115	7,401	1,866

Notes. * p < 0.10, ** p < 0.05, *** p < 0.01. Table shows regressions of students' z-standardized achievement scores on teachers' z-standardised values on the ESL scale. Fixed effects are at the student level. Each regression also contains subject fixed effects. In the upper panel, only the level of teachers' values on the ESL scale are considered. In the lower panel, the levels and squared terms of values on the ESL scales enter the models. All regressions are weighted by the students' sampling probability. Standard errors in parentheses are clustered at the classroom level. Control variables are listed in Table A.3. Data source: TIMSS/PIRLS 2011.

(recall that the data are standardized). Thus, in practical terms there does not seem to be a relevant non-linear relationship. This conclusion is corroborated in a model where teachers' intentions to engage their students are expressed in logs. Here, no significant negative effect is found for the squared value of ESL. Apart from children from low socio-economic backgrounds, no other subgroups seems to benefit from potentially engaging teaching practices. Thus, for children from high socio-economic backgrounds, boys, girls, children who mostly speak German at home and those who do not, the null results of the full sample are confirmed.

Generally, it is important to note that subgroup differences are less likely to be a consequence of sorting on unobserved teacher characteristics than the overall results. This is the case, because even if there is systematic sorting of teachers into certain teaching practices, such sorting will affect all subgroups in the same way. The only concern in this case would be subgroup-specific sorting. For such subgroup-specific sorting to happen, there would, for instance, have to be unobserved teacher characteristics, which are especially beneficial or detrimental to just a particular subgroup of

students and teachers would have to sort into certain teaching practices along these characteristics – something that has been deemed unlikely by other researchers in the past (see e.g. Lavy, 2011).

4.3. *Robustness of the Results*

In the following section, I present some analyses that (1) underscore the robustness of my results to alternative specifications and definitions of treatment and (2) and support their causal interpretation. As a first robustness check, I excluded all classes with less than 16 students from the analysis.[13] This reduces my sample by 324 observations to 9,915 student-subject pairings. The reason for excluding those very small classes is that the general classrooms dynamics may be very different in them as opposed to larger classes. Most importantly, in larger classes the potential for disturbances and interruptions increases, which may render the task of engaging students more important. However, one could also imagine that it is easier for teachers to "reach" their students with their attempts to engage them in small groups. Yet, the estimates presented in panel (a) of Table 4 suggest that different classroom dynamics in small classes do not decisively drive my results. The finding of no overall effects and modest gains for students from low socio-economic backgrounds holds. In fact, the point estimate of the coefficient for students from low socio-economic backgrounds is exactly the same as in the full sample regression (0.047).

As a second robustness check, I restricted my sample to those students that are not taught by the same teacher in all three subjects. This cuts my sample roughly in half to 5,187 students. Yet, it does not have a significant effect on the results, which are reported in panel (b) of Table 4. As compared to the estimates in Tables 2 and 3, they seem only marginally more positive. For example, the estimated benefit from a one-standard-deviation increase in the use of potentially engaging teaching practices for students from low socio-economic backgrounds is now 0.054 standard deviations of the test score distribution and thus 15% higher than in column 2 of Table 3. Apart from this significant effect, the general pattern of no significant effects in the full sample and for all other subgroups holds.

A third robustness check aims at manipulating my main independent variable, the score on the ESL scale. Recall that the IEA provides a ready-made scale based on item response theory in the dataset. I used the information from the six items to construct an alternative ESL scale based on factor analysis. First of all, it was encouraging that all items loaded on one factor. Visual inspection of the screeplot confirmed this result. The results of the regressions with the resulting factor score as principal regressor are provided in panel (c) of Table 4.[14] They further confirm the pattern of all previous regressions. Only children from low socio-economic backgrounds stand to gain from their teachers' intentions to engage them in learning. The point estimate suggests that a one-standard-deviation increase in the ESL scale is associated with a .063-standard-deviation increase in test scores. This effect is 34% larger than in the baseline model and 17% larger than in the specification without students who are taught by the same teacher in all three subjects.

[13] I chose 15 students as the cut-off point after visual inspection of the frequency distribution of class sizes. It becomes much denser starting with classes of 16 students. In total, there are 246 students in classes with 16 students, while there are only 90 in classes with 15 students.

[14] Note that I multiplied the factor score by -1 due to the way the questions are coded in the teacher questionnaire where a higher value signifies *less* frequent use of the technique in question. This manipulation makes the results easier to understand and compare to the other estimates.

Table 4

Robustness of the Estimated Effect of Teachers' Intentions to Engage Students on Standardised Test Scores for different subgroups, Student Fixed Effects Models

	Full Sample	Socio-economic background		Gender		Language mostly spoken at home	
		High	Low	Boys	Girls	German	not German
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(a)							
Excluding small classes (< 16 students)	-.000 (0.01)	-.016 (0.02)	.047* (0.03)	.006 (0.02)	.004 (0.02)	.008 (0.02)	-.012 (0.03)
Number of observations	9,915	4,203	3,207	4,953	4,962	7,164	1,794
(b)							
Only students with variation in teachers	.004 (0.02)	-.005 (0.02)	.054* (0.03)	.007 (0.02)	.011 (0.02)	.010 (0.02)	-.016 (0.03)
Number of observations	5,187	2,253	1,548	2,526	2,661	3,918	870
(c)							
Treatment based on factor analysis	-.004 (0.02)	-.025 (0.02)	.063** (0.03)	-.002 (0.02)	.005 (0.02)	.013 (0.02)	-.033 (.043)
Number of observations	10,122	4,250	3,254	5,072	5,050	7,313	1,850

Notes. * p < 0.10, ** p < 0.05, *** p < 0.01. Table shows regressions of students' z-standardized achievement scores on teachers' z-standardised values on the ESL scale. Fixed effects are at the student level. Each regression also contains subject fixed effects. In the upper panel, only the level of teachers' values on the ESL scale are considered. In the lower panel, the levels and squared terms of values on the ESL scales enter the models. All regressions are weighted by the students' sampling probability. Standard errors in parentheses are clustered at the classroom level. Control variables are listed in Table A.3. Data source: TIMSS/PIRLS 2011.

The fourth robustness check concerns the question whether my results are contaminated by subject-specific sorting and self-selection in some schools. It makes use of information from the school background questionnaire of the TIMSS and PIRLS studies. School principals are asked about the total number of students enrolled in grade 4 in their school. It turns out that the smallest school has only 6 students in grade 4, the largest 158. I use this information to separately estimate models for large and small schools. The cut-off point for being a small school is 31 students, which is equivalent to the largest number of students in one classroom in my data. This gives me one group of 1,311 students in

small schools and a significantly bigger group of 8,592 students in large schools. I expect that this procedure lends further credibility to my identification strategy, as there should be much less room for tracking in small schools. In fact, most schools should only have one classroom per grade. If there were significant within-school sorting on unobservables, I would expect that the results of the two groups starkly differ from one another. More precisely, if there were positive sorting of students to teachers by subject-specific ability and teaching practices, one would expect the estimate for large schools to be larger than the one for small schools (and vice versa if there were negative sorting). As the sample size gets rather small as a result of splitting the sample, I am not able to perform subsample analysis, e.g. for children from different socio-economic backgrounds. However, my main interest here lies in finding out whether or not my analysis *generally* suffers from omitted variable bias. Against this background, the results shown in columns 1 and 2 of Table 5 are encouraging. Neither the estimate for small schools nor the estimate for large schools turns out significant. This confirms the results depicted in Table 2 and suggests that my results are not biased by sorting within schools.

To further underscore this claim, I performed a fifth robustness check, which again splits the sample by schools that are more likely to sort students and schools that are less likely to do so. Again, the data is provided by school principals, who are asked a total of five questions about mainly teacher, student and parent expectations regarding academic success in their schools.[15] The answers are used to construct a so-called School Emphasis on Academic Success Index. This index has three different categories: very high emphasis, high emphasis, and medium emphasis. Note that more than 99% of all students visit a school that falls into one of the two latter categories. For practical reasons, I added all students going to a school with a very high emphasis on academic success to those visiting a school with a high emphasis. The resulting dichotomous indicator shows that more than two thirds of all students for whom there is information go to a (very) high emphasis school. My prior is that the former should be more prone to forming special ability groups and classrooms and, therefore, to endogenous sorting and selection. If the results were driven by such sorting, they should be different from the rest among these schools. Yet, both estimates turn out insignificant again.

[15] The five items belong to question 12 of the school context questionnaire, which had to be answered on a five-point-scale that ranges from very high to very low.

Table 5

Estimated Effect of Teachers' Intentions to Engage Students on Standardised Test Scores by Size of Grade 4 and Emphasis on Academic Success, Student Fixed Effects Models

	Grade size		Emphasis on academic success	
	Large (1)	Small (2)	(Very) High (3)	Medium (4)
(a)				
ESL	-.008	.044	.014	.064
	(0.02)	(0.04)	(0.02)	(0.09)
Number of observations	8,592	1,311	6,825	2,952

Notes. * p < 0.10, ** p < 0.05, *** p < 0.01. Table shows regressions of students' z-standardized achievement scores on teachers' z-standardised values on the ESL scale. Fixed effects are at the student level. Each regression also contains subject fixed effects. In the upper panel, only the level of teachers' values on the ESL scale are considered. In the lower panel, the levels and squared terms of values on the ESL scales enter the models. All regressions are weighted by the students' sampling probability. Standard errors in parentheses are clustered at the classroom level. Control variables are listed in Table A.3. Data source: TIMSS/PIRLS 2011.

5. Conclusion

In this paper, I have investigated the effects of teachers' intentions to engage students in learning on achievement outcomes as measured by standardised assessment studies. The object of my analysis was a nationally representative sample of fourth-graders in Germany. Teachers' intentions to engage their students were measured by the ESL scale as supplied by the IEA in connection with the TIMSS and PIRLS studies 2011. It is based on questions regarding how often teachers use questioning in class, bring interesting materials to the course, relate the course content to students' daily lives, give praise and encouragement and summarize the most important points of the lesson. The main finding is that potentially engaging teaching practices can yield small to modest achievement gains for students from low socio-economic backgrounds. In the full sample, no significant effects could be detected.

My identification strategy, which is based on within-student variation, reliably rules out unobserved heterogeneity stemming from individual or school-level characteristics. However, it has certain limitations regarding teacher sorting into specific teaching practices, which should be borne in mind when interpreting the results. Yet, the relative position of students from low socio-economic backgrounds as compared to students from high socio-economic backgrounds is much more likely to reflect a true causal mechanism then the full sample results. The reason for this is that any overall teacher-related bias would affect both groups in the same way unless sorting of teachers into teaching practices along unobserved characteristics is particularly beneficial or detrimental exclusively to particular subgroups.

From a policy perspective, the results of the present analysis can be understood as a possible vehicle to achieve greater equality of opportunity. After all, children from low socio-economic backgrounds stand to gain from engaging teaching practices. What is more, their gain does not come at the

expense of children from high socio-economic backgrounds who are not affected by these teaching practices. However, from an efficiency perspective, one would also have to assess the costs of implementing more engaging teaching practices in schools across the country. Yet, especially for future teachers, they would not seem to be prohibitively high, as they would mainly arise from slightly altering the focus of teacher training.

The results of this paper open up a number of fruitful avenues for future research. First of all, a lot remains unknown about what classroom actions are effective in conferring skills upon students. This is related to the question of teachers' time allocation between different teaching practices, as it is likely that a mix of different actions generates the best results. Secondly, this paper has shown that not all teaching practices need to be equally effective for all students. Against this background, more subgroup-specific analysis would be desirable. Of course, the feasibility of this hinges upon the provision of better data. For instance, the present work would have vastly benefited from subject-specific information on the use of potentially engaging teaching practices, as this would have allowed within-teacher estimations, which in turn would have generated more definite conclusions on the causal nature of the results.

Annex

Table A.1

Pairwise Correlations between Observable Teacher Characteristics and Teacher Scores on the ESL Scale

	Pearson's r
Objective characteristics	
Experience	0.24***
Sex	-0.00
Age	0.20***
Education	0.00
Field teacher	0.05
Interactions with other teachers	
Discuss how to teach a particular subject	0.20***
Collaborate in planning and preparing materials	0.05
Share teaching experiences	0.29***
Visit other classrooms to learn	0.12**
Work together to try out new ideas	0.17***
Job satisfaction	
Content with teaching profession	-0.07
Satisfied being a teacher at this school	-0.08
Had more enthusiasm when I began teaching	0.09
Do important work as a teacher	-0.06
Plan to continue as a teacher for as long as I can	0.07
Frustrated as a teacher	0.07
Relation to parents	
Individually discuss learning progress	-0.16***
Send home a progress report	0.01
Use of computers	
for preparation	-0.06
for administration	0.11*
for classroom instruction	-0.14**

* p < 0.10, ** p < 0.05, *** p < 0.01. Data source: TIMSS/PIRLS 2011.

Table A.2

Descriptive Statistics – Test Scores and ESL Scale

	Test scores	ESL scale
Mean	533.8	8.73
SD between students	64.7	1.36
SD within students	31.5	0.56

Data source: TIMSS/PIRLS 2011.

Table A.3

Summary Statistics and Definitions of Covariates

		Math		Science		Reading	
Variable	Definition	Mean	SD	Mean	SD	Mean	SD
Educational Achievement	Z-score based on first plausible value in dataset	0.01	1.00	0.01	0.99	0.00	0.99
Student and school controls							
Sex	Equals 1 for boys and 0 for girls	0.50	0.50	0.50	0.50	0.50	0.50
Age	in months	124.3	6.05	124.3	6.05	124.3	6.05
Test language spoken at home 1	Equals 1 if sometimes or never, 0 if always, almost always, or missing	0.18	0.38	0.18	0.38	0.18	0.38
Test language spoken at home 2	Equals 1 if always or almost always, 0 if sometimes, never, or missing	0.73	0.44	0.73	0.44	0.73	0.44
Parents ask about learning 1	Equals 1 if less than once or twice a week, 0 if more often or missing	0.11	0.31	0.11	0.31	0.11	0.31
Parents ask about learning 2	Equals 1 if at least once or twice a week, 0 if less often or missing	0.80	0.40	0.80	0.40	0.80	0.40
Socio-economic background (used for subsample analysis)	Equals 1 if at least one parent has a post-secondary degree, 0 if not	1.96	2.41	1.96	2.41	1.97	2.42
School size	No. of students in fourth grade	56.8	25.3	56.8	25.3	56.8	25.3
No. of computers	No. of computers in the entire school	14.9	9.22	14.9	9.22	14.9	9.22
School composition by student background	Scale from 1 (more affluent) to 3 (more disadvantaged)	2.02	0.64	2.02	0.64	2.02	0.64
No. of people living in school area	Scale from 1 (more than 500,000) to 6 (3,000 or fewer)	3.93	1.58	3.93	1.58	3.93	1.58

School emphasis on academic success	Scale ranging from 1 (very high emphasis) to 3 (medium emphasis)	2.30	0.46	2.30	0.46	2.30	0.46
School discipline and safety	Scale ranging from 1 (hardly any problems) to 3 (moderate problems)	1.62	0.56	1.62	0.56	1.62	0.56
Teacher and teaching-related characteristics							
Instruction affected by resource shortages	Scale provided by TIMSS/PIRLS (by subject)	10.59	1.55	10.65	1.55	10.63	1.59
Class size	No. of students in class	21.7	3.84	21.7	3.84	21.7	3.84
Instructional time	Weekly instructional time in minutes	249.1	52.4	134.3	77.2	373.4	151.3
Teaching experience	in years	19.3	12.6	18.6	12.5	18.8	12.4
Teacher sex	Equals 1 if teacher is female, 0 if teacher is male	0.80	0.40	0.89	0.31	0.92	0.27
Teacher age	Four categories (1 - under 30; 2 - 30-39; 3 - 40-49; 4 - over 49)	2.98	1.05	2.91	1.06	2.94	1.04
Teacher education 1	Equals 1 if teacher has no tertiary education, 0 if yes or missing	0.07	0.25	0.06	0.24	0.06	0.25
Teacher education 2	Equals 1 if teacher has tertiary education, 0 if not or missing	0.90	0.30	0.90	0.31	0.88	0.33
Field teacher 1	Equals 1 if teacher has not majored in the subject taught, 0 if yes or missing	0.46	0.50	0.37	0.48	0.14	0.35
Field teacher 2	Equals 1 if teacher has majored in the subject taught, 0 if not or missing	0.53	0.50	0.60	0.49	0.83	0.38

Interactions with other teachers							
Discuss how to teach a particular subject	Scale ranging from 1 (never or almost never) to 4 (daily or almost daily)	2.36	0.86	2.36	0.88	2.37	0.88
Collaborate in planning and preparing materials	Scale ranging from 1 (never or almost never) to 4 (daily or almost daily)	2.38	0.79	2.42	0.79	2.44	0.80
Share teaching experiences	Scale ranging from 1 (never or almost never) to 4 (daily or almost daily)	2.48	0.89	2.50	0.89	2.51	0.92
Visit other classrooms to learn	Scale ranging from 1 (never or almost never) to 4 (daily or almost daily)	1.14	.042	1.13	0.40	1.12	0.37
Work together to try out new ideas	Scale ranging from 1 (never or almost never) to 4 (daily or almost daily)	1.95	0.81	2.02	0.75	1.97	0.80
Job satisfaction							
Content with teaching profession	Scale ranging from 1 (agree a lot) to 4 (disagree a lot)	1.62	0.65	1.58	0.62	1.56	0.61
Satisfied being a teacher at this school	Scale ranging from 1 (agree a lot) to 4 (disagree a lot)	1.42	0.54	1.44	0.55	1.40	0.54
Had more enthusiasm when I began teaching	Scale ranging from 1 (agree a lot) to 4 (disagree a lot)	2.59	1.07	2.56	1.00	2.58	1.04
Do important work as a teacher	Scale ranging from 1 (agree a lot) to 4 (disagree a lot)	1.13	0.37	1.14	0.36	1.14	0.36
Plan to continue as a teacher for as long as I can	Scale ranging from 1 (agree a lot) to 4 (disagree a lot)	1.59	0.82	1.58	0.81	1.58	0.81
Frustrated as a teacher	Scale ranging from 1 (agree a lot) to 4 (disagree a lot)	3.31	0.73	3.38	0.67	3.38	0.67
Relation to parents							
Individually discuss	Scale ranging from	3.46	0.70	3.49	0.70	3.39	0.73

learning progress	1 (at least once a week) to 5 (never)						
Send home a progress report	Scale ranging from 1 (at least once a week) to 5 (never)	4.46	0.69	4.44	0.71	4.41	0.74
Use of computers							
for preparation	Equals 1 if yes, 0 otherwise	0.98	0.15	0.97	0.18	0.97	0.16
for administration	Equals 1 if yes, 0 otherwise	0.84	0.37	0.84	0.37	0.84	0.37
for classroom instruction	Equals 1 if yes, 0 otherwise	0.75	0.43	0.75	0.43	0.77	0.42

Data source: TIMSS/PIRLS 2011.

Table A.4

Estimated Effect of Teachers' Intentions to Engage Students on Student Achievement; Two Subjects at a Time

	Math + Science (1)	Math + Reading (2)	Science + Reading (3)
(a)			
ESL	.001	-.022	.010
	(0.01)	(0.04)	(0.08)
Subject dummies	yes	yes	yes
Teacher and teaching characteristics	yes	yes	yes
Number of observations	6,826	6,826	6,826

Notes. * p < 0.10, ** p < 0.05, *** p < 0.01. Table shows regressions of students' z-standardized achievement scores on teachers' z-standardised values on the ESL scale. Fixed effects are at the student level. Each regression also contains subject fixed effects. In the upper panel, only the level of teachers' values on the ESL scale are considered. In the lower panel, the levels and squared terms of values on the ESL scales enter the models. All regressions are weighted by the students' sampling probability. Standard errors in parentheses are clustered at the classroom level. Control variables are listed in Table A.3. Data source: TIMSS/PIRLS 2011.

Literature

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in Chicago public high schools. *Journal of Labour Economics, 25*, 95 – 135.

Ammermueller, A. & Pischke, J.-S. (2009). Peer Effects in European Primary Schools: Evidence from the Progress in International Reading Literacy Study. *Journal of Labor Economics, 27*, 315 – 348.

Ashenfelter, O. & Zimmerman, D.J. (1997). Estimates of the Returns to Schooling from Sibling Data: Fathers, Sons, and Brothers. *The Review of Economics and Statistics, 79*, 1 – 9.

Aslam, M. & Kingdon, G. (2011). What can teachers do to raise pupil achievement? *Economics of Education Review, 30*, 559 – 574.

Bietenbeck, J. (2014). Teaching practices and cognitive skills. *Labour Economics, 30*, 143 – 153.

Blazar, D. (2015). Effective teaching in elementary mathematics: Identifying classroom practices that support student achievement. *Economics of Education Review, 48*, 16 – 29.

Bloom, B.S. (Ed.) (1956). *Taxonomy of Educactional Objectives, Handbook 1: Cognitive Domain*. New York: Longmans, Green and Co.

Bos, W., Wendt, H., Koeller, O., & Selter, C. (2012a). *TIMSS 2011. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Muenster/New York/Munich/Berlin: Waxmann.

Bos, W., Tarelli, I., Bremerich-Vos, A., & Schwippert, K. (2012b). *IGLU 2011. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Muenster/New York/Munich/Berlin: Waxmann.

Chamberlain, G. (1982). Multivariate Regression Models for Panel Data. *Journal of Econometrics, 18*, 5 – 46.

Chamberlain, G. (1984). Panel Data. In: Z. Griliches & M.D. Intriligator (Eds.), *Handbook of Econometrics. Volume 2*. 1247 – 1318. Amsterdam: North Holland.

Chetty, R., Friedman, J.N., & Rockoff, J. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review, 104*, 2633 – 2679.

Clotfelter, C.T., Ladd, H.F., & Vigdor, J.L. (2010). Teacher Credentials and Student Achievement in High School: A Cross-Subject Analysis with Student Fixed Effects. *Journal of Human Resources, 45*, 655 – 681.

Clotfelter, C.T., Ladd, H.F., & Vigdor, J.L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *The Journal of Human Resources, 41*, 778 – 820.

Dee, T. (2005). A Teacher Like Me: Does Race, Ethnicity, or Gender Matter? *American Economic Review, 95*, 158 – 165.

Falck, O., Mang, C., & Woessmann, L. (2015). Virtually No Effect? Different Uses of Classroom Computers and their Effect on Student Achievement. *IZA DP No. 8939*.

Foy, P. (2013). TIMSS and PIRLS 2011 User Guide for the Fourth Grade Combined International Database. Boston: TIMSS & PIRLS International Study Center.

Fredricks, J.A., Blumenfeld, P.C., & Paris, A.H. (2004). School Engagement: Potential of the Concept, State of the Evidence. *Review of Educational Research, 74*, 59 – 109.

Goldhaber, D.D. & Anthony, E.A. (2007). Can teacher quality be effectively assessed? *The Review of Economics and Statistics, 89*, 134 – 150.

Goldhaber, D.D. & Brewer, D.J. (1997). Why Don't Schools and Teachers Seem to Matter? Assessing the Impact of Unobservables on Educational Productivity. *The Journal of Human Resources, 32*, 505 - 523.

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

Hanushek, E.A. & Rivkin, S.G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review, 100*, 267 – 271.

Hattie, J. (2009). *Visible Learning. A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. New York: Routledge.

Heineck, G. & Riphahn, R.T. (2009). Intergenerational Transmission of Educational Attainment in Germany – The Last Five Decades. *Journal of Economics and Statistics, 229*, 36 – 60.

Hidalgo-Cabrillana, A. & López-Mayan, C. (2015). Teaching Styles and Achievement: Student and Teacher Perspectives. Working Paper 2/2015. Economic Analysis Working Paper Series. Universidad Autónoma de Madrid.

Hill, L.G. & Werner, N.E. (2006). Affiliative Motivation, School Attachment and Aggression in School. *Psychology in the Schools, 43*, 231 – 246.

Hooper, M., Mullis, I.V.S., & Martin, M.O. (2013a). TIMSS 2015 Context Questionnaire Framework. In: I.V.S. Mullis & M.O. Martin (Eds.), *TIMSS 2015 Assessment Frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Hooper, M., Mullis, I.V.S., & Martin, M.O. (2013b). PIRLS 2016 Context Questionnaire Framework. In: I.V.S. Mullis & M.O. Martin (Eds.), *PIRLS 2016 Assessment Framework*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

IEA (2011a). TIMSS 2011. Teacher Questionnaire. Retrieved from http://timssandpirls.bc.edu/timss2011/international-contextual-q.html.

IEA (2011b). PIRLS 2011. Teacher Questionnaire. Retrieved from http://timssandpirls.bc.edu/pirls2011/international-contextual-q.html.

Joncas, M. & Foy, P. (2013). Sample design in TIMSS and PIRLS. In: M.O. Martin & I.V.S. Mullis (Eds.), *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Kane, T.J., Taylor, E.S., Tyler, J.H., & Wooten, A.L. (2011). Identifying Effective Classroom Practices Using Student Achievement Data. *Journal of Human Resources, 46*, 587 – 613.

Klaveren, C. van (2011). Lecturing style teaching and student performance. *Economics of Education Review, 30*, 729 – 739.

Koedel, C., Mihaly, K., & Rockoff, J.E. (2015). Value-added modeling: A review. *Economics of Education Review, 47*, 180 – 195.

Lavy, V. (2015). Do differences in schools' instruction time explain international achievement gaps? Evidence from developed and developing countries. *The Economic Journal, 125*, F397 – F424.

Lavy, V. (2011). What makes an effective teacher? Quasi-experimental evidence. *NBER Working Paper 16885*.

Lorena Comi, S., Argentin, G., Gui, M., Origo, F., & Pagani, L. (2017). Is it the way they use it? Teachers, ICT and student achievement. *Economics of Education Review, 56*, 24-39.

Martin, M.O., Mullis, I.V.S., Foy, P., & Stanco, G.M. (2012). *TIMSS 2011 International Results in Science*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martin, M.O. & Mullis, I.V.S. (Eds.). (2013). TIMSS and PIRLS 2011: Relationships Among Reading, Mathematics, and Science Achievement at the Fourth Grade – Implications for Early Learning. Boston: TIMSS & PIRLS International Study Center.

Martin, M.O. & Mullis, I.V.S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

McCoy, S., Smyth, E., & Banks, J. (2012). The Primary Classroom: Insights from the Growing Up in Ireland Study. Learning in Focus. The Economic and Social Research Institute. Dublin.

McLaughlin, M., McGrath, D.J., Burian-Fitzgerald, A., Lanahan, L., Scotchmer, M., Enyeart, C., & Salganik, L. (2005). Student content engagement as a construct for the measurement of effective classroom instruction and teacher knowledge. American Institutes for Research. Retrieved from http://www.air.org/sites/default/files/downloads/report/AERA2005Student_Content_Engagement11_0.pdf.

Metzler, J. & Woessmann, L. (2012). The Impact of Teacher Subject Knowledge on Student Achievement: Evidence from Within-Teacher Within-Student Variation. *Journal of Development Economics, 99*, 486 – 496.

Mullis, I.V.S., Martin, M.O., Foy, P., & Arora, A. (2012a). *TIMSS 2011 International Results in Mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I.V.S., Martin, M.O., Foy, P., & Drucker, K.T. (2012b). *PIRLS 2011 International Results in Reading*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mundlak, Y. (1978). On the Pooling of Time Series and Cross Section Data. *Econometrica, 46*, 69 – 85.

Murnane, R.J. & Phillips, B. (1981). What do Effective Teachers of Inner-City Children Have in Common? *Social Science Research, 10*, 83 - 100.

Paredes, V. (2014). A teacher like me or a student like me? Role model versus teacher bias effect. *Economics of Education Review, 39*, 38 – 49.

Perdue, N.H., Manzeske, D.P., & Estell, D.B. (2009). Early Predictors of School Engagement: Exploring the Role of Peer Relationships. *Psychology in the Schools, 46*, 1084 – 1097.

Piopiunik, M. & Schlotter, M. (2012). Identifying the Incidence of "Grading on a Curve": A Within-Student Across-Subject Approach. Ifo Working Paper No. 212. Munich.

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

Schwerdt, G. & Wuppermann, A.C. (2011). Is traditional teaching really all that bad? A within-student between-subject approach. *Economics of Education Review, 30*, 365 – 379.

Seidel, T. & Shavelson, R.J. (2007). Teaching Effectiveness Research in the Past Decade: The Role of Theory and Research Design in Disentangling Meta-Analysis Results. *Review of Educational Research, 77*, 454 – 499.

Shernoff, D.J. (2013). Optimal Learning Environments to Promote Student Engagement. New York: Springer.

Tyler, J.H., Taylor, E.S., Kane, T.J., & Wooten, A. (2010). Using Student Performance Data to Identify Effective Classroom Practices. *American Economic Review, 100*, 256 – 260.

Woessmann, L. (2016). The Economic Case for Education. *Education Economics, 24*, 3 - 32.

Woessmann, L. (2003). Schooling Resources, Educational Institutions, and Student Performance: The International Evidence. *Oxford Bulletin of Economics and Statistics, 65*, 117 - 170.