Budapest,
06.06.2016

The effect of teacher's gender on student's educational performance- An international evaluation

by Alfa Diallo

Submitted to Central European University
Department of Economics

In partial fulfillment of the requirements for
degree of master of Economics

Supervisors: Zoltán Hermann
Gábor Békés

Abstract

In this paper I investigate two questions. First, whether the quality of female or male teachers is better with respect to 8th grade students, and second, whether students with same sex teacher archive higher educational outcomes. For my estimation I use four waves of the TIMSS international dataset, which is a database that contains the test result of 8th grade children in mathematics and science. I investigate the two main questions of my paper, using twenty different European countries subjects. I use OLS and individual fixed effect regressions in order to answer the two main questions of the paper. The latter method is based on the idea to identify effects, from the individual test point variation across different subjects. The dataset is very well suited for this type of estimation, as the investigated subjects are very similar in nature and high in number. According to my results there is no economically relevant difference between men and women teachers. On the other there is a robust evidence, that having a same sex teacher can be beneficial, however the magnitude of this potential advantage is very heterogeneous across countries.

Table of contents

List of tables

Introduction

In the recent educational economic literature, a central question is that how important are teachers in determining the elementary and secondary students' educational outcomes. In this paper I will investigate a subtopic of this problem by analyzing whether the gender of the teacher has any effect for the pupil's performance.

The empirical findings in this topic are mostly centered around the question whether having a same sex teacher is beneficial for the educational outcome. The importance of this problem lies in the fact, that papers investigating this topic came to quite contradicting conclusions, as many researchers found that the same sex pairing of students and teachers are advantageous (Dee, 2004; Ammermüller-Dolton, 2006; Dee, 2007), but other scholar were not able to identify such relationship (Holmlund- Sund, 2008; Carrington et al., 2008). In my opinion however the pairing effect is already a second step if we want to analyze the effect of teacher's gender on students' performance as it's heavily interlinked with the question whether there is a quality difference between male and female teachers. It is quite likely that such a difference exists, as in most of the developed countries teacher profession is very contra selective, and the portion of women teachers are very high (Eide et al., 2004).

So during my research I will investigate two specific questions. First whether there is a quality difference between men a women teachers, and second whether the same sex pairing of students and teachers increases educational performance. On top of that as most of the cited researches used Anglo-Saxon dataset I am also interested in the fact whether the measured connections are stable or heterogeneous across different countries.

I will investigate the presented questions using the TIMMS database, which is an international dataset that measures the test performance of 8th grade students in mathematics, physics, chemistry, biology and earthly sciences. As a basic estimation, in twenty selected European

countries, I will regress the standardized test score of an individual in a given subject on a teacher gender dummy, a student sex dummy, a double interaction term which shows whether the sex of the student is the same as the teacher's and on background control characteristics.

However as probably the pairing of teachers and students are not random it is possible that the OLS estimation will not be able to determine the casual effect of the gender related variables on outcome. For this reason, I will estimate individual fixed effect regressions as well. In my opinion my dataset is better suited for these kind of estimations than the ones used by other scholars investigating this problem, as the subjects on which I will base my estimation strategy are very similar in nature, which is a basic requirement of this method, and high in number, which increases precision. In both type of estimations, I will use a model where the countries are pooled, in order to identify the basic mechanism and separate regressions for the different countries to test for heterogeneity.

My paper is structured as follows. In the second section I will show the basic mechanisms and the main empirical results, related to the topic. The third section consists the data description while the fourth introduces the basic identification strategies. In the fifth section I will present and interpret the results of the OLS and individual fixed effect estimations on the pooled sample and separately for the twenty participating countries. In the last section I will summarize my results and highlight some possibilities for further research.

Literature review

Theoretical explanations

In the economic literature many paper investigates the question, that how the gender of the teacher affects the students' performance in elementary and secondary school. The topic however is quite complex as in order to draw conclusion one most investigate the gender differences in both of the student and teacher side, as well as the pairing of pupils and teachers with respect to their sexes. In this section I will collect the main theoretical considerations and the corresponding empirical results in these topics.

As one of the main focus of my own evaluation will be, whether same sex pairing of students and pupils affect the educational performance of the students, first I will show the results that answer this question. There is a debate in the literature related to the problem, as the published results are somewhat contradictory. The reason for this, that many authors find that there is a positive effect of same sex teachers on school performance (Dee, 2004; Ammermüller-Dolton, 2006; Dee, 2007), while there is another set of researches which cannot identify such positive pairing effect (Holmlund- Sund, 2008; Carrington et al., 2008). Antecol (2015) on top of that shows that in some cases it is possible, that same sex teachers have negative effect on achievement.

Before I briefly present the papers which support the results, I would like to highlight the basic mechanisms that can stand behind the positive effect of same sex pairing. Dee (2007) identifies two basic mechanisms that can explain such phenomenon. The first affects the behavior of the students while, the second the attitude of teachers. Dee's (2007) first theory is that the effort level of students is dependent from the fact, whether the pupil considers, his/her teacher as a role model. In this case the child is more willing to put more effort into school work, which results in better performance. The main idea of Dee (2007) that it is more likely that a student

chooses someone who is similar to him as a role model (same sex), which idea is supported by the study of Bettinger and Long (2005), which gives evidence for this phenomena by investigating university course selection.

The other idea is that teachers tend to discriminate students who are different (different sex) from them (Dee, 2007). This negative discrimination can hinder the performance of the pupil through two channels. First, directly as the student possibly cannot receive as much attention, he can accumulate less human capital, and second, indirectly, if this possible "negative feedback" makes the student to invest less energy into learning. Greshenson et al. (2015) and Dee (2005) shows evidence that discrimination based on non-similarities exist. Although some researchers came to another conclusion as Terrier's (2015) result suggest, that teachers independently of their gender positively discriminate girls in grading in mathematics, while Lavy (2004) presents results that in matriculation exam teachers discriminate against boys.

From the above presented results, it is clearly visible that it is not yet decided, whether there is a positive benefit of same sex matching, neither the source of this potential bias. On top of that the difference in performance between sexes can both in the case of students and teachers complicate the situation. As I stated in the introduction, my main goal is to measure the effect of teacher's gender on student's performance. The answer to this question can be the identification of the above mentioned pairing effect, but it can be discussed in a more general level, whether there is a difference in the quality between women and men teachers. It can be easily seen that the two problems can be interlinked, as quality differences may result in different total pairing effects, but the question can be of interest independently of the pairing as well.

It is possible that there are some biological differences between males and females which predestinates one sex to be better in educating children, but the composition of elementary and secondary school teachers and the characteristics of the job may also result in different quality.

As Eide et al. (2004) states the portion of female teachers are much higher in elementary school, and secondary school than male ones. Eide et al. (2004) also show that teachers job is very contra selective, so not the most talented persons choose to study to become teachers, and many of them leaves their job in order to live better.

The question is whether the relative rate of contra selection is different between genders which can be the main source of potential differences between the effectiveness of men and women teachers. Two theoretical mechanisms are imaginable, which can result in different rate of contra selection. First, that in the case of men being a teacher is not a very prestigious job, so the small portion of teachers mostly consist of men whose life goal is to teach, which will result in the fact that male teachers are better. The second idea is that because of the unpopularity teaching only the worst skilled men choose this job as profession, which results in better female teachers on average. Eide et al. (2004) presents some proof that men are relatively more contra selected, than women.

Finally, the pairing effect can be also influenced by the relative difference in performance of students by gender, which are independent of the teacher's performance. As this question is not the main focus of my paper I only briefly present some potential ideas. An educational summary by OECD (2015) presented for example that, girls perform significantly better in reading, while boys outperform them slightly in math. As Antecol et al. (2015) argue, in mathematics girls based on historical reasons tend to face math anxiety which can result in the fact that boys outperform them. In my own estimation I will work with test results in mathematics and science, so this mechanism can give an explanation of the potential score advantage of boys.

Measurement difficulties and methods

In the empirical literature, the scholars who tried to measure the effect of teacher's gender on student performance faced two important methodological problems. The first one is that the pairing of teachers and students is not random (Kane et al.,2011). The sorting based on quality can happen between schools and within schools as well.

With respect to my research question it is not necessarily a problem that can bias the estimation results. There are two possible cases where the non-random sorting can be a methodological problem. The first one is if the quality of teachers is not independent of gender. The second possible case is if the quality of teachers is the same across sexes, but the sorting for some reason is not independent of it. To give an example it is possible, that due to historical reasons the expectation of school directors is that female teachers are better, so better schools which can choose from many applicants are more willing to hire women than men based on this belief.

There are many ways earlier studies tried to deal with this potential problem. One is to measure the effect of teachers in a randomized control trial environment. One example to these kind of research is Dee's (2004). In his evaluation he used the data of "Project Star" which program was originally designed to measure the effect of class size on school performance. The project had one important characteristics that within a school, the teachers were randomly paired with students. It allowed for Dee (2004) to estimate the effect of the teacher's gender on student's performance by a school fixed effect estimation. His main result where that same sex pairing is beneficial for both males and females in mathematics as well as in reading.

Unfortunately, in most of the cases it is not possible to make or evaluate a randomized control trial so researchers have to use alternative estimation methods in order to identify casual effect. One very common solution is to identify the effect of the teacher by comparing the results of the same individual across subjects. This methodology was used by Ammermüller and Dolton

(2006) who investigated the gender difference of 4th and 8th graders in the United States and Great- Britain based on a comparison of mathematics and science results and Dee (2007) who compared the test results of 8th grader American students in the subjects of mathematics, science, reading and social science. The latter found clear evidence of positive same sex pairing, while the former research identified the same effect but not in all specifications.

The other important methodological problem is that the students' performance is not only affected by their current teachers' quality but also affected by the work of their earlier ones (Ammermüller-Dolton, 2006). One solution can be, to measure the effect of the first teacher, so make the evaluation as early as possible in terms of school years, so the pupil did not have the opportunity to study from many teachers. The other possible solution is to use panel data, and individual variance in performance in time as a basis of identification. This method was used in the paper of Humlund and Sund (2008) who were not able to detect any beneficial effect related to same sex pairing of students and teachers based on Swedish data.

In this section I presented the main theoretical mechanisms that can affect the influence of teacher's gender on student's educational performance. I also presented the most important methodological difficulties, and the results of other scholars, related to the topic. In the upcoming sections of my paper I will focus on the description and presentation of my own evaluation.

Data

In this section I will describe the used dataset and present some basic descriptive statistics. Throughout my estimation I used the database of the Trends in International Mathematics and Science Study (TIMSS). This is an international assessment which started in 1995, and repeated since in every 4 years. The study measures the educational performance of 4^{th} and 8^{th} grade children on internationally standardized tests in mathematics and natural sciences. In the case of 8^{th} graders natural science is divided into four subsections, which are physics, chemistry, biology and earth science. Every wave of TIMMS consist of approximately twenty countries from all over the world, but mostly East- and Central- European, Scandinavian, Asian and American countries. One disadvantage of the dataset that the participant nations differ significantly in the different waves.

The use of TIMSS has many important advantages. First, that it is possible to link the participated students with their corresponding teacher even at the subsection (within natural science) level. Second, that consists of a huge variety of personal, teacher and school background variables, that are usable as controls. And third because of the international nature of the data it is possible to see whether the results that were presented in the literature review part, are heterogeneous across countries.

On top of these basic advantages TIMSS is useful for two other reasons. In some of the presented empirical literature, the outcome variable that were measured was not independent of the teacher itself. The most prominent example of this kind of dependence appears if the outcome variable are school grades (Terrier, 2015). This can be a problem because in these cases it is possible that the authors measured two different things. The first one, that how the work with the given teacher affected the pupil's human capital accumulation (the question I am

interested in), and second some other factors that bias just the teacher's evaluation. Because TIMSS is international student assessment the latter channel cannot bias the test result.

The other important aspect is that TIMSS consists of quite similar subject. As I mentioned in the theoretical section of my paper there are many evidence that the pairing of students and teachers are not random (Kane et al.,2011). For this reason, most of the identification strategies (where randomized control trials were not possible) are based on the variation of within student results, which is only a valid strategy if the potential outcome of students is the same between subjects. Because of data availability it can that not really similar subjects were considered, in which cases this assumption is more likely not to hold.

In my estimation I used the 1999, 2003, 2007, 2011 waves of TIMSS. In the choice of the relevant time period, two effects that were acting in the opposite direction affected my selection. One was based on my identification strategy and I required a lot of individual observations, because in the case of a short time period horizon the standard errors are quite high relative to the point estimates which results in imprecise zero results. On the other hand, in longer time horizon it is possible that there were changes in the educational system, which possibly affected the effect of teacher's gender on students' performance. As in 1995 wave the control variables were quite different from the following four waves I excluded those data from the estimation.

Generally, I considered only European countries in my regressions, because in my opinion it is important compare nations that are somewhat similar in their educational culture. In some European countries science is not divided into physics, chemistry, biology and earth science in 8^{th} grade education but taught as integrated science. In these countries it is not possible to benefit from the fact that the subject which were considered are possibly very similar, and because of the integration of natural sciences into one subject in the evolution of teachers' effect I loose precision. For this reason, in my estimation I only considered those countries where science is being taught as separate subjects.

Finally, in my evaluation I only worked with the results of 8th grade students. The first reason that in the case of 4th graders science cannot be divided into different subjects. The more serious problem however, is that as I mentioned in the theoretical part of my paper, teacher profession in most of the countries in elementary and secondary school level are heavily dominated by women teachers. This dominance is bigger in lower grades than upper grades which fact represents itself in TIMSS teacher data. In the case of the 4th graders in most of the countries more than 90% of the sample teachers are women, which does not provide enough variance in the gender of the teachers to estimate reasonable effects. This ratio in the case of the 8th graders is approximately 70%, for which reason I only used their results in my regressions.

Country	ID	Waves	Observations	Country	ID	Waves	Observations
Bulgaria	Bgr	3	44950	Macedonia	Mkd	3	56773
Bosnia	Bih	1	20603	Malta	Mlt	1	14996
Cyprus	Cyp	2	30792	Netherlands	Ndl	1	13011
Czech Rep.	Cze	2	39937	Romania	Rou	4	85014
Estonia	Est	1	17819	Russia	Rus	4	91283
Finland	Fin	2	27497	Serbia	Srb	2	40372
Hungary	Hun	4	76717	Slovakia	Svk	2	25843
Lithuania	Ltu	4	73398	Slovenia	Svn	4	46627
Latvia	Lva	2	24439	Sweden	Swe	3	29793[1]
Moldova	Mld	2	34979	Ukraine	Ukr	2	37965
				Full sample	Full	4	823777

1. Table: Basic statistics of the countries used in the estimation, own calculations

Table 1 summarizes the twenty countries that remained in the sample after the exclusions. As it seems the countries with which I worked are mostly post-socialist and Scandinavian states. The second column show the country ID of the given nation, that will be used in graphs and tables throughout this paper. As I highlighted earlier the composition of countries were quite different in the waves of TIMSS. For this reason, the third column consists the information that how many waves the given country was involved in, while the forth column presents the final

[1] In Sweden it is possible to study in 8th grade integrated science and the separated subjects as well, so I considered only those students who were learning in school where science was taught as different subjects.

observation number. It is important to note that the unit of observation throughout my paper is not one student, but one student linked with his/her teacher in a given subject.

In Table 2 I collected the information about the ratio of women teachers by country and subject. The second column consist of the country averages of women teachers all subject merged, while the third to seventh column shows the same ratio separately. In my estimation strategy it will be important that what portion of the students have teachers from both gender, which information is located in the last column of the table. It is clear from these statistics that in most of the countries the vast majority of teachers are women (the only exception is the Netherlands), but there are huge differences between countries. The ratio varies from approximately 50% to 80%. There is also a clear pattern that there are structural differences between subjects as well. While biology and chemistry are relatively more women dominated, in physics and earth science the portion of man teacher is much higher, while the same value in mathematics lies somewhere in-between.

In order to get a whole picture of the composition of our used sample I should show some statistics about the ratio of boy and girl students, by country. On the other hand, I won't report this table, because there is quite small variance in the portion of girls, with an average value of 0.5, and minimum and maximum of 0.49 and 0.52 respectively.

Country	Full Sample	Mathematics	Physics	Biology	Chemistry	Earth Sci	Variance
Bgr	0.82	0.86	0.81	0.87	0.88	0.72	0.35
Bih	0.63	0.58	0.57	0.7	0.73	0.55	0.86
Cyp	0.64	0.66	0.54	-[2]	0.72	0.65	0.75
Cze	0.72	0.76	0.55	0.81	0.84	0.62	0.71
Est	0.82	0.89	0.57	0.9	0.88	0.88	0.49
Fin	0.63	0.55	0.59	0.70	0.60	0.72	0.69
Hun	0.76	0.82	0.69	0.77	0.83	0.67	0.82
Ltu	0.86	0.94	0.7	0.91	0.94	0.83	0.56
Lva	0.84	0.92	0.60	0.94	0.90	-	0.89
Mda	0.75	0.80	0.57	0.83	0.85	0.71	0.41
Mkd	0.58	0.59	0.60	0.64	0.64	0.46	0.82
Mlt	0.61	0.6	0.54	0.78	0.64	0.5	0.82
Nld	0.26	0.33	0.17	0.37	0.17	0.27	0.72
Rou	0.73	0.6	0.72	0.83	0.85	0.65	0.62
Rus	0.93	0.96	0.82	0.96	0.96	0.93	0.27
Srb	0.7	0.62	0.62	0.81	0.78	0.67	0.63
Svk	0.79	0.81	0.72	0.84	0.85	0.71	0.52
Svn	0.81	0.84	0.63	0.9	0.91	0.78	0.42
Swe	0.52	0.51	0.41	0.58	0.58	-	0.40
Ukr	0.87	0.92	0.65	0.94	0.94	0.88	0.62
Full	0.75	0.75	0.65	0.83	0.82	0.70	0.52

2. Table: The portion of female teachers by country and subject, own calculations

Before I turn my attention toward the regression analyses that I have conducted, I think it is important to present some raw statistics about the distribution of test points by teacher's and student's gender. For this reason, in Table 3 I created a matrix that consist of the test points of all possible student-teacher gender combinations. In this table not just the waves, but the countries and the subjects were merged as well, in order to identify the basic statistical connections related to sexes. The vertical axis separates based on the teacher's while the horizontal on the student's gender. Each cell of the table shows the corresponding average test score (bold) and its standard deviation.

[2] If data is missing that means that the given subject is not taught separately in the corresponding country

	sex of teacher		
sex of student	male	female	average
boy	**501.87**	**515.16**	**513.07**
	88.435	85.276	85.915
girl	**498.68**	**511.29**	**509.35**
	82.041	79.88	80.347
average	**500.28**	**513.21**	**511.16**
	85.318	82.322	83.18

3. Table: The average value and standard deviation of test scores based on the sex of the teacher and the student

From this raw comparison three important conclusions can be drawn. First, that on average women teachers perform much better than men because both girls and boys archive better performance with a female teacher, by approximately 13 test points. Second that boys perform slightly better (4 tests points) than girls and this connection is stable across the sex of the teacher. The third conclusion comes from the combination of the earlier two. According to theses averages, there is no support for the theory that the pairing of student and teacher by gender has an additional effect on the test performance of the children.

Identification strategy

As I stated in the introduction of my paper I'm mainly interested in two questions. Whether the students of male or female teachers perform better, and whether those pupils who have teachers same gendered as themselves achieve better results or those whose teacher's sex is different.

In my first type of estimation I created a huge sample, where I pooled the different countries, the different waves and the different subjects into one dataset. By working with this pooled sample one can identify the average link between student and teacher gender. For this reason, I estimated the following regression:

$$(1)\ y_{i,s,m,c} = \beta_0 + \beta_1 * tfemale_{s,m,c} + \beta_2 * sgirl_{i,m,c} + \beta_3 * samesex_{i,s,m,c} + \beta_4 * X_{i,m,c} + \beta_5 * T_{s,m,c} + \beta_6 * S_{s,c} + \kappa_s + \forall_c + \varepsilon_{i,s,m,c}$$

In (1) i represent different individuals, s stands for the corresponding subject, m separates the different schools, while c appear for countries.

The dependent variable of the regression $y_{i,s,m,c}$ is the standardized test score of a given individual in a given subject, school and country. The standardized test score was created from the original test score of the students, with the (2) formula, where w is the year in which the test was taken, and p is the original test result of the individual in the given subject.

$$(2)\ y_{i,s,m,c} = \frac{p_{i,s,m,w,c} - \bar{p}_{s,w,c}}{s.d.(p_{s,w,c})}$$

The result of this transformation that in all wave, in all country in all subject the mean of the standardized test score is 0 while it's standard deviation of is 1. The standardization of the test score was necessary for several reason. The main purpose of TIMSS is to compare the school performance of different countries within and between waves. Because of this, the mean and the variance of test score in these two dimensions can differ greatly, which can cause problems by the interpretation of the results, because it is possible that instead of the intended teacher and

pairing effect I measure only changes in the averages. The standardization by subject would not be necessary at this point, but it becomes important in later presented estimation methods in my paper, so I will describe the importance of it by those regression, but for simplicity issues I used the same dependent variable in all of my estimations.

In regression (1) the main variables of interest are those, that represents the gender of either the student, teacher or the interaction between them. These variables are $tfemale_{g,s,c}$, which is a dummy variable which takes value one if the teacher of a given individual in a given subject is female and zero if male. $sgirl_{i,g,c}$ is also binary variable, which takes value one if the observed student is a girl and zero if a boy. $samesex_{i,s,m,c}$ can be considered as a double interaction term suggested by Ammermüller and Dolton (2006) because it shows whether the sex of the teacher of the given individual in a given subject is the same (value one) or different (value zero) from his or her own gender.

The estimation also contains background control variables. The reason for this, that in the theoretical part of my paper I presented, that the pairing of students and teachers are potentially not random, so I need to control for factors, that effect the performance of either the students or the teachers and can possibly correlate with the gender. The dataset allowed me to include three types of control variables. $X_{i,m,c}$ are individual level background characteristics, $T_{s,m,c}$ contains of class (subject) level variable that are mostly connected to teachers, while $S_{s,c}$ are variables that are different between schools. For the whole list of these controls see Appendix (A)1.

Finally κ_s is a categorical variable of the different subject and $\forall_c$ for the different countries, while $\varepsilon_{i,s,m,c}$ is the error term of the regression. In all of my estimations throughout this paper the standard errors were clustered at a class level.

Theoretically if I can control every characteristic that affect the students test score, and correlates with either the sex of the teacher, pupil or the double interaction term, than I will get a consistent estimate for the casual effect of $\beta_1, \beta_2, \beta_3$ from the OLS regression presented in equation (1). However, it is quite likely that due to the limitations of the available control, it is not possible to rule out completely, the biasing effects of non-random pairing.

To control for this possible distortion, I estimated another type of regression.

(3) $y_{i,s,m,c} = \gamma_0 + \gamma_1 * tfemale_{s,m,c} + \gamma_2 * samesex_{i,s,m,c} + \gamma_3 * T_{s,m,c} + \kappa_s + \sigma_i + \varepsilon_{i,s,m,c}$

The notations of (3) are the same as (1), and σ_i stands for the individual fixed effect. The idea of this estimation method is to identify the effect on the children based on the variance of a given individual between teachers. So the individual fixed effect estimation in my case compares the standardized test results of a given children in mathematics, physics, biology, chemistry and earthly science.

This estimation method however requires an additional assumption, that the potential outcome of the student is the same in the different subject, in other words the differences in the student's performance in different subjects controlled for the background characteristics is only a result of the teachers' work. The advantage of the TIMSS dataset that it consists of subjects of natural sciences so it is quite likely that this above mention assumption holds[3].

As the identification is based on the variance of test score of a given individual between subjects, this is a clarification of why the standardization between subject was necessary as well, because without it would be possible that I would measure differences in the test results due to the fact that the average score of the children differ by subjects.

[3] However, I cannot check it directly as there are no „counter factual" exist.

The individual fixed effect estimation absorbs any individual, and school level background characteristics, but not the class (subject) level variables, so in the estimation I included $T_{s,m,c}$ as well for the same reason as in equation (1).

By merging however, one can lose important information, because in the estimation, the regressions always assume homogeneous effects by the different dimensions. It is a relevant question however that how heterogeneous are the estimated coefficients between the twenty selected countries.

For this reason, I estimated the following regressions as well, which are really similar to (1) and (3).

$$(4)\ y_{i,s,m} = \beta_0 + \beta_1 * tfemale_{s,m} + \beta_2 * sgirl_{i,m} + \beta_3 * samesex_{i,s,m} + \beta_4 * X_{i,m} + \beta_5 * T_{s,m} + \beta_6 * S_s + \kappa_s + \varepsilon_{i,s,m}$$

$$(5)\ y_{i,s,m} = \gamma_0 + \gamma_1 * tfemale_{s,m} + \gamma_2 * T_{s,m} + \kappa_s + \sigma_i + \varepsilon_{i,s}$$

The only difference, that both in the case of the OLS estimation and the individual fixed regression I will estimate the coefficients separately for the 20 different countries of which my merged dataset consist. In the evaluation of these result I will comment on the measured differences between nations.

Results

OLS

In this section I will present the results of the three type of estimations that I introduced in the identification strategy section. In every estimation I made some small changes in the original equation that I will indicate specification by specification.

First I estimated the merged OLS regression (1) in three different forms:

$$(1a)\ y_{i,s,m,c} = \beta_0 + \beta_1 * tfemale_{s,m,c} + \beta_2 * sgirl_{i,m,c} + \beta_3 * samesex_{i,s,m,c} + \kappa_s + \forall_c + \varepsilon_{i,s,m,c}$$

$$(1b)\ y_{i,s,m,c} = \beta_0 + \beta_1 * tfemale_{s,m,c} + \beta_2 * sgirl_{i,m,c} + \beta_4 * X_{i,m,c} + \beta_5 * T_{s,m,c} + \beta_6 * S_{s,c} + \kappa_s + \forall_c + \varepsilon_{i,s,m,c}$$

Where in 1a the background control variables are not included, while in 1b the double interaction term is not present. Comparing (1) with (1a) shows that the inclusion of control variables at what magnitude changes the coefficients of the gender variable and the comparison of (1b) with (1) shows the cross dependence between the raw sex variables and the double interaction term.

Table 4 shows $\beta_1, \beta_2, \beta_3$ in the different specifications. Because of the large number of control variables (space limitations) I did not report the coefficients of the control variables, only indicated whether they were included in the regression or not.

4. Table: The results of the OLS estimations, pooled across countries, own calculations

VARIABLES	(1a) y	(1b) y	(1) y
tfemale	0.0731*** (0.0106)	0.0230*** (0.00870)	0.0232*** (0.00870)
sgirl	-0.0257*** (0.00685)	-0.0637*** (0.00533)	-0.0693*** (0.00597)
samesex	0.00942* (0.00521)		0.0116** (0.00455)
Controls	no	yes	yes
2X interaction	yes	no	yes
Observations	823,777	823,777	823,777
R-squared	0.001	0.174	0.174

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

In all specifications the coefficient of $tfemale$ shows that holding everything else constant, with how much standard deviation students with women teacher perform better, while the coefficient of $sgirl$ shows the same difference between girl and boy pupils. The coefficient of same sex can be interpreted as the relative difference in terms of standard deviation between students whose teacher's gender is the same and student's whose teacher's gender differ.

Regression (1a) shows a really similar pattern than the raw statistics I have presented in Table 3. There is a relatively high 0.073 standard deviation advantage of women teachers compared to men, while the difference is moderate 0.026 standard deviation in favor of boy pupils. Both coefficients are significant at a one percent level. There is a very weak support for a same sex pairing effect, because the coefficient of that variable is 0.009, which is only significant at a ten percent level.

After the inclusion of the mentioned control variables (1) the results changes dramatically. The better performance of women teacher decreases from 0.073 to a much more moderate level with a coefficient of 0.023, while the relative advantage of boys increases from 0.026 to 0.069. These result suggest that the pairing of students and teachers are not random, and this non randomness correlates with the sex of both participants.

In general, the decrease in the absolute value of teachers' between model (1a) and (1) suggests that good women teachers have higher probability to end up with more talented[4] students. One explanation can be to this phenomenon that as in most of the countries the vast majority of 8th grade teachers are women, schools trust more in the quality of women so they tend to employ more of them even if they are not as good as their men counterparts. It is also possible that within schools, women teachers are assigned more frequently to better classes. To investigate this question more in A2 I present separate regression results in which I excluded one type of the control variables in every specification. The results of this analysis suggest that from the change in the coefficient of $tfemale$, at least 0.01 standard deviation difference is explainable with the inclusion of school level control variables while 0.023 with individual level controls[5]. These results supports both hypothesizes.

A more surprising fact that not just the coefficient of the teacher gender but the coefficient of the sex of students changes dramatically with the inclusion of the control variables. The fact that in (1) boys achieve ceteris paribus much higher results than girl compared to (1a) means that similarly talented girls have higher probability to be paired with better teachers than boys, which covers the real difference between the two sexes. The possible explanation can be twofold as well, one that girls have better chances of admission to better schools, while it is

[4] I used the word "talent" in an econometric sense, referring to the fact as well, that children with lower socioeconomic status have lower potential test score.

[5] The difference does not sum up to 0.05 (the total difference in the coefficient of tfemale in the two specification) because the different types of control variables can take up the effect of each other.

also possible that girls have higher probability of getting a good teacher within school than boys. The analysis of A2 shows that the coefficient of $sgirl$ is only sensitive to the exclusion of individual characteristic which supports more the latter idea than the former.

As it is also visible the coefficient of the double interaction term remains constant across (1a) and (1) the difference between them is not significant. With the inclusion of the background characteristics the point estimate increased slightly to 0.012, significant at five percent level, which means that controlled for the differences in teacher and student performance by gender, holding everything else constant, those students achieve 0.012 standard deviation higher test results, whose teacher's sex is the same as themselves. This difference though is significant. It is very small in magnitude, which means that even if the presented mechanisms exists with respect to the pairing of pupils and teachers based on gender, these effect are not very relevant.

From the studies that I presented in the literature review section not many operates with standardized measures. One exception is the study of Dee (2007), who estimated the relative advantage of same sex pairing to be 0.042 standard deviation. From this it is evident, that I measured significantly lower effect as Dee (2007) in his own estimation.

It is also important to compare (1b) and (1). In estimation (1b) I did not include the double interaction term in order to see how does it biases the coefficient of $tfemale$ and $sgirl$. As it is visible from Table 4 the point estimates are not significantly different in the two types of specifications in both cases. This means that with respect to the control variables included, the double interaction term is not related statistically with the two gender dummies.

To conclude this subsection my main findings were that there is a selection based on quality which correlates with gender in the case of students and teachers as well. If I control for this selection by including background characteristics into the regression the relative advantage of women teacher decreases to a small level (0.023 standard deviation) while the advantage of

boys increases to relatively high level (0.069 standard deviation). The OLS regression present some proof that an additional pairing effect exist based on the sex of the student and the teacher as pupils with same sex teachers perform slightly better (0.012 standard deviation) than their different sex pair counterparts. On the other hand, this pairing effect is really small in magnitude, which questions its relevance. The magnitude of this pairing effect is not sensitive to the inclusion of background control variables.

Individual fixed effect

As I presented in the Identification strategy section of my paper, it is highly possible, that due to the limitations of the useable control variables the OLS estimation could not get rid of all the biases that can affect the estimation as the pairing of students and teachers are not random. For this reason, in this section I will present the result of the individual fixed effect estimation (3), in which the test scores of the same student is compared across teacher.

Similar to the OLS regression I present two more alternative specifications of equation (3):

$$(3a)\ y_{i,s,m,c} = \gamma_0 + \gamma_1 * tfemale_{s,m,c} + \gamma_2 * samesex_{i,s,m,c} + \kappa_s + \sigma_i + \varepsilon_{i,s,m,c}$$

$$(3b)\ y_{i,s,m,c} = \gamma_0 + \gamma_1 * tfemale_{s,m,c} + \gamma_3 * T_{s,m,c} + \kappa_s + \sigma_i + \varepsilon_{i,s,m,c}$$

Table 5 collects the coefficient of interest in the different specifications of equation (3) and the results of regression (1) as well, in order to make comparisons between the two types of models. As the main idea of the individual fixed effect estimation is to compare the same individual across different teachers, the sex of the student is included in the fixed effect, so only the coefficient of $tfemale$ and $samesex$ is reportable.

5. Table: The results of the individual fixed effect estimations, pooled across countries, own calculations

VARIABLES	(1) y	(3a) y	(3b) y	(3) y
tfemale	0.0232*** (0.00870)	0.0109*** (0.00413)	0.0102** (0.00419)	0.0106** (0.00419)
samesex	0.0116** (0.00455)	0.0236*** (0.00285)		0.0236*** (0.00285)
Controls	yes	no	yes	yes
2X interaction	yes	yes	no	yes
Observations	823,777	823,777	823,777	823,777
R-squared	0.174	0.849	0.849	0.849

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

In all specifications of (3) it can be concluded that the measured coefficients are the same. In the case $tfemale$ the estimated coefficient lies between 0.010-0.011, which means that according to the individual fixed effect regression, women teachers are approximately 0.01 standard deviation better than their male counterparts. This is a very small effect, and in the full model (3) it is only significant at five percent level. In the case of positive pairing of same sex teachers all estimation measures a higher but still quite small effect of 0.024 standard deviation, which means that students who have a teacher with same sex achieve 0.024 standard deviation higher test score ceteris paribus.

If I compare (3a) and (3) I can conclude, that the inclusion of control variables about class characteristics do not have effect on the regression outputs. Surprisingly the R^2 of the regression does not differ in the two specification despite the fact that from the included six control variables one was significant at a one percent, one at a five percent and one at a ten percent level. In order to see all coefficients of (3) check A3. The fact that the class level control

variables does not affect the coefficients of interest is not surprising, as in the OLS the exclusion of this variables did not have a significant effect either (A2). These findings are in line with the empirical results of other studies which shows that the quality of teachers are in many cases not correlatable highly with the education or experience level of the teachers, so it is really difficult to measure (Rivkin et al., 2005). The comparison between (3b) and (3) gives the same result as in the OLS case, that $tfemale$ and $samesex$ are quite independent.

On the other hand, the point estimates of (1) and (3) with respect to the variables of interest differs heavily. In the OLS regression I identified a mechanism that women teacher with the same quality of their men counterparts have higher probability of teaching more talented students. According to the results of the individual fixed effect regression including background control characteristic only terminates a portion of this non-random sorting, because the coefficient of $tfemale$ is just 0.011 standard deviation compared to the 0.023 in the OLS case. This means that the difference between men and women teachers shrinks almost to 0 if I take into account all the non-random pairing.

It is also important to note that the coefficient of $samesex$ was not sensitive to the different specifications of the OLS estimation. With the inclusion of individual fixed effect however it increased to 0.024 standard deviation from 0.012 in (1). This change can be the result of the fact, that the coefficient of $tfemale$ decreased in the fixed effect estimation, and the measured increase of men teachers between the two types of models were dominated by man-boy pairing. On the other hand, this is not necessarily the case. One disadvantage of the individual fixed effect estimation is that the gender of the student is absorbed by the fixed effect so it is possible, that the OLS estimation (1) did not captured correctly the performance difference between boys and girls, which was corrected in estimation (3).

In this subsection I showed an interpreted, the result of the individual fixed effect estimation. I found that in all specifications women teachers are approximately 0.011 standard deviation better then man, and if a pupil has a same sex teacher this will result in a 0.024 standard deviation advantage. From these results I concluded, that measures with which I tried to capture teacher quality and class characteristics are just slightly adequate for these goals (some of them are significant, but they do not change the R^2 of the regression significantly), and these variables are not correlated with gender or pairing effects. The inclusion of the double interaction term did not change the coefficient of the sex of teacher as well. On the other hand, the new result showed, that the included control variables in the OLS regression, were not able to account for all the non-random parring effect as the coefficients of interest changed to smaller relative advantage of women teachers and bigger same sex pairing effect.

Different countries

In the previous estimations I used a pooled dataset, which implicitly assumes that the quality difference between males and females, and the pairing effect is the same across countries. However, it is quite rational to assume that there can be country specific differences. For this reason, in this subsection I present the estimation results of equation (4) and (5), which are the same equations as (1) and (3), the only difference that instead of a pooled estimation I ran in each case twenty different regressions, for all included country separately.

The results of the OLS estimation already showed, that by simply controlling for different individual, class and school level characteristics cannot eliminate all of the non-random pairing effect, so in the pooled country case the individual fixed effect estimation seems to be superior relative to the OLS. The fact that controls are not enough, stayed valid as I estimated equations (4) and (5), because the coefficients of $tfemale$ and $samesex$ are not similar in the two types of estimation. For this reason, I do not report the results of the OLS estimations (4) in the main body of the text, it can be found in A4.

I can briefly summarize the OLS results by stating that the coefficient of $tfemale$ is insignificant in most of the countries, but this estimations are quite imprecise zeros as there are many coefficients which are much bigger than the estimated 0.023 in the pooled estimation (1), but insignificant at any conventional level. By simply looking to the point estimates[6] there is a very large heterogeneity between countries as they vary from -0.056 to 0.162. In the case of $samesex$ nearly half of the coefficient are significant at least at ten percent level all of them with a positive coefficient. The point estimates however are quite different as well as there are many countries where the coefficient are negative and relatively large in absolute terms, but not significant.

As the gender related variables of equation (4) did not capture the casual effects well, I will place more emphasis on the presentation of the individual fixed effect estimation. In Table 6 I collected the coefficients of $tfemale$ and $samesex$ from regression (5). Similarly to the pooled retrogressions γ_1 shows that controlled for everything else, how much standard deviation the pupils of female teachers achieve more points than the students with male ones, while γ_2 represents the relative advantage of same sex pairing over different sex pairing in terms of standard errors, ceteris paribus. The only difference that as I ran twenty separate regressions I have a point estimate for both parameters for every country.

[6] In every estimation where I separately investigated the twenty countries significance level can be misleading, as the observation numbers are very different because not all countries participated in all waves (for the exact observation numbers see Table 1.) For this reason, it can meaningful to compare the point estimates even in the case of non-significance.

6. Table: The results of the individual fixed effect estimations for separate countries, own calculations

Countries (5)	tfemale	samesex
Bgr	0.0361	0.0117
	(0.0309)	(0.0147)
Bih	-0.00303	0.0199
	(0.0181)	(0.0128)
Cyp	0.0194	0.0265**
	(0.0125)	(0.0111)
Cze	0.0148	0.0452***
	(0.0180)	(0.0116)
Est	0.0428	-0.0436*
	(0.0284)	(0.0229)
Fin	-0.00986	-0.0121
	(0.0209)	(0.0150)
Hun	0.00800	0.0413***
	(0.0101)	(0.00718)
Ltu	0.0101	0.0474***
	(0.0154)	(0.0111)
Lva	-0.0128	0.0861***
	(0.0297)	(0.0186)
Mda	0.0152	0.0333***
	(0.0214)	(0.0116)
Mkd	0.0238	0.0106
	(0.0159)	(0.00745)
Mlt	0.0321*	0.0258
	(0.0182)	(0.0176)
Nld	0.0148	0.0302
	(0.0269)	(0.0185)
Rou	-0.00182	-0.000182
	(0.0138)	(0.00743)
Rus	0.0345**	0.0493***
	(0.0170)	(0.0131)
Srb	0.0309***	0.0172*
	(0.0113)	(0.00941)
Svk	0.0276	0.0365
	(0.0268)	(0.0241)
Svn	-0.0101	0.0208
	(0.0208)	(0.0136)
Swe	0.00719	0.0287*
	(0.0179)	(0.0158)
Ukr	-0.000488	0.0590***
	(0.0184)	(0.0127)

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

First I will analyze the difference between male and female teachers in the twenty countries, by looking at the values of γ_1. In seventeen out of the twenty countries the coefficients $tfemale$ is insignificant. The only significant results can be found in Malta, Russia and Serbia, (at ten percent, five percent and one percent level respectively) where female teachers are approximately 0.03 standard deviation better. The only country where the point estimate is bigger, then the three significant value in absolute terms is Estonia, with 0.043, while in the case of Bulgaria I got a 0.036 standard deviation difference, both of them insignificant at any conventional level. Generally, the results are mostly positive and very close to each other as the lowest value is -0.012 while the biggest is 0.043, which means that there is not much heterogeneity between the relative performance of men and women.

From these the results the following conclusion can be drown. In the selected twenty countries there is no significant difference between male and female teachers. This result is supported both by the fact, that points estimates are very small as the average absolute effect[7] is 0.018 and the average effect is 0.014 standard deviation and that only three results were significant. It is also important to note, that during the estimation I ran twenty separate regressions so it is highly possible, that these three countries are significant because of multiple inference. Multiple inference can occur if a researcher estimates the same regression on many subsets of a sample because it is possible that some significant results will emerge not because of a real connection, but due to the statistical properties of hypothesis testing (Bland and Altman, 1995). It is possible to correct the p values of the regression, but there is no self-evident way of doing it, so correcting for this issue extends beyond the limits of this research, but it is important to note that this can weaken further the results. So despite the fact that teacher's profession is highly contra selective and the portion of female teachers varies heavily between countries (Eide et

[7] This is the average of the absolute value of the coefficients.

al., 2004), this doesn't result in quality difference between male female teachers based on their students' test points.

The situation is quite different in the case of $samesex$. In eleven countries the regressions estimated significant effect at a ten percent level[8]. In ten cases the connection is positive which means that same sex pairing of student and teacher results in higher test score, while according to my results in Estonia different gender pairing is beneficial. On top of that in the case of Slovakia, the Netherlands and Malta, the estimated coefficients are quite high, despite the fact that they are not significant even at a ten percent level. In order to give a more transparent interpretation of the results in Table 7 I categorized the twenty countries based on the point estimate of γ_2.

less than 0.02	0.02-0.04	0.04-0.06	more than 0.06
Bulgaria	Cyprus**	Czech Republic***	Latvia***
Bosnia	Moldova**	Estonia*	
Finland	Malta	Hungary***	
Macedonia	Netherlands	Lithuania***	
Romania	Slovakia	Russia***	
Serbia*	Slovenia	Ukraine***	
	Sweden*		

7. Table: Countries categorized based on the magnitude of pairing effect, own calculations

The columns of the table represent the ranges of the absolute value of the point estimate of variable $samesex$ in terms of standard deviations, so in every column all countries are listed where the absolute value of the coefficient falls in the given range. I marked those countries with red, where the coefficient is negative, and the stars after the name shows that at what percent the corresponding point estimate is significant[9]. From this table it is visible that the effect of same sex pairing is very heterogeneous, there are countries where the magnitude of this effect is modest, while in other places it is really close to zero. The average of the

[8] From these eleven countries, in eight the results were significant at five and in seven even at one percent level.

[9] The stars represent the same significance levels as in the regression throughout my paper.

coefficients is 0.027 which is very close to result of the pooled estimation (3), while the average absolute effect is 0.033 standard deviations. But if I calculate the same values for the countries where I got significant results these numbers increase 0.035 and 0.043 respectively. It is important to note that multiple inference can be a problem in this estimation as well, but as more than half of the coefficients are significant probably even a correction would not eliminate all effect. So according to these result it can be concluded, that positive same sex paring effect is not very high in magnitude, but is present in many of the investigated countries, however the size of it is very heterogeneous across nations.

In this subsection I estimated separately for all participating twenty countries the OLS (4), and the individual fixed effect (5) regressions. I already showed in the pooled case that the OLS regression because the lack of proper control variables, cannot account for all the non-random pairing effect of teachers and students, because it gives different results than the individual fixed effect estimation. When I ran the regressions separately for the different countries these, differences were more severe, for which I did not analyze the results of the OLS estimation (4), as it did not show the casual effect of being a women teacher and having a same sex teacher on test scores.

In the individual fixed effect estimations (5) I reported the coefficients of $tfemale$ and $samesex$. With respect to the benefit of having a women teacher, the separation of countries did not provide many additional information. The coefficients were insignificant in most of the case usually associated with the low point estimate value, which were very similar between countries. So I concluded that there is no significant difference between men and women teachers based on the results.

By the investigation of $samesex$ variable however the separation of countries provided additional insight. The results showed that same sex pairing effect is very heterogeneous across countries, there are places where it is very small or non-existent, but also lots of countries where the magnitude of it can be considered moderate. So my results robustly showed, that there is educational benefit for students who have same sex teachers, but the magnitude of this advantage is heavily dependent on which country the given student lives.

Conclusion and possibilities for further research

In this paper I investigated the effect of teacher's gender on student's performance. My main aim was to find an answer for two question. The first one was whether the quality of men or women teachers are better, while the second was whether same or different sex pupil teacher pairings leads to better educational results. The two questions are necessarily interlinked as if there is a performance difference between male and female teachers, it can directly affect the pairing effect.

I investigated the two questions with regression analysis using the TIMSS international dataset on twenty European countries. The dataset contains the test score of 8th grade children in mathematics, physics, chemistry, biology and earthly science. First I pooled the data across, waves, countries and subject and ran an OLS estimation, where the dependent variable was the standardized test score of a given student in a given subject, while the main explanatory variable of interest were a dummy which showed whether the teacher of the pupil in the given subject was female, and a double interaction term that showed whether the sex of the student is similar with the teacher's.

Unfortunately, in a real word setting, the pairing of students with teachers are not random, and even after controlling for individual level, class level and school level characteristics, the OLS estimation cannot identify properly the casual relationship between, the variables of interest and the measure of educational performance.

Because of this bias I turned to a common strategy which is used often in the literature, to identify the effect of teacher from individual variation of student test score, so I estimated a pooled individual fixed effect regression as well. TIMSS dataset is very appropriate for this purpose as an implicit assumption of this method is that the potential outcome of a given individual is the same in the different subject, which is as likely as possible to hold as I compared natural science subject. On top of that I had test results from 5 different subjects from

an individual, which highly increased the precision of the estimation relative to a pairwise comparison.

The results of the pooled estimation showed that the relative advantage of women teachers are 0.011 standard deviations which was significant at a five percent level, but in magnitude it is almost non-existent. I also concluded, that students with same sex teachers achieve 0.024 standard deviation better results.

In order to test for the heterogeneity of these two effects across countries, I estimated the individual fixed effect regression separately for the twenty participating countries. The results were quite similar in the case of female advantage with the pooled estimation, as in almost all countries the coefficient of female teachers were insignificant with very small point estimates. From these findings I concluded, that despite the fact that the teacher profession is very contra selective and that the portion of female teachers varies heavily between countries, there is no economically relevant difference, between male and female teachers in quality.

In the case of same sex pairing effect, the country level estimations provided additional information. First that in eighteen out of twenty countries the coefficient was positive indicating that same sex pairing is beneficial. Second, that the magnitude of this effect was very heterogeneous across countries (in absolute terms it varied between 0 and 0.08), so there were countries where the relevance of this pairing was close to zero, and also countries where moderate advantage was measured. By averaging out only the countries where I was able to measure significant results (11 out of 20), the average coefficients value increased to 0.035 while the average effect size (considering absolute values) was 0.043. These results are really close to the findings of Dee (2007) who estimated a 0.042 standard deviation large effect. So based on my regression I concluded, that having a same sex teacher is beneficial in terms of educational performance, however the size of this effect is quite different across countries.

However, my findings are important with respect to the problem whether same sex pairing is

beneficial or not, these results did not provide answer for the question what mechanism stands behind the positive same pairing effect. Dee (2007) shows two theoretical explanations for the phenomena, the first one is that the students are more likely to think about their teacher as role-model if their gender is the same, while the second is that teachers are more likely to discriminate pupils from different gender. Sadly, based on my results it is not possible to distinguish between the two type of mechanism. To make this distinction future researches should try to explain the heterogeneity of the coefficient of same sex pairing by country level characteristic, which are related to the willingness of discrimination, or rerun the estimations in smaller subsamples where only one of the above mentioned mechanisms can be convincible.

Appendix

A1-Control variables used in the regressions:

Class level control variables (1), (3)

- age of teacher: a categorical variable with 6 categories[10], which are less than 25, 25-29, 30-39, 40-49, 50-59 and more than 60 years.
- years taught by the teacher
- completed education level of the teacher: A categorical variable with 3 categories. In most of the countries the education level of 8th grade teacher are the same but the exact level (in isced categories) are different by countries. This variable shows whether the given teacher's education is below, equal or higher than this country specific level.
- type of teachers education: A binary variable which shows for teachers whether there primary education were mathematics or science[11] teacher respectively.
- number of students participating in the class in the given subject
- classes minutes held/week in the given subject

Individual control variables (1), (3):

- immigrant: Binary variable, which is one if the pupil was not born in the given country.
- the number of books at home: categorical variable with 5 categories, which are less than 10-nél, 10-25, 26-100, 101-200 and more than 200 books.
- Education level of parents: Different variable for the mother and the father, categorical variable based on isced levels.

[10] This is a categorical variable in the original dataset as well.
[11] In the case of this variable science is not divided into physics, chemistry, biology and earthly science.

School control variables (1):

- size of the settlement where the school is located: categorical variable with 6 categories, which are less than 3000, 3001-15000, 15001-50000, 50001-100000, 100001-500000 and more than 500000 inhabitants.
- the ratio of economically disadvantaged children: A categorical variable with four categories which are 0-10%, 11-25%, 26-50% and more than 50%.
- the ratio of economically affluent children: The categories are the same as in the disadvantaged case.
- days spent with education in a year
- total number of students of the given school

In the case of categorical variables, I created a different group for the missing observations while if the variable was continuous I replaced the missing value by the corresponding mean by country and subject.

A2- The importance of the different type of control variables in OLS

In all cases I estimated a regression, where one type of the control variables were missing.

Estimated equations:

$$(1c)\ y_{i,s,m,c} = \beta_0 + \beta_1 * tfemale_{s,m,c} + \beta_2 * sgirl_{i,m,c} + \beta_3 * samesex_{i,s,m,c} + \beta_5 * T_{s,m,c} + \beta_6 * S_{s,c} + \kappa_s + \forall_c + \varepsilon_{i,s,m,c}$$

$$(1d)\ y_{i,s,m,c} = \beta_0 + \beta_1 * tfemale_{s,m,c} + \beta_2 * sgirl_{i,m,c} + \beta_3 * samesex_{i,s,m,c} + \beta_4 * X_{i,m,c} + \beta_6 * S_{s,c} + \kappa_s + \forall_c + \varepsilon_{i,s,m,c}$$

$$(1e)\ y_{i,s,m,c} = \beta_0 + \beta_1 * tfemale_{s,m,c} + \beta_2 * sgirl_{i,m,c} + \beta_3 * samesex_{i,s,m,c} + \beta_4 * X_{i,m,c} + \beta_5 * T_{s,m,c} + \kappa_s + \forall_c + \varepsilon_{i,s,m,c}$$

VARIABLES	(1c) y	(1d) y	(1e) y	(1) y
tfemale	0.0453*** (0.0101)	0.0215** (0.00862)	0.0321*** (0.00890)	0.0232*** (0.00870)
sgirl	-0.0306*** (0.00659)	-0.0692*** (0.00599)	-0.0691*** (0.00605)	-0.0693*** (0.00597)
samesex	0.00828* (0.00500)	0.0120*** (0.00455)	0.0118** (0.00463)	0.0116** (0.00455)
X	no	yes	yes	yes
T	yes	no	yes	yes
S	yes	yes	no	yes
Observations	823,777	823,777	823,777	823,777
R-squared	0.044	0.171	0.165	0.174

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

A3-All coefficients of the individual fixed effect estimation (3)

VARIABLES	(3) y
Gender variables	
tfemale	0.0106** (0.00419)
samesex	0.0236*** (0.00285)
Subject dummies	
physics	0.0272*** (0.00566)
biology	0.0209*** (0.00620)
chemistry	0.0212*** (0.00600)
earth sci	0.0402***

	(0.00626)
Age of teacher	
25-29	0.000323
	(0.0127)
30-39	0.000829
	(0.0121)
40-49	-0.00660
	(0.0128)
50-59	-0.0118
	(0.0144)
60+	-0.0165
	(0.0169)
Students in class	0.00107*
	(0.000635)
Years taught	0.000714**
	(0.000306)
Education level	
average	0.00819
	(0.00690)
high	0.00995
	(0.00881)
Primary education subject teacher	0.00299
	(0.00438)
Minutes/week	0.000173***
	(4.60e-05)
Constant	-0.0750***
	(0.0215)
Observations	823,777
R-squared	0.849

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

A4- Results of OLS estimation for different countries (4)

Countries	tfemale	sgirl	samesex
Bgr	-0.0515	-0.0238	-0.0169
	(0.0483)	(0.0314)	(0.0261)
Bih	0.0215	-0.0356	0.0341*
	(0.0393)	(0.0288)	(0.0185)
Cyp	0.00868	0.0725***	0.0328**
	(0.0165)	(0.0208)	(0.0133)
Cze	-0.0607	-0.255***	0.0461**
	(0.0399)	(0.0279)	(0.0217)
Est	-0.00113	0.0952**	-0.0375
	(0.0556)	(0.0375)	(0.0283)
Fin	-0.0177	-0.0936***	-0.0222
	(0.0305)	(0.0266)	(0.0204)
Hun	0.0295	-0.190***	0.00803
	(0.0223)	(0.0152)	(0.0112)
Ltu	-0.0364	-0.0563**	0.0324**
	(0.0309)	(0.0222)	(0.0158)
Lva	0.0383	-0.151***	0.0360
	(0.0482)	(0.0304)	(0.0233)
Mda	-0.0716	-0.0303	0.0308*
	(0.0467)	(0.0228)	(0.0164)
Mkd	0.162***	-0.0160	0.0289***
	(0.0298)	(0.0158)	(0.0108)
Mlt	0.146***	-0.229***	0.157***
	(0.0514)	(0.0661)	(0.0512)
Nld	-0.0340	-0.216***	-0.0130
	(0.0789)	(0.0256)	(0.0191)
Rou	-0.0563**	-0.0229	-0.00691
	(0.0273)	(0.0179)	(0.0120)
Rus	0.0206	-0.0951***	0.00364
	(0.0413)	(0.0238)	(0.0197)
Srb	0.0800***	-0.0174	0.00959
	(0.0258)	(0.0200)	(0.0138)
Svk	0.0323	-0.231***	0.0608***
	(0.0460)	(0.0251)	(0.0184)
Svn	0.0799**	-0.0627**	0.0302*
	(0.0343)	(0.0260)	(0.0171)
Swe	-0.0221	-0.0503***	0.0106
	(0.0279)	(0.0179)	(0.0163)
Ukr	0.0240	-0.115***	0.0347
	(0.0363)	(0.0319)	(0.0212)

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

References

Ammermüller, Andreas – Dolton, Peter **(2006):** Pupil-teacher gender interaction effects on scholastic outcomes in England and the USA, *ZEW-Centre for European Economic Research Discussion Paper*, 06-060, http://www.econstor.eu/bitstream/10419/24515/1/dp06060.pdf

Antecol, Heather - Eren Ozkan – Ozbeklik, Serkan **(2015):** The Effect of Teacher Gender on Student Achievement in Primary School, *Journal of Labor Economics*, 33, 1, pp. 63-89, http://www.econstor.eu/bitstream/10419/58932/1/715806394.pdf

Bettinger, Eric P. - Long, Bridget T. **(2005):** Do faculty serve as role models? The impact of instructor gender on female students, *The American Economic Review*, 95, 2, pp. 152-157, http://www.jstor.org/stable/pdf/4132808.pdf?acceptTC=true

Bland, J. Martin - Altman Douglas G. **(1995)**: Multiple significance tests: The Bonferroni method, *British Medical Journal*, 310, 6973, p. 170, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2548561/pdf/bmj00576-0038.pdf

Carrington, Bruce - Tymms, Peter – Merrell, Christine **(2008):** Role models, school improvement and the 'gender gap'—do men bring out the best in boys and women the best in girls?, *British Educational Research Journal*, 34, 3, pp. 315-327, https://vpn.ceu.edu/+CSCO+0h756767633A2F2F6A6A6A2E77666762652E626574++/stable/pdf/40375493.pdf?acceptTC=true

Dee, Thomas S. **(2004):** Teachers, race, and student achievement in a randomized experiment, *Review of Economics and Statistics*, 86, 1, pp. 195-210, http://media.hoover.org/sites/default/files/documents/ednext20042unabridged_dee.pdf

Dee, Thomas S. **(2005):** A teacher like me: Does race, ethnicity, or gender matter?, *American Economic Review*, 95, 2, pp. 158-165, http://www.jstor.org/stable/pdf/4132809.pdf?acceptTC=true

Dee, Thomas S. **(2007)**: Teachers and the gender gaps in student achievement, *Journal of Human Resources*, 42, 3, pp. 528-554, https://core.ac.uk/download/files/153/6853897.pdf

Eide, Eric - Goldhaber, Dan – Brewer, Dominic **(2004)**: The teacher labour market and teacher quality, *Oxford Review of Economic Policy*, 20, 2, pp. 230-244, http://phd.mshaffer.com/projects/teacherQuality.pdf

Gershenson, Seth - Holt, Stephen B. - Papageorge, Nicholas W. **(2015)**: Who Believes in Me? The Effect of Student-Teacher Demographic Match on Teacher Expectations, *Institute for the Study of Labor*, Discussion paper, No. 9202, https://www.econstor.eu/bitstream/10419/114079/1/dp9202.pdf

Holmlund, Helena – Sund, Krister **(2008)**: Is the gender gap in school performance affected by the sex of the teacher?, *Labour Economics*, 15, 1, pp. 37-53, https://vpn.ceu.edu/+CSCO+00756767633A2F2F6E702E7279662D7071612E70627A++/S0927537106000947/1-s2.0-S0927537106000947-main.pdf?_tid=a90c865c-f0a3-11e5-9ac6-00000aab0f6b&acdnat=1458702323_91b46ce0cc32f42470f871b89ef4dcea

Kane, Thomas. J.- Taylor, Eric S. - Tyler, John H. - Wooten, Amy L. **(2011):** Identifying effective classroom practices using student achievement data, *Journal of human Resources*, 46, 3, pp. 587-613, http://danielsongroup.org/wp-content/uploads/2014/04/IdentifyingEffectiveClassroomPractices.pdf

Lavy, Victor **(2004)**: Do gender stereotypes reduce girls' human capital outcomes? Evidence from a natural experiment, No. w10678, *National Bureau of Economic Research*, https://core.ac.uk/download/files/153/6502016.pdf

OECD **(2015)**: *The ABC of Gender Equality in Education: Aptitude, Behavior, Confidence*, OECD, PISA, OECD Publishing, ISBN 978-92-64-22994-5, http://dx.doi.org/10.1787/9789264229945-en

Rivkin, Steven G. – Hanushek, Eric A. – Kain, John F. **(2005)**: Teachers, schools, and academic achievement, *Econometrica*, 73, 2, pp. 417-458, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.322.4872&rep=rep1&type=pdf

Terrier, Camille **(2015)**: Giving a little help to girls? Evidence on grade discrimination and its effect on students' achievement, *Center for Economic Performance*, Discussion Paper, No. 1341, http://eprints.lse.ac.uk/61696/1/dp1341.pdf

Data:

Trends in International Mathematics and Science, 2011, 2007, 2003, http://timssandpirls.bc.edu/