LUND UNIVERSITY
School of Economics and Management

More is more?

The effect of increased instruction time on students' performance – evidence from TIMSS 2019

Student: Felix Persson

Supervisor: Jan Bietenbeck

Course: NEKN01, Essay 1, Spring 2021

2021-08-18

Abstract

Students performs differently in international comparable school evaluations such as TIMSS and PISA. One possible explanation discussed in economic research literature is that differences in the amount of instruction time across countries are the reason for international gaps in student performance. Using a within-student between-subject fixed effects model and the data from the TIMSS 2019 evaluation, I find a positive and statistically significant effect for increased instruction time on students' performance on test scores in math and science in high income countries which is in line with what previously papers has shown. I also find that students with different background characteristics benefit differently from increased instruction time. However, when extending the analysis to other groups of countries and using different types of weights, the effect seems to be lower and statistically insignificant. I conclude that although the effect of increased instruction time on students' performance is positive in high income countries, the effect seems to vary between countries.

Keywords: instruction time, student achievement, TIMSS, teacher characteristics, PISA

Table of contents

1. Introduction

How well students perform in school is important for students themself, both in respect of understanding the world and in respect of earnings since higher education can lead to higher income (OECD, 2015). It is also important for a country since high education levels can be of importance for growth and for development of societies (Goldin & Katz, 2009).

For many countries, education is one of the biggest government expenditures (OECD, 2015). However, it is notable that educational systems are different in several aspects between countries and students´ performance in international comparable tests also differ. The reasons for these performance differences are the subject of a research literature in economics, and one possible explanation discussed in this literature is that differences in the amount of instruction time across countries are the reason for international gaps in student performance (see for example Lavy (2015), Rivkin & Schiman (2018) and Bingley et al (2018)).

One of the more influential papers on the subject was written by Lavy (2015), and he found a positive and statistically significant impact of increased weekly instruction time on test scores, using a within-student between-subject fixed effects model with data from the PISA 2006 evaluation, which is an international evaluation conducted by OECD where thousands of students participate by taking comparable tests and by providing information on several characteristics. Lavys estimated baseline effect is that one additional hour of weekly instruction time increases test scores by 0.058 standard deviations (SD).

In this essay, I use a similar fixed effects model to the one that Lavy used and investigate the same key question, whether increased weekly instruction time effects students' performance in math and science, but applying it on new data from the TIMSS 2019 evaluation (which is another international evaluation like PISA)[1]. Assuming that students are not sorted into classes based on their previous performance in math or science, the within-student between-subject fixed effects model allows for controlling if differences between students' performance in these subjects are systematically correlated with differences in instruction time between the subjects, since both observed and unobserved characteristics are fixed for each student but instruction time differs.

[1] More information about TIMSS follows in the section "Data".

I will also extend the model to include interaction variables with teacher characteristics, to examine whether the effect of instruction time differs by different teacher characteristics. As a further extension, I will also perform the baseline regression on specific subsets of students depending on certain socioeconomic and background characteristics. As a last extension I will change the sample of countries to see if the baseline results hold. We know from previous literature, for example Lavy (2015), that the effect can vary between different types of countries. My main focus will be on high income countries since it is more homogenous group of countries, and their education systems are therefore more alike.

Using a similar model as Lavy (2015) and the data from TIMSS 2019, I find positive and statistically significant result for instruction time on test results for the baseline specification, with my baseline coefficients being 0.0137 SD without control variables and 0.0117 SD with control variables. These coefficients are lower than what Lavy (2015) presented, but more in line with other previous studies who also studies the effects of instruction time on test scores and get positive and statistically significant results but smaller coefficients than Lavy (2015).[2]

When interacting with teacher characteristics I use variables for professional development, teachers' degrees and experience. My results for teacher characteristics show that for professional development there is a negative but statistically insignificant effect when interacting with instruction time, therefore no strong conclusions can be drawn about the additional effect of professional development. For experience the coefficient is positive although statistically insignificant as well. When it comes to teachers who has majored in a relevant subject for their teaching there is a positive and statistically significant effect although the coefficient is low, suggesting that students with a teacher that has majored in the subject that they teach performs slightly better.

The results for different subgroups of students indicates that students benefit differently from increased instruction time depending on background characteristics and socioeconomic factors. For example, girls have a smaller coefficient than boys, indicating that boys benefit more from increased instruction time, which is the opposite of what Lavy (2015) finds but the same result as Bingley et al (2018) get. When comparing results between students who has guardians with different education levels, I find that students with guardians that has a lower education level

[2] A literature review with examples follows this section.

benefit more from instruction time than students with parents who has high education levels. Lavy (2015) gets a similar result. The results when looking at students who are first- and second-generation immigrants are also similar with what Lavy (2015) presents. Second generation immigrants seem to benefit more from increased instruction time since the coefficient is larger than for first generation immigrants, although they are both statistically insignificant. When changing the sample of countries, the baseline results are positive and statistically significant for high income countries, but statistically insignificant although positive for OECD countries, all countries (the whole sample) and for non-high-income countries. For the non-high income countries, the coefficient is negative.

Apart from various extensions, I also perform various robustness checks to see if the results hold in different settings. I find that changing weights (or not using weighs at all) changes the baseline results in some settings. The results for high income countries are positive when using senate weights and when not using any weights at all but negative, although statistically insignificant, when using total weights. My interpretation of these results is that although the results are positive and statistically significant in several contexts, there are differences between countries, i.e., the results does not hold for all countries included in the sample.
I also change the way I create the variable for test scores, resulting in only minor differences in the baseline coefficient which suggests that the original way of using the test score values is adequate.

The disposition of the essay will be as follows. In the literature review I will present some of the results that previous similar studies have shown. In the section “Data” I will present more information on TIMSS in general and on the variables that will be used. After that, the empirical strategy and models will be covered in more detail. In the sections that follows, the results will be presented and discussed. “Conclusion” will conclude and summarize.

2. Literature review

Since Lavy (2015) published his paper, others have both replicated and extended his model in various directions. Bingley et al (2018) performs a similar study but in a Danish setting with data from Denmark both for both instruction time and for characteristics. They do not use instruction time for one year which is done with data from PISA or TIMSS (for example 8th or 9th grade), but they use accumulated instruction time throughout the whole education (year 1 to

9). They find positive effects on instruction time on test scores with a coefficient of 0.0360 SD, which is lower than what Lavy (2015) finds. They also find bigger effect for boys than for girls and for students that are defined as being socioeconomic disadvantaged as well as for immigrants.

Rivkin & Schiman (2015) uses a similar model but data from PISA 2009, and apart from the effect of instruction time on test scores, they also examine whether classroom quality effects the student achievements. They find positive effects on both accounts, i.e., a positive effect on instruction time on test scores and that better classroom quality yields better student performance. Their baseline coefficient for instruction time on test scores is 0.0230 SD which is also smaller than Lavy (2015).

A replication and extension of Lavys paper has also been conducted, including both PISA and TIMSS data (with 2015 being the latest for TIMSS). The replication using the PISA 2006 data yields the same result as for Lavy (2015), but they also extend their analysis to include all available PISA studies following the one conducted in 2006 and find that the average effect for the following years is only 0.022 SD for the baseline case using the same model. This result, combined with the fact that other studies also get a lower coefficient for the baseline regressions, indicates that the effect in more recent years might not be as high as Lavy originally suggested (Bietenbeck & Collins, 2020).

By using a similar model as the studies mentioned above but new data, this essay complements the current literature which examines the effect of instruction time on student performance in international evaluations. By using similar models, the results become more comparable between studies. To my knowledge, there has not been any papers published using data from TIMSS 2019.

3. Data

3.1. TIMSS

The data comes from the 2019 Trends in International Mathematics and Science Study (TIMSS), which is an international assessments of student achievement in mathematics and science conducted by the International Association for the Evaluation of Educational Achievement, IAE. TIMSS has been conducted every fourth year since the start in 1995. In the

2019 assessment, students in both 4th and 8th grade from totally 54 countries participated. (TIMMS, 2019)

The tests contain a large variety of questions and therefore students only answer a subset of the questions, called booklets. The content of the tests is about the same despite which version of the test you take. For example, the math part for eight grade students consists of numbers (30%), algebra (30%), geometry (20%) and data & probability (20%). The science part consists of biology (35%), physics (25%), chemistry (20%) and earth science (20%). This procedure ensures that the test scores are comparable between students even if they have used different booklets (TIMMS, 2019).

The students answers to the tests results in five plausible values per student and subject, and I will use the first plausible value for my test score variable since that is standard in the literature and also what Lavy (2015) uses. The plausible values are not actual test scores, but rather estimations based on both the overall achievements and on contextual information. They are used so that valid comparisons on the group level can be done (Martin & Mullis, 2017). In addition to the tests themselves, there are questionnaires containing questions about attitudes towards the subject, teachers experience level, instruction time for each subject, socioeconomic information etc. that are completed by students, guardians, teachers and principals (TIMMS, 2019).

The schools and classes that participate in TIMSS are selected in a way that should ensure that the sample is representative for the country. The process has two stages, one for choosing the school and one for choosing the classes. The schools and classes are chosen with probability proportional to school size which should results in student samples with equal selection probabilities (Martin & Mullis, 2017). Since there might be some selected schools that after selection do not participate for some reason, the students in the sample might not have equal selection probability in practice. To overcome this, I use senate weights which can be used when countries form the unit of analysis, like high income countries in this case (Jerrim et al, 2017). Senate weights gives every country the same weight. Lavy (2015) does not use any weights, but Jerrim et al argue that weights should be used when using international educational achievement data such as TIMSS to minimize the risk of the results or particular characteristics from some students or schools being either under or overrepresented within the analysis (Jerrim et al, 2017).

3.2. Dataset and variables

TIMSS is conducted for both 4th and 8th grade students. I restrict my sample to the eight grade students, since eight grade students are about the same age as the students participating in PISA studies, which is what Lavy (2015) uses. This is also what most other previous studies has focused on, which makes my results more relatable to those studies as well.

To be able to use a fixed effect model like the one used by Lavy (2015), I wanted a balanced cross sectional panel with two observation per student - one for math and one for science. Students in classes that only contained information about one of the subjects were therefore dropped. As in any study with self-reporting there were some observations with values that did not make sense, for example there were a few observations that answered that they had one minute of instruction time per week (when they presumably meant one hour per week). To avoid outliers to give misleading results, I restricted my sample to only include observation with at least 30 minutes instruction time and at most 450 minutes per subject. To be able to do various extensions from the baseline I also needed all my observations to contain information on characteristics. This left me with a sample containing in total 39 countries, 585 schools, 2´532 classes and 181´024 students, providing 362´048 observations since each student is observed twice. Out of these observations, 228´258 were high income according to the World Banks definition of high income which is defined as having a GNI per capita larger than 12´536 USD (World Bank, 2021).

For the baseline regressions the variables I use are students test score, which are based on the 1st plausible values, and instruction time. The dependent variable is the test score that the student performs. Using the 1st plausible value for each subject as the variable for test score is what Lavy (2015) does as well, and the difference on the result of using the 1st or all five plausible values are usually trivial (Jerrim et al, 2017)[3]. The test scores are then standardized, so that the mean is 0 for all students and the standard deviation is 1. This is in line with previous literature such as Lavy (2015), and the implication is that a student with a positive standardized test score has a test score that is above the mean and vice versa.

[3] To make sure that this is true also in this case, I also perform the baseline regression with the test score value being based on all five plausible values as a robustness check. The results shows that the differences between the two methods are indeed very small. The results are presented in section “Robustness check”.

The main independent variable is instruction time. As a part of the teacher's questionnaire, the time spent on instruction time, submitted as minutes per week, is reported for each subject by the respective teacher. Some students have more than one teacher per subject. This is especially common for science since some countries does not teach ”science” as an integrated subject but rather divide it into different subjects: physics, biology, chemistry and earth science. Therefore, I summed the amount of instruction time for each teacher, generating one observation for each subject (math and science) irrespective of the number of teachers the student had. I also converted the answers from minutes to hours per week.

To explore if instruction time has different effects on test scores depending on the characteristics of the teacher, I also included some variables exploiting this. The variable "prof. development", which stands for professional development, is a dummy variable that takes the value 1 if the teachers during the last two years have taken part in some sort of activity involving professional development regarding the content of the subject, pedagogy and curriculum. If not, it takes the values 0. Teachers experience is also being controlled for with the variable "experienced", which takes the value 1 if the teacher has more than 10 years of experience. "Education (major)" takes the value 1 if the teacher has majored in the subject that they are teaching.

To evaluate if instruction time effects different students test results in different ways, I created some new variables so that I could run the baseline regression for certain subgroups of students. The variable "1st gen. immigrant" takes the value 1 if the student not born in the (high-income) country which the test was conducted in and "2nd gen. immigrant" takes the value 1 if both students' guardians not born in the (high-income) country which the test was conducted in. "Parents high degree" takes the value 1 if both guardians have a degree on at least ISCED level 3, which is equal to a degree in high school-level and "low degree" if none of the guardians has a degree on at least high school level. "Female student" takes the value 1 if the student is female.

The variables and their summary statistics are presented in Table 1 below. They will all be used in the analysis and more results are presented in part 4. of the essay. Again, it is worth noting that the variable for test score is standardized for the whole sample so that the mean is 0 and a standard deviation of 1. When only looking at the high income countries, we see that the mean is 0.24, i.e. above the total mean.

Table 1: **Summary statistics**

	Mean	Std. Dev.	Min	Max	N
Instruction time	3.785	1.234	.5	9	228 858
Test score	0.24	0.95	-4.3	3.9	228 858
Female student	0.494	0.50	0	1	228 858
Experienced	0.643	0.479	0	1	228 858
Education (major)	0.871	0.336	0	1	228 858
Prof. development	0.332	0.471	0	1	228 858
Parents high degree	0.453	0.498	0	1	228 858
1st gen. immigrant	0.129	0.336	0	1	228 858
2nd gen. immigrant	0.167	0.373	0	1	228 858

Notes: High income countries. Instruction time: hours per week. Test score: standardized.

The average amount of instruction time is about 3.8 hours per week for high income countries, and the minimum and maximum are 0.5 and 9 since that was my restriction on order to get rid of outliers (suspected misreported information). It also noticeable that the majority of the teachers has a major in the subject that they are teaching, 87 %, and most of the teachers, 64 %, has more than 10 years' experience teaching. Only a third of the teachers, 33 %, has participated in professional development. For the students we see that less than half of the students' guardians, 45 %, has a degree defined as being "high level". First generation immigrants sum up to about 13 % in the high income countries and second generation immigrants sum up to 17 % of the sample.

4. Empirical strategy

To investigate the effect of instruction time on students test scores, one must be careful so that unobserved factors does not affect the results. As described in the previous section, each student is observed twice in the dataset. Every student has a unique ID and includes one row with variables for each subject, including variables for test results, instruction time and characteristics. I use the fact that each student is observed two times in a within-student between-subject fixed effects model which is similar to the one that Lavy (2015) uses. Since both observed factors such as school environment and unobserved characteristics, for example the students´ general academic ability, is the same (i.e. fixed) for both subjects but instruction time and teacher characteristics differ between subjects, it is possible to examine if differences between students performance between math and science are systematically correlated with differences in instruction time between the subjects. (Lavy 2015)

Using only a naive binary regression of test scores on instruction time would not control for the unobserved factors the way a fixed effects model does, which would lead to biased results. These biased results caused by omitted variables could both be biased upwards and downwards, i.e. result in coefficients that would either be too large or to small. For example, not taking students ability into account like the fixed effect model does would give misleading results.

Although I use a fixed effects model there might still be some unobserved factors that is not controlled for that could occur if the assignment of instruction time is not random, which could be the case if there are self-selection into classes by either students or students' parents. For example, a student who is great in math and likes the subject could choose a class with more instruction time in math which could lead to upward biased results. On the other hand, one could also imagine the opposite scenario where a student with weak performance in math gets enrolled in a class with more instruction time in math by its parents wishing to improve the students' performance which would lead to downward biased results (Bingley et al 2018). However, for this to take place in practice, parents must be aware of the relative amount of planned instruction time in different subjects when they choose a school for their children and schools must admit pupils based on subject-specific considerations and in most countries, eight grade students are still in an early stage of their education where "specialization" has not taken place yet (Bingley et al, 2018). While this may happen in a few cases, I assume that it is not likely that it takes place at a larger scale.

Using the same model as Lavy (2015) for the baseline regressions also makes it easier to compare the results with both his results and with other similar studies, such as Rivkin and Schiman (2015) and Bingley et al (2018), that uses similar models. The fixed effects model that is used has test score as the dependent variable and weekly instruction time as the (main) independent variable, and is presented as the equation below:

$$testscore_{ijk} = \beta_1 instruction_{jk} + \mu_i + \eta_k + u_{ijk}$$

where $testscore_{ijk}$ is the test score value for the student *i*, *j* is the school and *k* is the subject (either math or science). " $instruction_{jk}$ " is the weekly instruction time measured in hours per week in school *j* in subject *k*. The expressions that follow represent the student fixed effects (μ_i) which controls for unobserved aspects that effect performance at the student-level, but which does not vary between math and science, and unobserved subject specific characteristics (η_k). u_{ijk} is an error term. I also extend the model adding interactions with teacher characteristics, which gives the following equation, denoted as the following equation:

$$testscore_{ijk} = \beta_1 instruction_{jk} + \beta_2 instruction_{jk} * C_{lj} + \mu_i + \eta_k + u_{ijk}$$

where C_{lj} represent teacher characteristics for teacher *l* in school *k* which is then interacted with instruction time.

Although the fixed effect model controls for unobserved factors, there are still a few things to hold in mind. One is possible measurement errors in the data that could occur since the amount of instruction time and all information regarding the various characteristics are self-reported. This means that there could be some observations with values that are incorrect for some reason. For the instruction time variable, I dropped some variables that I suspected to be misreported but there could be some misreported values for the other variables as well. However, since there are a lot of observations and since most other variables has little room for misinterpretation and since the incentives to consciously lie are minimal, I don’t expect it to be a problem for the results. Another thing to keep in mind is that the model by construction assumes that the effect of increased weekly instruction time is the same for both subjects, although it might be possible that the effect differs between subjects.

5. Results

5.1 Baseline results

I perform a regression described in the previous section as my main regression, where test score either in math or science is the dependent variable and weekly instruction time is the independent variable. In addition to that, I perform a ”naive” OLS regression with the same variables to get an estimate of what the result would be without the fixed effects model. As previously mentioned, my main interest is on high income countries, so the sample is restricted to those countries[4]. As mentioned in section 2.1, senate weights are used in all regressions that are presented. The standard errors are clustered on school level, which is what Lavy (2015) does as well.

[4] As a robustness check, I perform the same baseline regression on different samples of countries. These results are presented “Robustness check”.

My baseline results are presented in Table 2. The result from the ”naive” OLS-regression is presented in column 1 and shows a negative effect of instruction time on test results. This is intuitively strange and not what previously studies have shown, which is why the fixed effects model is used instead. The OLS regression also has a very low R^2. When applying the fixed effect model (denoted FE in the tables) but not using any control variables, the results shows that an additional hour of instruction time per week increases test scores with 0.0137 SD in high income countries (column 2). Performing the same regression but including all confounding variables that are presented in Table 1 as controls, the effect is still positive and statistically significant although the coefficient is slightly lower at 0.0117 SD as seen in column 3.

Table 2: Baseline results

	(1) Test score	(2) Test score	(3) Test score
Instruction time	-0.0322*** (0.00162)	0.0137** (0.00163)	0.0117* (0.00168)
N	228 858	228 858	228 858
R^2	0.002	0.916	0.916
FE	No	Yes	Yes
Controls	No	No	Yes

Notes: Standard errors in parentheses: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Controls: all confounding variables that are presented in Table 1. Senate weights are used. The dependent variable, test score, is standardized (mean: 0, standard deviation: 1).

The effects from the baseline regressions are significant both with and without controls but lower than what Lavy (2015) presented using his dataset containing observations from 22 OECD countries in the PISA 2006 survey. As mentioned earlier, Lavys (2015) estimated baseline effect is 0.0580 SD. My result for the baseline case shows a lower coefficient which is more in line with studies following Lavy (2015). For example, using samples from the same 22 OECD countries and data from the latter PISA evaluations, the average effect is only 0.022 SD (Bietenbeck and Collins, 2020). Rivkin and Schiman (2015) also looked at a later PISA evaluation, PISA 2009, and got a baseline result of 0.0230 SD while Bingley et al (2018) gets a baseline estimate of 0.0360 SD using school- and register data from danish students. It is also worth noting that Lavy (2015) does not use any weights or any control variables for his baseline estimates.

5.2 Interactions with teacher characteristics

As an extension to the baseline regressions, I allow for heterogenous effects within the high income countries. The effect of instruction time on test results might vary not just between high income and low income countries, but also within the sample of high income countries depending on the teachers characteristics, such as the teachers qualification. It could be that a better teacher has a higher quality on the instruction time and that students effect of instruction time therefore differs because of the quality of teacher they have. To investigate this, I extended the model to include teachers' characteristics interacted with instruction time and performed new regressions. The results are presented in Table 3.

Table 3: Teacher interactions

	(1) Test score	(2) Test score	(3) Test score	(4) Test score
Instruction time	0.0117* (0.00168)	0.0119* (0.00169)	0.00854 (0.00178)	0.00919 (0.00197)
* Prof. development		-0.00240 (0.00127)		
* "Major (degree)"			0.0000110*** (0.00000117)	
* Experienced				0.00326 (0.00127)
N	228 858	228 858	228 858	228 858
R^2	0.916	0.916	0.916	0.916
FE	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes

Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. High income countries. Senate weights are used. The dependent variable, test score, is standardized (mean: 0, sd: 1). Controls variables are used in all regressions (as in table 2).

The first column (1) is the baseline regression without any interaction variables, using the fixed effect-model and controls. As also seen in Table 2, the baseline effect is 0.0117 SD. The second column (2) includes the variable "prof. development" which is an interaction variable with professional development interacted with instruction time. It is a dummy variable that takes the value 1 if the teachers during the last two years have taken part in some sort of activity involving professional development regarding the content of the subject, pedagogy and curriculum. If not, it takes the values 0. If a teacher continuously develops and require new skills within their subject, it might affect the impact the quality of the instruction time and therefore the impact instruction time have on the students. When interacting with instruction time I get a negative but statistically insignificant effect at -0.00240 SD. The negative coefficient indicates that instruction time with teachers that have participated in professional development activities has a negative impact on students test scores, but since the result is statistically insignificant no such definite conclusions can be drawn. Intuitively, it seems strange that the effect would be negative. Even if it would not be negative, this result does not suggest that the effect is positive when professional development is defined as it is in this case. The coefficient for test score stays positive and statistically significant and the coefficient becomes slightly larger, 0.0119 SD instead of 0.0117 SD, than in the baseline case.

In column (3) the effect of a teachers educational background is examined. The interaction variable “major (degree)” takes the value 1 if the teacher majored in the subject that they are teaching and 0 otherwise. The level of the degree could be either a bachelor’s degree, a master’s degree or a degree in educational science with an emphasis on the specific field. The idea is that a teacher that is properly qualified for the job education-wise might have a larger impact on the students than a teacher that is less qualified. For science teachers, I define the teacher as havening majored in science if the teacher has majored in either biology, physics, chemistry or earth science as a specific subject or if they have a degree in educational science with an emphasis on science. For math teachers, I define the teacher as having majored in the subject if the teacher has majored in math or if the teacher has a degree in educational science with an emphasis on math. The reason for this definition, that might seem broad, is twofold.
Firstly, as discussed in the section 2. Data , I defined science as one subject and not as several and therefore I define a degree as being in “science” as one here as well. Secondly, countries have different education systems throughout the whole education systems, including university level, and the ways of becoming a teacher can therefore differ. Since I wanted to include all teachers with a relevant degree, the definition is quite broad.

The result from the regression shows a positive and statistically significant effect of the interaction variable, although the coefficient is quite low at 0.0000110 SD. The positive effect indicates that students with a teacher that has majored in the subject that they are teaching preforms better at the tests, although not by much. The baseline regression is still positive although not statistically significant, and the coefficient is slightly lower at 0.00854 SD than in the baseline case.

The fourth column (4) shows the results for the variable indicating experience, “experienced”, taking the value 1 if the teacher has 10 years of experience or more. The idea is that a more experienced teacher might have a higher quality on its instruction time, resulting in higher test scores for its students. The coefficient is positive at 0.00326 SD but not statistically significant. A positive coefficient suggests that instruction time with an experienced teacher has a larger effect on students test scores, but since it is not statistically significant, we cannot say that that is the case based on these results. As in the case with “major (degree)”, the result for instruction time is also positive but not statistically significant, and the coefficient is slightly lower than in the baseline case at 0.00919 SD.

Interacting teacher characteristics with instruction time gave both expected and unexpected results. The baseline coefficients stay positive and statistically significant for all regressions as expected. Experienced teachers and teachers with a degree in the subject that they teach shows a positive effect, which sounds reasonable. Experience seems to be the most important factor according to these results. What is a bit surprising is the results for professional development where the coefficient was negative. But since the result was not statistically significant, no strong conclusions can be drawn.

5.3 Students background and socioeconomic effects

The effect on instruction time on test scores might also vary between students, for example students with different background characteristics may benefit differently of instruction time. This is also done by Lavy (2015). To explore this, I perform the baseline regression as before, still in high income countries, but now only for specific subsets of students. The results are presented in Table 4.

The first column (1) presents the results for girls and the second column (2) present the results for boys. Both has positive effect of instruction time, but while it is statistically significant foy boys it is not so for girls. The effect is larger for boys than it is for girls, with a coefficient of 0.0203 SD compared to 0.00658 SD for girls. This implies that male students seem to benefit more from instruction time, but since the result for girls is not statistically significant it hard to draw any definitive conclusions from these outcomes. Previously studies have presented mixed results. Lavy (2015) finds that girls have a slightly higher coefficient of 0.056 SD than boys at 0.050 SD, and Bingley et al (2018) present the opposite result where boys seem to benefit more from increased instruction time. It is therefore difficult to draw any conclusions to why that is based on this data alone. What can be said is that girls perform better on the test, both in high income countries and in the overall sample. This is displayed in Table 5. In high income countries, girls have a mean standardized test score of 0.278 SD while boys have 0.227 SD. The median score is 0.312 SD for girls and 0.276 SD for boys. It is also notable that both the person performing the worst and the best test score is a boy, both for high income countries and for all countries, while girls seem to perform at a higher level in general.

Table 4: Socioeconomic factors

	(1)	(2)	(3)	(4)	(5)	(6)
	Girls	Boys	Low ed	High ed	1st gen imm.	2nd gen imm.
Instruction time	0.00658	0.0203^{***}	0.0120^{*}	0.0114^{*}	0.00222	0.0153
	(0.00233)	(0.00242)	(0.00227)	(0.00251)	(0.00675)	(0.00554)
N	113 016	113 734	125 226	103 632	29 586	38 206
R^2	0.903	0.925	0.913	0.908	0.902	0.902
FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Standard errors in parentheses: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Senate weights are used. The dependent variable, test score, is standardized (mean: 0, sd: 1).

Table 5: Test scores

	Mean	Median	Std. Dev.	Min	Max
High income: girls	.278	.312	.892	-3.158	3.332
High income: boys	.227	.276	1.004	-3.543	3.536
All countries: girls	.024	.043	.954	-3.778	3.332
All countries: boys	-.024	-.017	1.043	-3.974	3.536

The educational level of students’ guardians are also taken into account, and in column three (3) the result for students with guardians that has ”low degree” are presented while the results for students with guardians having high degrees hare presented in column four (4). I define low degree when none or only one of the guardians has a degree on at least ISCED level 3, which is the equivalent of a high school degree. High degree is defined as if both guardians have a degree on at least ISCED level 3 or higher. Both coefficients for instruction time are positive and statistically significant, with a coefficient of 0.0120 SD for “low degree” and 0.0114 SD for “high degree”. The instruction time seems to have a higher effect on students with parents that are in the ”low education”-category than for those who has parents in the ”high education”-category, although the difference is not that large as large as one could expect. This result is similar to what Lavy (2015) finds, since he also finds that students whose parents has less education benefits from more instruction time in a higher degree than those students whose has parents with higher degrees.

In the fifth (5) and sixth (6) column the results for students who are 1st or 2nd generation immigrants. In column five we have the result for 1st generation immigrants, defined as students that are not born in the country which the test was conducted in. In column six, results for students with both guardians not born in the country which the test was conducted in. The result for students not born in the country which the test was conducted in shows a small and statistically insignificant effect of instruction time, and the result for students whose parents are born in a different country than the one that the test was conducted in also has a positive and insignificant. It is worth noting that the sample sizes for these regressions are quite small in comparison with the other regressions. Also, the students within these sample groups can be quite different since the reasons for immigration can differ greatly between countries within the high income countries. For example, there might be a big difference between a student who is a refugee from a country far away and a student who moved a shorter distance between countries where the culture and languages are similar. Lavy (2015) also find that the effect for 2nd generation immigrants (0.076 SD) is larger than for 1st generation immigrants (0.064 SD).

Students with different characteristics or backgrounds seem to benefit in different degrees of increased instruction time. My results shows that girls and boys seem to have different effects of instruction with boys benefiting more than girls, which Bingley et al (2018) also get but Lavy (2015) gets the opposite result. The differences in size of the coefficients also differs, I have bigger differences between boys and girls than Lavy (2015) does. When it comes to educational differences among the parents, we get similar results. This also goes for 1st and 2nd generational immigrants.

As the final extension, I perform the same baseline regression as before, including control variables, but also on different samples than high income countries. The results are presented in Table 6.

Table 6. Different samples of countries

	(1) High Baseline	(2) All countries	(3) Non high	(4) OECD
Instruction time	0.0117*	0.00346	-0.00933	0.00777
	(0.00168)	(0.00139)	(0.00242)	(0.00195)
N	228 858	362 048	133 190	116 458
R^2	0.916	0.923	0.923	0.917
FE	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes

Standard errors in parentheses: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Senate weights. Controls: all.

Column 1 shows the baseline for high income countries with the positive and significant effect of 0.0117 SD as seen before. For OECD countries, presented in column four (4), there is a positive although statistically insignificant effect, and the coefficient is smaller than for high income countries with the effect being 0.00777 SD. Lavy (2015) also uses OECD countries, but he has 22 countries in his sample and this dataset only has 15, so even if they both are defined as “OECD” they don’t include the same countries. When performing the baseline regression for the whole sample, which includes observations from all countries, there is also a positive but statistically insignificant effect, and the coefficient is smaller at 0.00346 SD than it is for high income countries. This indicates that the effect is larger for high income countries than for the whole sample. This result is presented in column two (2). When only the non high income countries are considered, the effect is negative but statistically insignificant, which is seen in column three (3). Lavy (2015) also gets a lower coefficient on countries that he defines as developing countries, although he still gets a positive effect at 0.030 SD compared to his baseline 0.058 SD.

It is a strength that the results are in line with what previous studies has found for the baseline case, although I use different data and samples. However, since the effect is not statistically significant for the non-high income countries, OECD countries or the whole sample, it is hard to draw any definite conclusion about the effect in those samples. What these results suggests is that the effect of increased instruction time on students test scores differs between countries.

5.4 Robustness checks

To see if the (baseline) results holds in different settings, I perform various robustness checks. As mentioned in section 2. Data, I´m using the first plausible value as the test score value in my regression. As a robustness check, I run the baseline regression again with both the first plausible value as seen before and with the test score variable calculated as a mean of all five available plausible values per student. This is done both with and without control variables. The results are presented in table 7. The first and second column shows the results without any controls, and the third and fourth columns displays results with controls. As expected, the results are very similar for both cases, using the first plausible test score value or using a mean of all five available test scores. With that said, using the mean value gives a slightly higher coefficient than using the first plausible value. However, the difference is small and changing the value for test score from the first plausible to the mean would not have changed the conclusions drawn from the regressions done in the previous sections (Jerrim et al, 2017).

Table 7. Different test score values

	(1) 1st pl value	(2) Mean value	(3) 1st pl value	(4) Mean value
Instruction time	0.0137** (0.00163)	0.0140** (0.00147)	0.0117* (0.00168)	0.0120* (0.00153)
N	228 858	228 858	228 858	228 858
R^2	0.916	0.931	0.916	0.931
FE	Yes	Yes	Yes	Yes
Controls	No	No	Yes	Yes

Standard errors in parentheses: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Senate weights are used.

As a further robustness check, I perform the baseline regression but with different or no weights [5]. Not using any weights means that the size of the sample that the country contributes with is what matters, no matter how much it what intended that the country should contribute with in the first selection process of choosing representative schools and classes. When not applying any weights at all, the effect is positive although statistically insignificant and the coefficient is smaller than with the senate weight (which is the baseline case), as is seen when comparing the coefficients in column (1) and (2). It is worth noting that Lavy (2015) does not use any weights. In the third column (3), total weights are used. Total weight scales the size of the sample to size

[5] There is a list in the Appendix which includes all the included countries and their respective weight when using senate-, total- and no weights.

of the population of each country, so bigger countries receive a larger weight than smaller countries. The results show a negative effect of increased instruction time on test scores, although it is statistically insignificant. The interpretation of the negative sign would be that increased instruction time has a negative effect on test scores, but since the effect is statistically insignificant no strong conclusions can be drawn. However, this is an additional indication of what was mentioned when changing the sample of countries – the effect of instruction time on test scores seems to differ between different countries.

Table 8. Different weights

	(1) Senate weights	(2) No weights	(3) Total weights
Instruction time	0.0117* (0.00168)	0.00686 (0.00136)	-0.00145 (0.00287)
N	228 858	228 858	228 858
R^2	0.916	0.920	0.928
FE	Yes	Yes	Yes
Controls	Yes	Yes	Yes

Standard errors in parentheses: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

6. Conclusion

By using a fixed effects model that exploits within-student between-subject variation on data from the TIMSS 2019 evaluation, I get a positive effect of increased weekly instruction time on students' achievement in high income countries. This main result is in line with previous literature such as Lavy (2015), although the coefficient is lower. The lower coefficient is also what other studies following Lavy (2015) has presented. The results from interacting with teacher characteristics were not as convincing as expected, since they were either statistically insignificant or had very low coefficients. When examining different subgroups of students, I found that students benefit differently depending on background- and socioeconomic characteristics which is both interesting and in line with what Lavy (2015) and others also finds, although the differences between boys and girls seems to differ between different papers.

As mentioned in the introduction, improved student performance has previously shown to affect growth (Goldin & Katz, 2009). Therefore, it is an important result that increasing instruction time can improve learning. However, since the positive and statistically significant effect disappears when the sample changes and when different weights are applied, it is quite clear that the effect of instruction time on test scores are different in different types of countries, and even among high income countries depending on the weight that is applied. If the effect was universal, it would not matter so much which countries or which weights that were applied because the effect would be about the same either way. And since no paper using data later than PISA 2006 gets a coefficient as high as Lavy (2015), one could ask if the magnitude of the effect of increased weekly instruction time on students' performance really is as big as Lavy (2015) suggests.

Given that there is a positive effect of increased instruction time on test scores in high income countries, a policy implication could be to give more instruction time in math and science even if the effect does not seem to be as large as suggested by Lavy (2015) and even if the effect seems to differ between countries. However, it should probably not be the only solution although it might work for some subgroups of students since I found that different subgroups of students benefit differently from increased instruction time. One should also be aware of that time is a limited resource and a student can only have so many productive hours per week, meaning that while increasing instruction time in math and science may increase performance in those subjects slightly, you might be forced to reduce instruction time in other subjects such as social science or language which could lead to weakened performance in those subjects.

It also worth noting that the focus is on eights grade students here and for students about that age in studies that uses PISA. It might be the case that the effect of increased instruction time is larger for younger students. It is also possible that the effect can be different in other subject, such as language. All in all, increasing instruction time should be considered by policy makers to increase students' performance, but it also important to consider other elements as well to get as good results as possible overall.

7. References

Bietenbeck, Jan & Collins, Matthew, 2020. "New Evidence on the Importance of Instruction Time for Student Achievement on International Assessments" Working Papers 2020:18, Lund University, Department of Economics. Available: https://ideas.repec.org/p/hhs/lunewp/2020_018.html

Bingley, P., E. Heinesen, K.F. Krassel, and N. Kristensen. 2018. The Timing of Instruction Time: Accumulated Hours, Timing and Pupil Achievement. *IZA Discussion Paper No. 11807*. Available: https://www.iza.org/publications/dp/11807/the-timing-of-instruction-time-accumulated-hours-timing-and-pupil-achievement

Goldin, C.; L. F. Katz (2009). *The Race Between Education and Technology*. Harvard University Press. Available: http://www.nber.org/papers/w12984

Jerrim, J., Lopez-Agudo, L. A., Marcenaro-Gutierrez, O. D., & Shure, N. 2017. What happens when econometrics and psychometrics collide? An example using the PISA data. *Economics of Education Review.* 61: 51-58. Available: https://doi.org/10.1016/j.econedurev.2017.09.007

Lavy, Victor. 2015. Do Differences in Schools' Instruction Time Explain International Achievement Gaps? Evidence from Developed and Developing Countries. *The Economic Journal.* 125(588): 397-424. Available: https://doi.org/10.1111/ecoj.12233

OECD. 2015. *Education at a Glance 2015: OECD Indicators*. OECD Publishing. Available: http://dx.doi.org/10.1787/eag-2015-en

Mullis, I. V. S.; Martin, M. O. 2017. *TIMSS 2019 Assessment Frameworks*. Available: https://timss2019.org/wp-content/uploads/frameworks/T19-Assessment-Frameworks.pdf

Rivkin, Steven; Schiman, Jeffery. 2015. Instruction Time, Classroom Quality, and Academic Achievement. *The Economic Journal.* 125(588): 425-448. Available: https://doi.org/10.1111/ecoj.12315

TIMSS. 2019. About TIMSS & PIRLS International Study Center. Available: https://timssandpirls.bc.edu/about.html

TIMSS. 2021. TIMSS 2019 International Database. Available: https://timss2019.org/international-database/

World Bank. 2021. World Bank Country and Lending Groups. Available: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519#High_income

Appendix

Table A1:

ISO code	Country	No wgt	Tot wgt	Sen wgt	OECD	High income
36	Australia	3.81	1.55	2.32	Yes	Yes
48	Bahrain	2.88	0.12	2.91	No	Yes
152	Chile	1.68	1.31	2.32	Yes	Yes
158	Taiwan	2.66	1.50	3.05	No	Yes
196	Cyprus	1.29	0.04	2.08	No	Yes
246	Finland	2.51	0.39	2.93	Yes	Yes
250	France	1.53	4.48	2.30	Yes	Yes
268	Georgia	1.65	0.29	2.72	No	No
344	Hong Kong	1.40	0.32	2.46	No	Yes
348	Hungary	2.28	0.60	2.82	No	Yes
364	Iran	3.30	8.05	3.11	No	No
372	Ireland	1.74	0.38	2.42	Yes	Yes
376	Israel	1.84	0.69	2.76	Yes	Yes
380	Italy	1.81	3.60	2.79	Yes	Yes
392	Japan	2.44	8.10	3.09	Yes	Yes
398	Kazakhstan	2.24	1.56	2.80	No	No
400	Jordan	3.53	0.98	2.80	No	No
410	Korea	1.91	2.98	2.80	Yes	Yes
414	Kuwait	2.17	0.30	2.68	No	Yes
422	Lebanon	1.86	0.36	2.25	No	No
440	Lithuania	2.02	0.18	3.00	Yes	Yes
458	Malaysia	3.55	2.79	2.84	No	No
504	Marocco	4.24	3.30	2.85	No	No
512	Oman	2.83	0.30	2.36	No	Yes
554	New Zealand	2.66	0.34	2.43	Yes	Yes
578	Norway	1.46	0.26	1.79	Yes	Yes
620	Portugal	1.77	0.73	2.95	Yes	Yes
634	Qatar	1.51	0.10	2.23	No	Yes
642	Romania	2.16	1.18	2.68	No	Yes
643	Russia	1.99	9.33	2.86	No	No
682	Saudi Arabia	2.14	2.00	2.14	No	Yes
702	Singapore	2.56	0.28	2.96	No	Yes
710	South Africa	8.32	4.91	2.32	No	No
752	Sweden	1.95	0.72	2.81	Yes	Yes
784	United Ar.Em.	7.16	0.30	1.83	No	Yes
792	Turkey	1.93	7.28	2.66	No	No
818	Egypt	3.36	9.25	2.64	No	No
840	United States	3.04	17.35	1.97	Yes	Yes
926	Kosovo	0.82	1.79	1.28	No	No
Total		100	100	100		
High income		*63.2*	*50.1*	*68.9*		