CHARLES UNIVERSITY
FACULTY OF SOCIAL SCIENCES
Institute of Economic Studies

Gender Gap in Math Score: Does Teacher Gender Matter?

Master's thesis

Author: Bc. Šimon Scharf
Study program: Economics and Finance
Supervisor: Mgr. Barbara Pertold-Gebicka M.A., Ph.D.
Year of defense: 2023

Declaration of Authorship

The author hereby declares that he or she compiled this thesis independently, using only the listed resources and literature, and the thesis has not been used to obtain any other academic title.

The author grants Charles University permission to reproduce and to distribute copies of this thesis in whole or in part and agrees with the thesis being used for study and scientific purposes.

Prague, January 2, 2023

Šimon Scharf

Abstract

Even though quality and equal opportunities are regarded as generally desirable in education, major differences in the study outcomes of girls and boys still exist. In this thesis, we try to assess the effect of a teacher's gender on the educational outcomes of pupils. Specifically, we use TIMSS data from 36 countries to evaluate this effect on 4^{th} grade students. To our knowledge, we are the first to utilize the Propensity Score Matching (PSM) approach to overcome the selection bias in this context. The results of the pooled analysis suggest that there is no significant effect of teachers' gender on girls but we observe a negative effect for boys. When considering each country separately, only in 4 countries do we find a significant effect of teacher's gender on students' test scores for boys. Of the 4 countries, only boys in Montenegro prosper with a same-sex teacher, while in 3 countries boys' achievement is hampered by a same-sex teacher. For girls, we find a robust positive effect in 4 countries and a negative effect in 3 countries. For both boys and girls, we find no significant robust effect of having a same-sex teacher in the majority of countries. Our findings contribute to the literature on the effects of teachers' gender, as well as, to the broader discussion of differences in the educational attainment of boys and girls.

JEL Classification	I2, I21, I24, I20, F00
Keywords	Gender gap in math achievement, Propensity score matching, TIMSS standardized tests, Effect of teachers' gender, International analysis
Title	Gender Gap in Math Score: Does Teacher Gender Matter?

Abstrakt

Přestože jsou kvalita a rovné příležitosti ve vzdělávání vnímány jako obecně prospěšné, stále existují značné rozdíly ve studijních výsledcích chlapců a dívek. V této práci se pokusíme vyhodnotit jaký má efekt pohlaví učitele na studijní výsledky žáků. Konkrétně využijeme data TIMSS z 36 různých zemí abychom posoudili tento efekt na žáky a žačky čtvrtých tříd. Jsme první, kdo používá metodu Propensity Score Matching (PSM) k překonání výběrové chyby v tomto kontextu. Výsledky analýzy sdružených dat naznačují, že pohlaví učitele nemá žádný významný vliv na dívky, ale pozorujeme negativní efekt na chlapce. Když jsme uvažovali každou zemi zvlášť, pouze ve 4 zemích jsme našli statisticky významný efekt pohlaví učitele na výsledky studentů u chlapců. Z těchto 4 zemí pouze chlapci v Černé Hoře benefitovali z učitele stejného pohlaví, zatímco ve zbylých 3 zemích to jejich výsledky zhoršilo. U dívek je efekt robustní a pozitivní ve 4 zemích a negativní ve 3 zemích. Jak pro dívky tak pro chlapce jsme ve většině zemí nenašli žádný statisticky významný efekt. Naše výsledky přispívají k literatuře zabývající se vlivy pohlaví učitele na výsledky žáků ale také k širší diskuzi ohledně rozdílů v dosaženém vzdělání mezi chlapci a dívkami.

Klasifikace JEL	I2, I21, I24, I20, F00
Klíčová slova	Genderová nerovnost v matematických výsledcích, Propensity score matching, TIMSS standardizované testy, Efekt pohlaví učitele, Mezinárodní analýza
Název práce	Genderová Propast v Matematických Výsledcích: Má Pohlaví Učitele Vliv?

Acknowledgments

I wish to express my deepest gratitude to Mgr. Barbara Pertold-Gebicka M.A., Ph.D., for her patience and wise counsel, as well as expert comments and general guidance throughout the process of writing this thesis.

Furthermore, I am thankful to my family for their support throughout the entire period of my studies. I would also like to thank my friends for motivating and inspiring me.

Typeset in LaTeXusing the IES Thesis Template.

Bibliographic Record

Scharf, Šimon: *Gender Gap in Math Score: Does Teacher Gender Matter?*. Master's thesis. Charles University, Faculty of Social Sciences, Institute of Economic Studies, Prague. 2023, pages 112. Advisor: Mgr. Barbara Pertold-Gebicka M.A., Ph.D.

Contents

List of Tables

List of Figures

Acronyms

TIMSS Trends in International Mathematics and Science Study
UN United Nations
SDGs Sustainable Development Goals
STEM Science, Technology, Engineering, and Math
OECD Organization for Economic Cooperation and Development
OLS Ordinary Least Squares regression
PSM Propensity Score Matching
NELS88 National Education Longitudinal Study of 1988
ECLS—K Early Childhood Longitudinal Study Kindergarten Cohort
NLSY79 National Longitudinal Survey of Youth
FLDOE Florida Education Data Warehouse
MPR Mathematica Policy Research, Incorporated
NETFA National Evaluation of Teach for America
PISA Programme for International Student Assessment
NAEP National Assessment of Educational Progress
TFA Teach for America
IEA International Association for the Evaluation of Educational Achievement
CIA Conditional Independence Assumption
NN Nearest Neighbour
ATT Average Treatment Effect on the Treated
MICE Multiple Imputation by Chained Equations
MCAR Missing Completely at Random

Master's Thesis Proposal

Author	Bc. Šimon Scharf
Supervisor	Mgr. Barbara Pertold-Gebicka M.A., Ph.D.
Proposed topic	Gender Gap in Math Score: Does Teacher Gender Matter?

Motivation The difference between girls and boys in standardized test scores — commonly known as the gender gap has sparked the interest of researchers as hypothetically there should be none. In math and science, this gap historically favoured boys, while in reading or writing girls were more successful. As the prevailing ambition among policymakers is to provide equal educational opportunities, there is an acute need to explain these systematic differences. I am planning to focus mainly on the gap in Mathematics. Various explanations of this phenomenon have appeared in the literature. For example, Guiso et al. (2008) found a link between test score differences and indicators of gender equality. On the contrary, later revisitation by Anghel et al. (2020) shows that this link vanishes once the country fixed effects are accounted for, yet the link still holds for poor countries. Apart from the societal inequality, some authors tried to explain the gap by cultural family background. For instance, Dossi et al. (2020) used fertility stopping rules to show that girls in families with boy preference score lower than girls in other families, as well as finding that maternal gender role attitudes have a similar impact. In summary, Dossi et al. (2020) claim that family background may explain part of the observed gap. Conversely, Kim and Law (2011) found little support for the family background effect and also showed the non-trivial impact of single-sex schooling. I shall deviate from these explanations and instead, I shall add to another vast branch of literature that tried to evaluate the effect of a teacher's gender on students' performance. So far, the literature offers several mixed results. Krieg (2005) followed 3rd graders in the state of Washington for two years but found no significant impact of a same-sex teacher on student performance. Dee (2007) first exploited the matching pairs strategy to control for student fixed effects in longitudinal data (National Education Longitudinal Study of 1988) and found that assignment to same-sex teachers significantly

improves students' results regardless of student gender. Carrell et al. (2010) examined the topic in college settings, the results suggest that although there is little effect of teacher gender on male students, female students are significantly affected. Winters et al. (2013) found no significant impact of a same-gender teacher on student performance in Florida panel. As for some cross-country comparisons, Cho (2012) utilized the data from the Trends in International Mathematics and Science Study (TIMSS) to evaluate the effect of same-gender teachers in 15 OECD countries and found large heterogeneity and overall little support for the significance of this effect. The identification strategy is similar to the one used by Dee (2007) and accounts for student fixed effects. Diallo and Hermann (2017) also use TIMSS data on 20 European countries to evaluate the differences between Western and Eastern Europe. Their results suggest that same-gender teacher benefits mostly students in Western Europe. As mentioned above the results are still rather inconclusive and point to large cross-country heterogeneity. The papers mentioned in the previous paragraph mainly use the first difference identifying strategy. An advantage of this method is that it identifies a causal effect, however, due to its nature, it is not applicable to be used for 4^{th} graders in the TIMSS environment. Since the results of TIMSS 2019 seem to point out to expansion of the gender gap for 4^{th} grade I believe it is worth examining the teacher-student same gender effect on younger pupils. Specifically, if the student fixed effects are correlated with the assignment to a same-sex teacher (treatment) then the coefficient of this variable would be biased. Therefore, I plan to utilize (a different identification strategy that should reduce the selection bias) Propensity Score Matching to examine whether the gender of the teacher matters for the gender math gap among 4^{th} grade students.

Hypotheses As mentioned above, the main goal of this thesis is to assess the effect of being taught by a same-gender teacher for girls and boys among 4^{th} grade students using TIMSS data. To get an unbiased estimator using OLS, the selection to treatment should be random - uncorrelated with student characteristics. However, there exists a concern that this is not the case, as in Cho (2012): "For example, if students in lower academic tracks are more likely to be assigned to female teachers, this nonrandom assignment creates a negative correlation between teacher gender and unobservable student ability and causes a bias in the coefficient for the gender matching variable." Hence, I plan to test the randomness of selection to treatment using observable data available at the time of assignment to treatment.

Hypothesis #1: Having a male teacher improves the math score for boys.

Hypothesis #2: Having a female teacher improves the math score for girls.

Hypothesis #3: Selection to treatment (having a male teacher) is non-random.

Methodology The intended identification strategy is Propensity Score Matching. This method aims to replicate the randomized experiment by balancing covariates. In the first stage, the probability of assignment to treatment, i.e. to having a male teacher, will be calculated using data from the Early Learning Survey. This survey is part of TIMSS and is intended to be completed by students' parents. The questions are specifically aimed at children's abilities and characteristics before they start school. For their timing, I believe that these data are exactly the ones that would determine the probability of being selected for treatment. I shall estimate this probability (or propensity score) using logistic regression. The first stage results should also provide us with a test of hypothesis 3. I expect the selection to be non-random, but in the unlikely case that all explanatory variables in the logistic model are jointly insignificant (selection to treatment is random) I could move to a simple OLS approach to test hypotheses 1 and 2. After obtaining the propensity score from the fitted model, I move to the second stage. There are several approaches of how to proceed in the second stage. The most straightforward is a one-to-one matching of observations from the treatment and control group with the most similar propensity score using the nearest neighbour technique. Another option is to use stratification - dividing the propensity score distribution into several groups and weighting. Lastly, I may use a weighting algorithm (Hogrebe and Strietholt, 2016) - for example, kernel matching. I plan to perform leave-one-out validation to choose the best matching algorithm. Finally, to obtain standard errors bootstrap techniques will be used. A preliminary inspection of TIMSS 2019 data showed large heterogeneity among countries when it comes to the proportion of students in 4^{th} grade taught by a male teacher. While in Latvia it is less than 0.5% of students, in Saudi Arabia it is over 49%. Such heterogeneity suggests that the effect may not always be easily generalizable to a country's population and so the results will probably be of higher value to countries where a larger proportion of the population is treated. Therefore, I intend to provide a cross-country comparison.

Expected Contribution The literature on teacher gender effects so far examined mainly teenage students and used the first difference estimating strategy to identify the effects. Such an approach does not allow researchers to estimate the effects for 4^{th} graders, because 4^{th} grade students are usually taught by one teacher in all subjects. The novelty of my work is a new identification strategy that should allow us to estimate the effect of being taught by a same-gender teacher for the TIMSS data for 4^{th} grade. So far, this effect has not been evaluated for these students using TIMSS. Also, the latest data from TIMSS 2019 will be used. The results should add to the literature on the effect of teachers' gender on student performance as well as to more general literature trying to explain gender gaps in standardized test scores.

Finally, the results may provide useful implications for actual policy.

Outline The thesis will follow this structure.

1. Introduction - the motivation behind studying this topic will be disclosed as well as its' relevance, moreover, possible contributions will be highlighted.
2. Literature review - both the literature on the gender gap and the effect of the teachers' gender will be reviewed.
3. Data - description of the data used for testing each of the hypotheses.
4. First stage model and results - in this section, the approach to determine the probability of being treated will be described and the results will be presented.
5. Second stage model and results - based on the results from the previous section the model will be chosen and estimated, then the best model will be evaluated.
6. Discussion - the results, their meaning and implications for practice, and the relevance of the employed method will be discussed.
7. Conclusion - the main findings of the thesis will be highlighted alongside key policy implications and possible avenues for future research.

Core bibliography

ANGHEL, Brindusa, Núria RODRÍGUEZ-PLANAS and Anna SANZ-DE-GALDEANO. Is the math gender gap associated with gender equality? Only in low-income countries. Economics of Education Review. 2020, 79(C).

CARELL, Scott E., Marianne E. PAGE and James E. WEST. Sex and Science: How Professor Gender Perpetuates the Gender Gap. The Quarterly Journal of Economics. 2010 125(3), p. 1101-1144.

CORDERO, José M., Víctor CRISTÓBAL and Daniel SANTÍN. Causal Inference on Education Policies: A Survey of Empirical Studies Using PISA, TIMSS AND PIRLS. Journal of Economic Surveys. 2017, 32(3), p. 878-915.

CHO, Insook. The effect of teacher-student gender matching: Evidence from OECD countries, Economics of Education Review, 2012, 31(3), p. 54-67.

DEE, Thomas S. Teachers and the Gender Gaps in Student Achievement. The Journal of Human Resources, 2007, 42(3), p. 528-554.

DOSSI, Gaia, David FIGLIO, Paola GIULIANO and Paola SAPIENZA. Born in the Family: Preferences for Boys and the Gender Gap in Math. CEPR Discussion Papers 13504. 2019.

GUISO, Luigi, Ferdinando MONTE, Paola SAPIENZA and Luigi ZINGALES. Culture, Gender, and Math. Science. 2008, 320(5880), p. 1164-1165.

HERMANN, Zoltán and Alfa DIALO. Does teacher gender matter in Europe? Evidence from TIMSS data. Budapest Working Papers on the Labour Market 1702, Institute of Economics, Centre for Economic and Regional Studies. 2017.

HOGREBE, Nina a Rolf STRIETHOLT. Does non-participation in preschool affect children's reading achievement? International evidence from propensity score analyses. Large-Scale Assessments in Education. 2016, 4(1), p. 1-22.

KIM, Doo Hwan a Helen LAW. Gender gap in maths test scores in South Korea and Hong Kong: Role of family background and single-sex schooling. International Journal of Educational Development. 2012, 32 (1), p. 92-103.

KRIEG, John M. Student gender and teacher gender: What is the impact on high stakes test scores?, Current Issues in Education. 2005, 8(9).

WINTERS, Marcus A., Robert HAIGHT, Thomas SWAIM and Katarzyna A. PICKERING. The effect of same-gender teacher assignment on student achievement in the elementary and secondary grades: Evidence from panel data,. Economics of Education Review. 2013 34(C), p. 69-75.

Chapter 1

Introduction

Quality education and gender equality are two of the 17 United Nations (UN) Sustainable Development Goals (SDGs), which serve as directions and ambitions for the future development of the global community. For education, the goal is to: "Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all." Regarding gender equality, the aim is to: "Achieve gender equality and empower all women and girls." (UN SDGs, 2022). Despite these goals, gender differences pertain even in the most advanced economies. In 2020 the gender pay gap in the EU was 13%. As this measure is an unadjusted pay gap it also reflects differences in the average type of employment for men and women (European Commission, 2022). Indeed, women are under-represented in Science, Technology, Engineering, and Math (STEM) field workforce — for example, in the U.S. women make up only 34% of the STEM workforce[1] and these jobs are among the best paid. Under-representation in STEM fields is painfully apparent in tertiary education — according to OECD, 2017 women across OECD account for 37% of new entrants to science programmes; only 24% in engineering, manufacturing, and construction programmes; less than 20% in computer science. However, the career paths of girls and boys start to diverge already by the age of 15 (OECD, 2017). A possible cause for this divergence could be a lower achievement of girls in mathematics. The latest edition of Trends in International Mathematics and Science Study (TIMSS) documents that this phenomenon starts even sooner and can be observed already in 4^{th} grade: Although not by a large margin, boys outperform girls when the international average is considered. When looking at

[1]National Science Board, National Science Foundation. 2022. Science and Engineering Indicators 2022: The State of U.S. Science and Engineering. NSB-2022-1. Alexandria, VA. Available at https://ncses.nsf.gov/pubs/nsb20221

the countries individually, of the 58 participating countries, only in 4 countries girls outperformed boys in math tests. Conversely, boys scored higher on the standardized test in 27 countries (Mullis *et al.* 2020).

This thesis aims to shed light on the reasons behind this gap in math achievement among children. So far, the literature offered several possible explanations for this phenomenon. First, the influence of society: the gap in math achievement is affected by general gender inequality in society, the hypothesis being that less equal societies can pass on negative stereotypes and hence discourage girls from focusing on math (Guiso *et al.* 2008; Fryer & Levitt 2010; Anghel *et al.* 2020) or by general societal inequality (Breda *et al.* 2018) — authors hypothesize that women have a lower status compared to men and hence more unequal societies mean status differences between boys and girls which in turn affects performance. Second, several authors studied the effects of family background (Dossi *et al.* 2021; Kim & Law 2012). Specifically, the authors show that negative stereotypes can easily transmit through families, parents can direct girls away from STEM fields, or invest less in their education. Third, the influence of the school environment: for example, the effect of sorting and tracking (Bedard & Cho 2010) or single-sex schooling (Kim & Law 2012) — differences in educational attainment between boys and girls are larger for countries where streaming starts sooner.

The next strand of literature focuses on the role of the teacher. This is also the focus of this thesis. Specifically, the objective of this thesis is to evaluate the effect of the student-teacher gender match on the results of standardized math tests. There are several possible mechanisms through which the gender of a teacher can affect students. First, a role-model effect — students can relate more easily to a teacher of the same gender and be motivated by their example (Cho 2012). Second, as in a family or a society, negative stereotypes can also be passed on by teachers (Cho 2012). Third, the self-fulfilling expectations — the teacher can more effectively communicate higher ambition to a student which then becomes self-fulfilling (Cho 2012). Next, Andersen & Reimer (2019) suggest that male and female teachers use different class organization and the study environment has different effects on boys and girls. Finally, Parker-Price & Claxton (1996) points to a difference in the learning experience of boys and girls — e.g. boys are more likely visual learners and male teachers may better relate to that experience as they themselves are likely visual learners.

Although the effect of being taught by a teacher of matching gender was studied in the existing literature, the evidence is still inconclusive. While some

authors find a positive effect of the student-teacher gender match (Dee 2007; Andersen & Reimer 2019; Carrell *et al.* 2010) others find no significant effect (Krieg 2005; Cho 2012; Winters *et al.* 2013). Some authors even find a detrimental effect of same-sex teachers (Antecol *et al.* 2015; Beilock *et al.* 2010). These studies vary in the geographic coverage, grade of the analysed pupils, and used methodology.

In this thesis, to provide more robust evidence, we analyse a large sample of countries using the newest data from a large international survey TIMSS 2019. We study the effect on 4^{th} grade students in 36 countries. Since the selection of students to teachers may be non-random (Cho 2012; Dee 2007) a more complex methodology than a simple Ordinary Least Squares regression (OLS) has to be used to reliably estimate the effect. We use an approach to overcome the issue of possible selection bias that was not yet used in this context: Propensity Score Matching (PSM). We are able to use this methodology thanks to the rich design of TIMSS which also collects exhaustive background information and, mainly, information on students before they start school. So far, the effect for 4^{th} grade students was not studied in international settings.

The analysis is run both on pooled data and within countries. The results from the pooled analysis suggest that there is no significant effect of the teacher's gender on the math results of 4^{th} grade students for girls — in line with Krieg (2005), Cho (2012), and Winters *et al.* (2013). Having a same-sex teacher seems to be detrimental for boys which is a result that is not found by other authors. When considering the results for individual countries, we find a significant effect in 4 countries for boys. In Montenegro, the effect is positive while in South Korea, Poland, and South Africa the effect is negative. As for the girls, we find a positive effect of same-sex teachers in 4 countries (Malta, Singapore, South Africa, and Macedonia) and a negative effect in 3 countries (Germany, Kuwait, and the United Arab Emirates). Existing literature reports positive effects of gender match in the U.S. (Dee 2007), in Denmark (Andersen & Reimer 2019), for boys in Spain but also for girls in France and Greece (Cho 2012), and in Western Europe (Hermann & Diallo 2017). Negative effects were, to our knowledge, so far only reported in the U.S. (Beilock *et al.* 2010; Antecol *et al.* 2015). Still, in the majority of the countries in our sample, there seems to be no effect of the student-teacher gender match corresponding to the results of Cho (2012) who also analyzes TIMSS data but for 8^{th} grade students using different methodology. To sum up, we investigated a possible explanation of the gender gap in math scores but, based on our results, gender interaction

between the student and the teacher is likely not the only reason behind the gaps.

The contribution of this thesis to the existing literature is multi-fold. First, we bring forward further evidence from international data on the effect of being taught by a teacher of a matching gender on achievements in math. Second, the employment of Propensity Score Matching is new in this context and it allows us to study the effect on data where it is not possible with methods used in literature so far (international TIMSS data for 4^{th} grade students). The method seems to be suitable for this kind of data as they offer extensive information on students and their background so that researchers are not too constrained in the choice of matching variables and can follow economic theory. In our analysis, we were able to find good counterfactuals for the treated individuals and hence synthetically approximate experimental settings. On the other hand, the treatment and control groups are relatively balanced, so the OLS method may also provide reliable estimates. Third, we contribute to the general debate on gaps in educational achievement between boys and girls, we show that teachers' gender is likely not the cause of this phenomenon universally across countries.

The remainder of this thesis is organized as follows: Chapter 2 reviews in more detail the existing literature on gender gaps in education and the effects of teacher gender, Chapter 3 presents the source of the data and discusses the variable and sample selection. In Chapter 4 we describe the methodology used for the analysis. In Chapter 5 the results are reported and discussed and in Chapter 6 we run some robustness checks. Lastly, Chapter 7 summarizes our findings and concludes the thesis.

Chapter 2

Literature Review

The educational outcomes of children are naturally a great concern for the parents, but also educators, and, eventually, since these outcomes shape the future of each country, to politicians. For policymakers in the area of education, the stakes are extremely high (generations of children), and hence evidence-based policy is needed. A plethora of papers has been written on the topic of the economics of education. In this literature review, we shall summarize papers that tried to explain the differences in educational outcomes (especially mathematics) between boys and girls — gender gaps.

The studies can be classified into two main groups: national (regional) studies and cross-country studies. The most frequent results found in national (regional) studies suggest girls score lower than boys in standardized tests in math (Dee (2007) — National Education Longitudinal Study of 1988 (NELS88), nationally representative study in the US; Fryer & Levitt (2010) — Early Childhood Longitudinal Study Kindergarten Cohort (ECLS—K) also administered in the US; Dossi *et al.* (2021) — National Longitudinal Survey of Youth (NLSY79); Carrell *et al.* (2010) — US Air Force Academy data; Andersen & Reimer (2019) — a dataset on Denmark combining survey and register data). At the same time, girls score higher than boys in reading/language skills (Dee 2007; Fryer & Levitt 2010; Andersen & Reimer 2019). However, these results are not universal. For example, Winters *et al.* (2013) did not find differences in average scores of girls and boys in both math and reading using data from Florida — Florida Education Data Warehouse (FLDOE). Antecol *et al.* (2015) found that girls score higher than boys in both math and reading using US data focusing on the disadvantaged part of the population — Mathematica Policy Research, Incorporated (MPR) and National Evaluation

of Teach for America (NETFA).

Similarly, when considering the cross-country studies a lower achievement of girls in math is the most frequent result (Guiso *et al.* (2008) — Programme for International Student Assessment (PISA) 2003; Kim & Law (2012) — PISA 2006 (only Hong Kong and South Korea); Anghel *et al.* (2020) — PISA 2003, 2006, 2009, 2012, 2015; Cho (2012) — TIMSS 1995, 1999, 2003, 2007), as well as higher achievement of girls in reading (Guiso *et al.* (2008)). Interestingly, Breda *et al.* (2018) found no gap in math results using PISA data when comparing the average results, but a lower number of girls among top performers. A similar result was found by Bedard & Cho (2010) in cross-country settings (TIMSS 1995, 1999, 2003), and Pope & Sydnor (2010) for the US using data from the National Assessment of Educational Progress (NAEP). Finally, Meinck & Brese (2019) use data from TIMSS to show the trends and development of the gender gap in standardized test performance in the last 20 years and it is suggested that girls are catching up in terms of math performance.

The literature offers several explanations as to why this gap exists. Guiso *et al.* (2008) relate general gender equality in society to the gender gaps in math scores in an attempt to show that it is not biological differences that drive the gap. Exploiting cross-country variations using PISA 2003 data they find that a higher level of societal gender equality should lead to lower gaps in math achievement. Fryer & Levitt (2010) revisit the explanation offered by Guiso *et al.* (2008) using an additional source of data: TIMSS in addition to PISA which enables them to analyse a slightly different set of countries. It is shown that once Muslim countries are included in the pool of countries the relationship between societal gender equality and a low gender gap in math scores disappears. Anghel *et al.* (2020) offer the most recent evidence on this topic again using only PISA data using waves between the years 2003 to 2015 to account for country fixed effects. It is shown that the link between societal gender equality and the gender math gap holds only for the poorer countries (bottom quartile of GDP), i.e., in developed countries there seems to be no link between gender equality in society and the gender gap in standardized test scores. A slightly different explanation is offered by Breda *et al.* (2018) who relay the gender gap in math achievement to general societal inequality rather than societal gender inequality.

The next strand of literature focuses on the influence of family and societal background on the results in standardized test scores as a potential way to explain the gender gaps in these tests. Studying the US data Pope & Sydnor

(2010) use the measure of attitude toward gender stereotypes to show that in states with stereotypical attitudes the gender gaps in math are larger. Such attitudes usually transmit easily in families, as shown by Dossi *et al.* (2021) who use fertility stopping rules to identify boy-biased families and show that girls' achievement is hampered when growing up in such a family. The specific mechanisms of how gender stereotypes can affect girls' achievement are the following: less investment in girls by boy-biased families, directing girls away from STEM fields, and transmission of the stereotypical gender attitudes to children (girls do not try themselves because they do not believe to have the necessary ability or that they should try to invest in STEM field learning) (Dossi *et al.* 2021). On the other hand, when considering data from outside the US Kim & Law (2012) found no detrimental effect of family background on girls in South Korea and Hong Kong.

Other authors try to relate the gender gap in math achievement to the school environment. Bedard & Cho (2010) explores the effects of sorting and shows that if girls are disproportionately placed in better classes they can close the gap in math achievement. Moreover, the authors show an important effect of tracking/streaming on the achievement gap — in countries with a higher degree of academic streaming i.e., in countries where streaming starts earlier, the educational gap between boys and girls is higher. Kim & Law (2012) consider the effect of single-sex schooling and find that while boys benefit from single-sex schooling, the effect for girls varies across countries in scope.

The next explanations focus on the role of the teacher. Specifically, the teacher's gender and its effects on the educational outcomes of boys and girls. This explanation is of interest to our work so we will provide a bit more exhaustive review of the literature on this topic.

2.1 Effects of teacher's gender

First, a brief background on the specific mechanisms through which the gender of a teacher can affect students. Cho (2012) mentions the following three mechanisms: role-model effect — if someone who looks like me can do it, I can do it as well; negative stereotypes — like in the case of stereotypes transmitted in society or a family, these stereotypes can also be transmitted by teachers to their students; self-fulfilling expectations (i.e., Pygmalion effect — the teacher can more effectively communicate higher ambition to a student which then becomes self-fulfilling). Antecol *et al.* (2015) and Beilock *et al.* (2010) explore

further how the attitudes are transmitted from teachers to students and find that especially teachers without a strong background in mathematics can transmit "math anxiety" to students and hamper their results. Andersen & Reimer (2019) propose a different mechanism through which teachers' gender affects students' outcomes. Specifically, they suggest that male and female teachers use a different class organization. Then, male and female students react differently to an environment created by teachers of the two genders, and hence their educational outcomes can be affected. Meece (1987) indeed documents that female teachers tend to be more supportive while male teachers are more authoritative. Similarly, Etaugh & Hughes (1975) find that male teachers create a positive environment for boys meanwhile female teachers create a positive environment for all. Einarsson & Granström (2002) also suggest that female teachers are more attentive to boys and male teachers are more attentive to girls but only as the girls get older. Next, Lavy & Megalokonomou (2019) and Terrier (2020) explore further the effects of teachers' gender favouritism by comparing blind and non-blind test scores and find a significant effect on students' performance. Lastly, a mechanism proposed by Parker-Price & Claxton (1996) points to a different learning experience for girls and boys. Specifically, it is suggested that boys are more sort of visual learners, so male teachers can relate to that experience and offer explanations suited for visual learners as they themselves can likely be visual learners. There are other possible specific mechanisms through which teachers' gender affects educational outcomes, but we shall leave further details to specialized journals. For the purposes of this work, it is enough to present this basic overview of the possible mechanism.

Having briefly outlined the specific mechanisms, let us now turn to the results of the studies on the effect of teachers' gender. So far, the results are rather mixed. Dee (2007) in the US using the NELS88 data; Andersen & Reimer (2019) for Denmark; Carrell *et al.* (2010) in US Airforce Academy data; found support for the hypothesis that students benefit from having a teacher of a matching gender to theirs. On the other hand, Antecol *et al.* (2015) using NETFA data in the US, and Beilock *et al.* (2010) using small scale survey of specific schools in the US found that having a same-sex teacher can be detrimental in terms of math achievement, especially, for girls. Krieg (2005) utilizing data from the state of Washington; Cho (2012) using TIMSS data; and Winters *et al.* (2013) using FLDOE and NAEP data, do not find any significant effect of having a same-sex teacher. The above-mentioned studies differ not only in their geographic coverage and grade attended by the analyzed

pupils. They also use different methods to identify the effect of teachers' gender on students' performance.

Generally, the assignment of the pupils to teachers or teachers to pupils may be non-random. Hence simple OLS is prone to produce biased estimates. There are three main methods that authors use to overcome this issue and identify the effect of teachers' gender on the educational outcomes of boys and girls: differences within students across subjects, differences within students over time, and random assignment. In what follows I will briefly present each method and a bit more detailed description of the studies using the method to see the reason behind mixed findings.

2.1.1 Differencing within students across subjects

As mentioned above, simple OLS may produce biased estimates due to non-random selection of pupils to teachers. Another potential problem of the OLS approach is the omitted variable bias. Although educational surveys usually include extensive information on pupils, teachers, and schools, some unobserved variables correlated with observed ones may be a valid predictor of the outcome and their non-inclusion may bias the estimates of interest. Therefore, the authors implemented a smart solution to this problem. When outcomes from more than one subject are observed, it is possible to difference out the unobserved student fixed effects. The key assumption in this approach is that unobserved student heterogeneities do not vary across subjects. As for the data requirements, it is necessary to observe outcomes for a student from at least two subjects. Also, for the purpose of estimating the effect of teachers' gender, there needs to be a different teacher for each subject.

The papers utilizing the within student across subject differencing approach are mostly focused on 8^{th} and 9^{th} graders. A positive effect of a same-sex teacher is found in Dee (2007) — NELS88 data USA; Andersen & Reimer (2019) — Denmark, mixed results are found in Hermann & Diallo (2017) — TIMSS (2003, 2007, 2011), and no effect of a same-sex teacher is found in Cho (2012) — TIMSS 1995 to 2007.

A more detailed description follows: A seminal paper by Dee (2007) is one of the first to utilize the first difference within students across subjects approach to evaluate the effect. The data from the National Educational Study (NELS) from the USA are used, in total over 21 thousand 8^{th} grade students. For each student, scores in math or science and English or history tests are observed,

such data allows the author to use the first differences as follows: assume that the unobserved heterogeneity is the same for math and English, math and history, science and English, or science and history. Since one of the four pairs of outcomes is observed for each student it is possible to difference out the unobserved student heterogeneity, assuming it remains unchanged across subjects. The results suggest that both boys and girls benefit from being assigned to a same-sex teacher, yet these results vary quite significantly across subjects (suggested negative effect in math for girls, otherwise positive) and they also differ between boys and girls. A specification check indeed suggests that girls with a propensity for lower achievement are assigned to female teachers in math. This observation supports the use of the first differencing approach to estimate the effect of a teacher's gender on students' test scores. The rest of the paper attempts to evaluate the effect of gender match on teacher perceptions of student performance and student engagement. As the assignment in math is non-random, other subjects are used to evaluate this question. For both teacher perceptions and student engagement, a positive effect of student-teacher gender match is found.

Cho (2012) used a similar identification strategy as Dee (2007) but used international (TIMSS) data for 15 Organization for Economic Cooperation and Development (OECD) countries. An advantage of using the TIMSS data is that both science and math results are observed for one student (in Dee (2007) either math or science score was observed). Again, differencing the two cancels out the unobserved student effects — while in Dee (2007) the identifying assumption was that these unobserved effects are the same for science or math and English or history the identifying assumption in this study is that the unobserved effects are the same for math and science. This assumption is arguably more plausible as math and science are closer to one another than math and history/English (or science and history/English). Four waves of TIMSS are used (1995, 1999, 2003, 2007), the focus is on 8^{th} graders and in total there are over 200 thousand students. The results are far from universal, but in most countries, there is no support for the hypothesis that same-sex teacher improves the educational outcome of a student. Moreover, for the few countries where a positive effect was found, a robustness check was done which suggests that this effect is driven mainly by teachers' quality (especially for girls). The divergence from the results of Dee (2007) is striking and may lie either in different data sources, geographic origin of the data, different time periods, or slightly different identification strategies.

Interestingly, Hermann & Diallo (2017) also use international TIMSS data for 20 European countries and find a positive effect of teachers of the same gender in Western Europe. They also find that female teachers tend to increase test scores for both girls and boys, but this effect is stronger for girls. Although the identification strategy is exactly the same as in Cho (2012) the results are quite different. Possible explanations are observing a slightly different pool of countries (but the overlap is quite large) and using more recent TIMSS waves (2003, 2007, 2011). Additionally, it is found that the effect is stronger for girls, low achievers, and students with a low socio-economic background or immigrant background (especially in Western Europe). The authors suggest that the reason for differences in teacher effectiveness may lie in the selection to the profession, specifically, it is suggested that female teachers are more effective in countries where relative teacher wage difference is more in favour of women.

In a similar manner, Andersen & Reimer (2019) combine survey and register data on 7700 Danish students to evaluate student-teacher gender interactions. The data on educational outcomes are the results of standardized school-leaving exams at the end of the 9^{th} grade in math and Danish. These data are linked to a survey on teacher practices from a Danish national survey. The authors argue that in the Danish educational system the selection to classes is virtually random due to the institutional settings, which would overcome the selection problem. Otherwise, the author uses the first differencing approach like Dee (2007) to account for unobserved heterogeneity. The results indicate that students benefit from being assigned to a same-sex teacher, this effect is larger for girls. The authors also include variables to account for teacher class management strategy and show that for boys it is this class organization rather than the sole gender of the teacher that matters. The role of gender remains significant for girls.

2.1.2 Differencing within students over time

The differencing within students over time is based on the idea that students' unobserved characteristics are accounted for by including their initial ability. While the previous method assumed that unobserved student effects remain fixed across subjects, now it is assumed that they remain fixed over time. So all changes in students' study outcomes are determined by the study process, which may be afflicted, among others, by the teacher. The main challenge of

this method is the data collection process, it is necessary to have panel data and each student has to be observed at multiple points in time.

For this reason, only studies from individual states in the US apply this method. Krieg (2005) uses a large dataset of almost 50 thousand 3rd graders in Washington who were followed for two years. This data structure allows the author to control for test scores in the 3rd grade when evaluating the effect of teacher gender on the results at the end of the 5th grade. The study has several results, notably, students of male teachers perform worse than students of female teachers. This impact is similar for boys and girls. The author also explores the possibility that caring parents may select a teacher of a specific gender for their children which would suggest that the teacher gender variable is affected by unobserved parents' situation. The author re-runs the analysis with a subset of schools where parents are unable to choose the specific gender of the teacher, but results remain unchanged suggesting little such selection.

Winters *et al.* (2013) use data from public schools in Florida over a five-year period which contains in total 1.7 million students between 3rd and 10th grade and around 13 thousand teachers from 3 thousand schools. The focus of the study is performance in maths and reading. For neither a significant gender gap is found unlike in nationally representative results at that time from the National Assessment of Educational Process (NAEP). To account for unobserved student heterogeneity, an approach that accounts for prior student proficiency is used, similar to Krieg (2005). In other words, the authors control for student fixed effects. Winters *et al.* (2013) find that students benefit from being assigned to a female teacher. This effect is larger for girls and it starts to be significant only once students enter middle/high school (6th to 10th grade).

Similarly, Hwang & Fitzpatrick (2021) use data between 2010 and 2017 in Indiana for over 760 thousand pupils between 3rd and 8th grade. Like in Winters *et al.* (2013), the student fixed effects are controlled for and the authors also find a positive effect of having a female teacher. This effect is again larger for girls and especially in maths. Unlike in Winters *et al.* (2013), these effects are found to be significant already in elementary school. There is also no evidence that boys would benefit from being assigned to a male teacher.

2.1.3 Random assignment

The last method used to identify a causal effect of teacher gender on educational outcomes is the random assignment of pupils to teachers or a randomized ex-

periment framework. With random assignment or random experiment, authors do not need to use more sophisticated methods to identify causal effects, but such data are usually quite difficult to find. One example of such data is presented by Carrell *et al.* (2010) who use data from U.S Air Force Academy (an undergraduate institution for higher education) where students are randomly assigned to professors. These data cover over 9 thousand students between graduating classes of 2001 and 2008. The results suggest a positive effect of student-teacher gender matching on performance — better in-class grades. This effect is increasing with female initial maths skills, so female students who are better at math are more sensitive to the effect of teacher gender. The authors also consider long-term effects like performance in follow-up courses or graduating with a STEM degree. This time the effect is significant especially for girls with higher initial math skills. The authors also evaluate the mechanism of the effect and find that the effect is not exclusively driven by overall teacher quality (value-added), so there exists an effect attributable to gender.

Antecol *et al.* (2015) use data from a randomized experiment carried out to evaluate the effectiveness of the Teach for America (TFA) program. Students were randomly assigned to TFA or "normal" (control classes). This random assignment allowed for the evaluation of the effectiveness of TFA, but it also allows the authors to evaluate the gender interactions. More than 1,900 primary school students are included in the study (the sample is focused on a deprived part of the population and therefore is not nationally representative). Surprisingly, the authors find that having a female teacher is detrimental to girls' math achievement.

2.2 Summary

To sum up, there are various mixed results in the literature on gender gaps and their explanations. While it seemed that the gap between boys and girls in math achievement was closed (Breda *et al.* 2018 — PISA 2003 to 2015) or was closing over time (Meinck & Brese 2019 — TIMSS 1995 to 2015), this gap opened again massively in the latest edition of TIMSS 2019 (Mullis *et al.* 2020). Hence it is crucial to understand the reason behind this gap. The effects of teachers' gender seem to be a promising avenue but so far the evidence is inconclusive. This thesis should add to the existing literature by exploring a new methodological framework, which would allow us to study the effects on 4^{th} graders (the gap opens as early as 3^{rd} grade, according to Fryer & Levitt

2010) in cross-country settings. To the best of our knowledge, the effect of teachers' gender on fourth graders has so far been only studied using national datasets.

Chapter 3

Data

The purpose of this chapter is to present the datasets used for the analysis to the reader. In the first section of this chapter, we provide a brief background on the source of the data. In the next section, the data selection process is described alongside descriptive statistics of the dependent variable. We also present chosen independent variables.

3.1 Trends in International Mathematics and Science Study (TIMSS)

In this thesis, we aim to compare results across countries hence Trends in International Mathematics and Science Study (TIMSS) data are used. TIMSS is an international comparative study carried out by the International Association for the Evaluation of Educational Achievement (IEA). TIMSS assessment evaluates the achievements of 4^{th} graders and 8^{th} graders in mathematics and science by administering standardized tests. It started in 1995 and it is conducted every four years. For the purpose of this work, we use the last wave of TIMSS which was administered in 2019. In total 64 countries participated in TIMSS 2019 (Mullis *et al.* 2020).

Working with TIMSS data has several advantages. Firstly, it allows for both national level analyses and cross-country comparisons, these may be important to exploit structural relationships that are not visible using solely national datasets (Cordero *et al.* 2017). Secondly, apart from student characteristics TIMSS also collects exhaustive information about the classroom, school, and national contexts. Moreover, information about the children's family background and home is collected which we intend to rely on in our work. Thirdly,

a rigorous sampling strategy ensures that results can be generalized for the whole population. Fourthly, the advantage of using the latest data probably does not need to be further elaborated on, but the long history of administering TIMSS ensures that both the tests and sampling strategies are top quality since they are continuously updated and re-evaluated throughout time (Mullis *et al.* 2020). Lastly, the data are freely available on the TIMSS website.[1]

There are six questionnaires used to extract contextual information on 4^{th} grade students participating in TIMSS. Student Questionnaires are filled out by students themselves and they are used to obtain basic demographic information, information about students' home environment and school climate for learning, as well as students' own attitudes towards learning. In 2019 TIMSS also initiated the possibility of a computer based version of the test, so for those who participated in this manner there were additionally several questions about their experience with digital devices. The third questionnaire is the Home Questionnaire (or Early Learning Survey) which asks the parents or guardians of participating children about attitudes towards their children's education, home resources, and level of support in learning. Moreover, parents are asked about their own education and employment situation. Crucially for us, the parents are also asked about literacy and numeracy activities before their children attended school and when starting school. Next, teachers fill in the Teacher Questionnaires which are used to obtain information on teachers' education, experience, and professional development. In addition, teachers provide their views on the students they teach, topics covered in the curriculum, classroom amenities, and, for example, how much time they spend preparing their lessons. The principals of participating schools provide school specific contexts via School Questionnaires. Lastly, TIMSS National Research Coordinators fill in Curriculum Questionnaires that provide insights into countries' education systems and national curriculum policies (TIMSS 2019 Context Questionnaires).

3.2 Data selection

3.2.1 Dependent variable and variable of interest

The dependent variable used in our analysis is the score in mathematics of 4^{th} grade students. We are mainly concerned with how this score is affected by

[1] https://TIMSS2019.org/reports/download-center/

teachers' gender and whether a student-teacher gender match increases achievement. According to TIMSS[2], of the 58 countries where 4^{th} grade students were tested, in 27 there was no gap between male and female students in math achievement, in 27 countries the gap in math favoured boys, and in 4 countries the gap favoured girls.

The key question that this thesis is attempting to answer is what is the effect of teachers' gender on the performance of students. The methodology we use is Propensity Score Matching (PSM). The idea of this method is to match together students of male and female teachers with similar characteristics (same propensity to be treated) and then identify the effect of treatment (having a same-sex teacher) within each pair. For PSM to work it is necessary that the common support condition is satisfied which basically means we need to observe enough students that are taught by a male teacher and enough students that are taught by a female teacher. As some countries do not have enough of these observations the analysis is limited to a subgroup of countries.

Let us now present a more detailed description of how countries for our sample were selected. We are purposely selecting only countries with a relatively balanced distribution of teachers so that the results can be interpreted as relevant for these countries. A preliminary inspection of the data showed that male teachers are more scarce than female teachers. In some countries the scarcity is so great that it prohibits any reasonable analysis — the small number of observations hinders reliable inference, moreover, obtained results would be hardly generalizable for the population, hence they would not be very representative. Specifically, in these countries, the percentage of students taught by a male teacher is lower than 5%: Azerbaijan, Armenia, Bulgaria, Croatia, Georgia, Hungary, Italy, Kazakhstan, Latvia, Lithuania, Oman, and Russia. Therefore these countries are not included in our analyses. Additionally, in Saudi Arabia girls are almost exclusively taught by female teachers while boys are mainly taught by male teachers. This makes Saudi Arabia not suitable for our analysis. Moreover, as stated above, we intend to rely largely on data from the Early Learning Survey in our strategy. In the following countries, the Early Learning Survey was not administered so they were also excluded from our analysis: Australia, England, Netherlands, Northern Ireland, and the USA. Next, Japan is lacking information on pre-primary education which we intend to use for matching and hence is also discarded from our analysis. Similarly, Norway is excluded because of missing information on parents' occupation.

[2]https://TIMSS2019.org/reports/

Then, in Austria and Belgium, we are missing information on teachers' math education which is not used for matching but serves as a crucial control in the second stage of the alternative model. Lastly, we eliminated observations where the student's or teacher's gender was unknown. The summary of the discarded countries is in Table 3.1.

Table 3.1: Countries discarded from the analysis

Country	Reason to be discarded
Azerbaijan Armenia Bulgaria Croatia Georgia Hungary Italy Kazakhstan Latvia Lithuania Oman Russia	Less than 5% of students taught by a male teacher
Saudi Arabia	Almost perfect student-teacher gender match for all observations
Australia England Netherlands Northern Ireland USA	Early Learning Survey not administered
Japan	Missing data on pre-primary education
Norway	Missing data on parents' occupation
Austria Belgium	Missing data on teachers' math education

Hence, 36 countries remain for the analysis. In Figure 3.1 we can see the difference between girls' results in mathematics and boys' results in mathematics (gender gap) for the subset of countries we work with. If the gap is positive (to the right of the zero line) it means that girls scored, on average, more on the standardized test than boys. The 95% confidence intervals for the estimates are also displayed in Figure 3.1. We see that in most countries the

gap favours boys, in 8 countries there is no significant gap, and in 5 countries the gap favours girls.

Figure 3.1 also includes the percentage of students taught by male teachers. At first glance there is no clear trend, maybe a weak relationship suggesting that the larger the number of students taught by a male teacher the more the gap favours girls (lower gap or gap in their favour). To explore the effects of teachers' gender on student achievement further a summary table in Appendix A is constructed. It shows how many students in each country sample are male/female and how many are taught by male/female teachers. Average achievement for each student-teacher gender combination is also reported. We divide the sample into two subsamples — one for girls and one for boys. This is necessary as we plan to test the hypothesis separately for boys and girls. Nevertheless, combining the results into one table should be easier for a reader to grasp and immediately see the differences. The girls' dataset has in total 144,998 observations and the boys' dataset has 149,825 observations.[3] Despite preliminary selection, we still see that in some countries the number of boys/girls taught by a male teacher is relatively low. This increases the risk that the matching procedure will not work well — specifically the common support condition (enough treated: taught by same-sex teacher; enough untreated: taught by opposite sex teacher) may be violated.

The purpose of including a summary table in Appendix A is to see the effect of the matching gender on student achievement. For example, when we look at the United Arab Emirates we can see that girls taught by male teachers score on average higher than girls taught by female teachers. Such observation would be in line with the findings of Antecol *et al.* (2015) and Beilock *et al.* (2010) that female teachers hamper girls' math achievement. On the other hand, Antecol *et al.* (2015) suggest no effect of teachers' gender on boys, but if we look at the data for United Arab Emirates in Appendix A, we see that boys with female teachers score, on average, higher than boys with male teachers. Such effect is not universal across countries and so this example serves as another motivation to study the effect in more detail.

Figure 3.2 is a graphical representation of gender matching explained above. It shows a gap in the achievements of treated and control students — treatment means having a same-sex teacher i.e., a male teacher for boys and a female teacher for girls. Again, when estimates are positive (to the right of the zero

[3]These numbers include incomplete cases which will not be used for the main analysis but will be used as part of the robustness check.

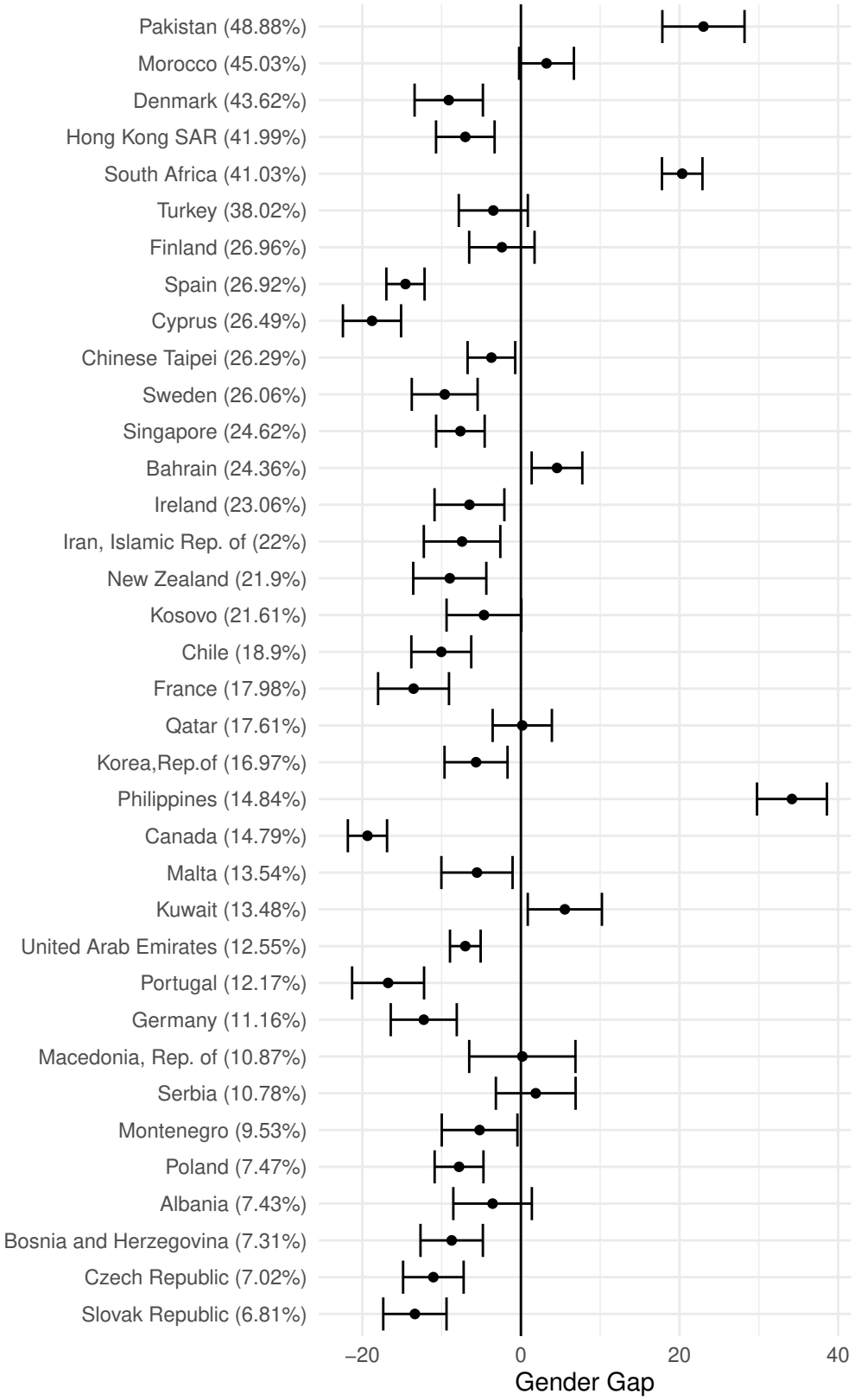

Figure 3.1: Gender gap in mathematics

Source: Author's calculations based on TIMSS 2019 data.
Note: When the estimate is positive girls score higher in math than boys. Estimates are shown with their 95% confidence intervals. The percentage of male teachers is reported for each country.

line) it suggests that students benefited from having a same-sex teacher. Figure 3.2 suggests that in 23 countries girls benefit from having a same-sex teacher, while boys benefit from having a same-sex teacher only in 14 countries. For a better understanding let us interpret the results, for example, for the Czech Republic. It is suggested that Czech girls perform equally well regardless of the gender of their teacher. Czech boys on the other hand seem to perform better when having a male teacher. Canada may serve as another, more representative, example: While girls benefit from being assigned to a same-sex teacher, boys' achievement seems to be hampered when having a same-sex teacher. These observations can serve to provide an overview and further motivation to explore the effect, however, they cannot be yet interpreted as causal. Let us now turn to the selection of independent variables.

3.2.2 Control variables

Regarding the choice of matching variables, as mentioned above, TIMSS data offer a plethora to choose from. Bearing in mind that the intended estimation strategy is Propensity Score Matching (PSM) the variables should describe the information available to those deciding about the selection to treatment at the time of selection to treatment — for now, we will assume that whether a student will be treated (will have a same-sex teacher) is decided at the start of the first school year. Although the researcher does not directly observe the decision making process, it is reasonable to assume that the Early Learning Survey well approximates the information available to those deciding on treatment selection at the time of the selection. At the same time, these variables should be affecting the outcome (Cordero *et al.* 2017). Cho (2012) suggested that sorting to treatment (sorting of students to teachers of a specific gender) may be non-random, for example, worse students can be assigned to female teachers. We try to address this issue by utilizing as much information about students prior to the assignment as possible.

For both girls' and boys' datasets, we extract information on both the students' early learning achievement as well as their family background. Moreover, in the alternative models, we use the teacher background and school background to complement the analysis in the second stage (not used for matching). The following list shows the specific variables used for matching and independent variables used for the second stage of the alternative model:

- Early Numeracy Activities Before School (ASBHENA) derived from Home

Figure 3.2: Gap in math achievement between students in treatment and control group

Source: Author's calculations based on TIMSS 2019 data.
Note: When the estimate is positive students in the treatment group score higher in math than students in the control group. Treatment — having a same-sex teacher. The percentage of male teachers is reported for each country.

Questionnaire completed by parents. This variable sums up information from a series of questions that ask about pre-school activities. For example: "Before your child began primary/elementary school, how often did you or someone else in your home do the following activities with him or her? Count different things" The answers are mapped into a scale where a higher number means better skills. (T19 UG Supplement1, T19 UG Supplement3)

- Early Numeracy Tasks Beginning School (ASBHENT) derived from Home Questionnaire completed by parents. This variable intends to quantify information at the start of the school by compounding several questions like: "Could your child do the following when he/she began the <first grade> of primary/elementary school? Count by himself/herself" This variable is also mapped to a scale. (T19 UG Supplement1, T19 UG Supplement3)

- Student Attended Preprimary Education (ASDHAPS) is also derived from Home Questionnaire but intends to provide us with details about students' pre-school education (kindergarten etc.). (T19 UG Supplement1, T19 UG Supplement3)

- Parents' Highest Education Level (ASDHEDUP) is self-explanatory and has six levels from primary or no school to "University or Higher". (T19 UG Supplement3)

- Parents' Occupation (ASDHOCCP) has six levels between "Never Worked for Pay" to "Professional" (T19 UG Supplement1, T19 UG Supplement3)

- Students' Age Starting School (ASBH05) "How old was your child when he/she began the <first grade> of primary/elementary school?" (T19 UG Supplement1)

- *School Composition by Socioeconomic Background (ACDGSBC) from school context data with three levels affluent — disadvantaged.*[4] (T19 UG Supplement3)

[4]This and the following variables are not used for matching but are used in the second stage of the alternative procedure as additional controls and to control for potential differences between male and female teachers.

- *Teachers Majored in Education and Mathematics (ATDMMEM) is derived from Teacher Questionnaire and describes teachers' educational background.* (T19 UG Supplement3)
- *Teacher Age (ATBG03) also serves to control for teachers' characteristics.* (T19 UG Supplement1)

The nature of the first three variables ensures that the information was available at the time of selection to the treatment. Students' age was also known at the time. Parents' education and occupation can change throughout the four years, but as argued by Hogrebe & Strietholt (2016) these are "rather stable measures". The summary statistics for these measures can be found in Appendix B, again the results from the boys' dataset and girls' dataset are combined into one table.

Chapter 4

Methodology

In this section, the empirical approach of this thesis will be presented. The key question we are trying to answer in this work is whether student-teacher gender interactions affect the educational outcomes of 4^{th} grade students. As mentioned in the literature review, the standard in estimating this effect on TIMSS data is to use differencing within student across subjects to "difference away" the unobserved fixed student effects. However, such a strategy is generally available only for higher grade students who are taught by different teachers in different subjects (in the case of TIMSS 8^{th} grade). Since 4^{th} grade students are usually taught by one teacher, such a strategy is not applicable. Hence, we propose to use the Propensity Score Matching (PSM) which would allow us to evaluate the effect on 4^{th} graders.

The idea behind this approach is to simulate a randomized experiment using observational data (Cordero *et al.* 2017). The treatment in our case is having a same-sex teacher. Ideally, we would compare the outcomes of one individual once with treatment and once without it. Yet, this is not possible in practice, since we only observe one outcome for one individual i.e., the individual cannot be treated and not treated at the same time. Comparing mean outcomes of treated and untreated (henceforth control) group is possible, but the inference is generally valid only if the two groups do not differ in terms of other predictors of outcome (Hogrebe & Strietholt 2016). Such a condition usually holds for randomized experiments which, unfortunately, is not the case with our data.

Generally, with observational data, we are at risk of selection bias due to non-random selection. Usually, the treatment and control group are fundamentally different from each other, and hence simple comparison may produce a biased estimate. It is possible to solve the selection issue simply by including

variables on which the selection is done into an OLS regression model. This is a valid approach however it has some pitfalls. Specifically, if there is a big average difference between the control and the treatment group, OLS may produce a biased estimate because of extrapolation — the model is extrapolated even to the region where there is not enough data. Matching, on the other hand, explicitly requires the satisfaction of the Common Support assumption so that only similar individuals are compared and extrapolation error is avoided. Moreover, the intrinsic linearity of the OLS model can also be regarded as a weakness although it is possible to model the non-linearities even within this framework. Conversely, matching is more flexible in this regard (Schafer & Kang 2008).

Therefore, we implement PSM that aims to recreate the treatment group within the control group. Basically, we find individuals who have the same propensity score (same conditional probability of being treated given the observed covariates), then if the two individuals have the same propensity score but are in different treatment groups, the assignment can be assumed to be random (Cordero *et al.* 2017). So, for each pair, we get an estimate of the treatment effect as the difference in the outcome of the treated individual and the individual from the control group. Finally, the treatment effect is computed as the weighted average of these differences.

Under the assumption that unobservable covariates are correlated with observed covariates and individuals in the treatment and control group have the same distribution on unobservables, PSM solves the endogeneity problem (Caliendo & Kopeinig 2005). These assumptions are rather strong ones to make, their justification follows in the latter part of this chapter and in the discussion of results. PSM is usually performed in two stages, let us now describe each of the stages in further detail.

4.1 First stage

In the first stage, we estimate the propensity score — we estimate the probability of being treated (having a same-sex teacher) based on the selected matching variables. We perform the matching on two sub-samples: girls and boys. This should ease the computational intensity, but it could also provide a key insight into possible heterogeneity of the effect i.e., the effect may only be significant for boys or only for girls or have the opposite direction. Although the main goal of the first stage is to estimate the propensity score which is then used to match similar individuals from treatment and control group together, we

can also check the validity of Hypothesis #3: Selection to treatment (having a same-sex teacher) is non-random.

To estimate the propensity score we use a standard logistic model — logit. Where the dependent variable is binary and equals 1 if the individual has a same-sex teacher and 0 if not.

$$Treatment_i = f(\beta_0 + \beta_1 X_i + \beta_2 Z_i) \tag{4.1}$$

There is an individual i in whose treatment status we are interested. Vector X represents a student's i characteristics derived from the Early Learning Survey. Specifically, it includes the variables Early Numeracy Activities Before School (ASBHENA) and Early Numeracy Tasks Beginning School (ASBHENT) which are compound variables that capture the math skills of the individual i before starting school and at the time of starting school. Moreover, Student Attended Preprimary Education (ASDHAPS) and Students' Age Starting School (ASBH05) are part of the vector X. These variables capture whether and how many years the individual i spent in institutions like kindergarten and how old the student i was at the beginning of the first grade. Vector Z represents the characteristics of the parents of student i and is composed of two factor variables capturing Parents' Highest Attained Education Level (ASDHEDUP) and Parents' Occupation (ASDHOCCP).

Generally, there have been mentioned two main hypotheses in the literature as to why the sorting of children to teachers may be non-random. First, Cho (2012) suggested that it may be the case that worse students get assigned by the school to specific gender of a teacher (for example, bad students are assigned to female teachers). Second, Dee (2007) proposes that it may be possible for some parents to choose the gender of the teacher of their children.

Hence, the choice of the variables included in the first stage model accounts for both the "initial" skills and family background. The choice of variables to include is a peculiar matter, because: "...omitting important variables can seriously increase bias in resulting estimates." — Caliendo & Kopeinig (2005). At the same time, the matching variables should not be affected by participation in treatment. In our case, we choose to work with variables from the Home survey that asks about the period before the start of treatment, and the occupation and education of parents are, as argued by Hogrebe & Strietholt (2016), relatively stable.

Another consideration when choosing variables to include in the model concerns two crucial assumptions needed for performing PSM. The first one is the Unconfoundedness or Conditional Independence Assumption (CIA) which states that potential outcomes are independent of treatment assignment conditional on the balancing (propensity) score (Caliendo & Kopeinig 2005). Basically, we do not want the treatment decision to be influenced by variables other than those included in our model. It also means that the matching variables should not be affected by the treatment. The second assumption is Common Support, it rules out perfect predictability and ensures that for people with the same propensity score it is possible to participate or not in the treatment (Black & Smith 2004). In other words, it ensures we can find a good counterfactual. Unfortunately, there seems to be a certain trade-off between the plausibility of the two conditions. When trimming the covariates to a minimum, the Common Support is not a problem, but the plausibility of CIA becomes rather unlikely. When the model is too rich the CIA is likely to hold but the Common Support may be a problem (Caliendo & Kopeinig 2005). Therefore we stick to the "economic theory" model in the main specification.

4.2 Second stage

In the second stage, we should finally be able to answer the main question of this work and assess the validity of hypotheses 1 and 2.

Hypothesis #1: Having a male teacher improves the math score for boys.

Hypothesis #2: Having a female teacher improves the math score for girls.

There are several ways how to obtain the PSM estimator, and each of them uses a different matching algorithm. They differ in handling the common support condition but also in weights assigned to different observations. The first one is Nearest Neighbour (NN) matching. In this approach, the counterfactual for the treated individual is the individual from the control group with the closest propensity score. There are several important decisions to be made when using this algorithm: whether to use it with or without replacement, whether to use oversampling (two controls for one treated individual) and if so how to assign the weights. All these decisions unfortunately include a certain trade-off. Specifically, the trade-off between bias and variance (Caliendo & Kopeinig 2005).

The next matching algorithm is Caliper and Radius Matching which basically imposes a tolerance level within which the observations can be matched, but choosing a reasonable tolerance level may be difficult. Another possibility is Stratification and Interval Matching. This method divides the sample into strata and compares the results within strata — under normality, dividing the sample into just five strata should be sufficient to remove most of the bias associated with covariates. Another option is Kernel and Local Linear Matching which use the weighted averages of all observations in the not treated group to construct the counterfactual outcome for those treated. The main considerations for this method are the choice of bandwidth and kernel function, where bandwidth choice again brings forward the dilemma between bias and variance. Finally, it is also possible to perform Weighting on propensity score to balance the control and treatment groups (Caliendo & Kopeinig 2005).

There is no clear guideline as to which of the mechanisms is the best. Therefore we plan to follow the strategy proposed by Caliendo & Kopeinig (2005) which suggests trying and comparing several algorithms and unless the results vary significantly there is no need for further investigation. We will use the Nearest Neighbour (NN) algorithm in our main model.

Once the matching is done, which in the case of NN means observations with the most similar propensity score from treatment and control group are paired together (we perform an exact match on country variable meaning that only observations from the same country can be paired together). Then for each pair, we take the difference between the math score attained by the treated and control observation which gives us the treatment effect for each pair. Averaging these effects either naively or using the weights provides us with the Average Treatment Effect on the Treated (ATT). The analysis is performed in the R software, *MatchIt* package is used for matching.

As suggested by Zhao *et al.* (2021), it is also common in the literature to run regression on the covariates using the matched (balanced) samples instead of just comparing the outcomes of treated and control individuals. This method controls for any remaining disbalances after matching and thus is said to be "doubly robust". Moreover, as the regression is only done on the balanced sample it avoids the possible extrapolation error that could be attained by an OLS analysis on the original sample. Therefore, we also report the results from this method when presenting the results in the following chapter. Except for the matching variables we use teacher and school characteristics in this regression. These variables are solely connected to the outcome and not to the treatment

decision. So, we are combining matching with a regression-based model. This is sometimes called dual modelling and it was found to be less sensitive to misspecification as well as resulting in less biased estimates (Schafer & Kang 2008). Package *intsvy* is used for the regression. The advantage of *intsvy* package is that it produces standard errors using the jackknife replication technique which takes into account the peculiar design of TIMSS: the dependency of observations within one school and the imputation of the plausible values. Hence, the standard errors are produced taking into consideration both the sampling error and the imputation error (Caro & Biecek 2017).

Chapter 5

Results

In this section, we shall present the results of the Propensity Score Matching. The logic in which the results are presented is as follows: first, general pooled results are presented, and then country-specific results will be shown. The results will be presented and discussed along the way in the text. As mentioned in Chapter 4 (Methodology) the matching is done separately for boys and girls and hence the results for boys and girls vary. We try to present the results so that they are as easily graspable for the reader as possible. After presenting the main results from matching we also show an alternative approach that combines matching with a regression model.

5.1 Matching

5.1.1 Pooled results

The apparent advantage of pooling is that the number of observed students that can be used for the analysis is massively increased. This should help to increase the precision of the estimate of interest — effect of student-teacher gender match. On the other, hand this aggregate gives only limited information about the effect, especially, in terms of potential policy recommendations. Mainly, because the resulting estimate is a between country average, and so country-specific heterogeneities remain hidden. Although the analysis was done separately, we will present results for both boys and girls in this section.

As mentioned in Chapter 4 (Methodology) the matching was on the following covariates: Math skills before the start of school (ASBHENA), Math skills at the start of school (ASBHENT), Pre-primary education (ASDHAPS), Parents' education (ASDHEDUP), Parents' occupation (ASDHOCCP), Student's

age (ASBH05) when starting school. In the first stage, a logit model is used to determine the propensity score i.e. the probability of being assigned to treatment (having a same-sex teacher). Then matching is performed using 1:1 NN without replacement and 1:1 NN with replacement. The effect of treatment (having a same-sex teacher) is obtained by comparing the outcomes of treated and control individuals within the matched pair and then averaging across the pairs.

When looking at the results of the first stage for boys, the hypothesis that the selection to treatment (having a male teacher) is non-random is confirmed. Most variables used for matching significantly affect the selection to treatment at least at the 5% significance level. The only variable that seems to not have any effect on having a male teacher is the age of the student.

The results are similar when considering girls (this time the treatment is having a female teacher). Again most of the variables are significant at least at the 10% significance level. This time, except for students' age also the parents' occupation seems not to matter for the selection.

As for some specific results, which are however not the main focus of this work so they will be presented very briefly. It appears that more skilled boys and girls before the start of school are more likely to be assigned to female teachers. While boys with lower pre-primary education are more likely to have a male teacher, girls with lower pre-primary education are more likely to have a female teacher. Boys from families with higher parental education and better occupation are more likely to have a female teacher. For girls, higher parental education increases the chance of having a female teacher while better occupation increases the chance of having a male teacher, yet, the effect of parental occupation is statistically insignificant. These findings suggest that without taking selection into account we could find biased estimates of the effect of teacher's gender on student performance.

In figures 5.1 and 5.2 we can see the distribution of propensity scores for boys and girls respectively. For each gender, we distinguish the distribution of propensity scores for the control and treatment group. For both boys and girls, we see that the distributions are quite similar regardless of the treatment status. This visual check reassures us that the common support condition is not problematic when pooling the data. Still, there are some minor differences between control and treatment group distributions. For boys, the mean propensity score of the treatment group is 0.257 while for the control group it is 0.238. Girls are generally more likely to be treated but again the difference between

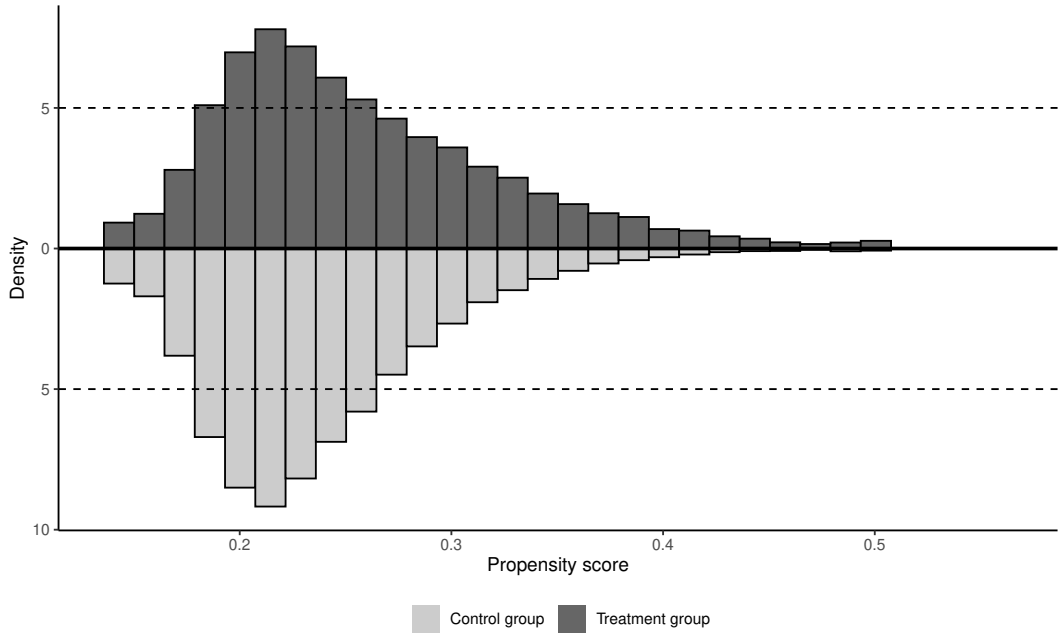

Figure 5.1: The distributions of the propensity scores for boys

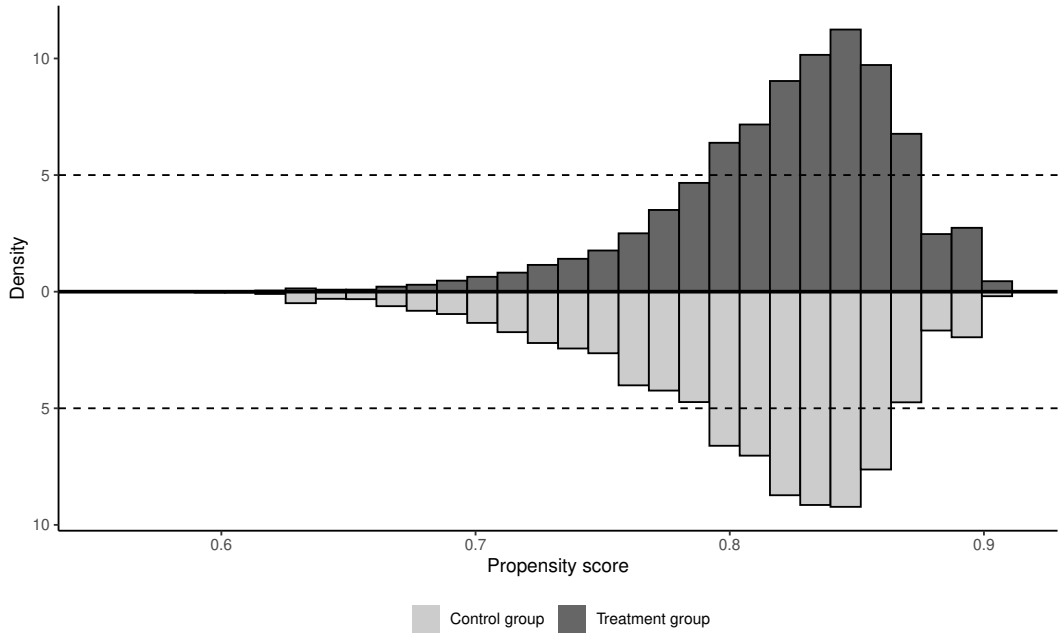

Figure 5.2: The distributions of propensity scores for girls

the control and treatment group mean propensity score is not big: 0.822 for the treated and 0.806 for the control group. This also suggests that even a simple OLS may work well with our data since the treatment and control groups are similar in terms of the propensity score. We present an OLS model as part of the robustness check in the next chapter.

As for the matching itself, as mentioned in the methodology, there are several different matching procedures. We will focus on the Nearest Neighbour matching which renders big strength since it is intuitive and easily graspable even for uninterested readers. Within the Nearest Neighbour framework, we will compare two different mechanisms: 1:1 without replacement, and 1:1 with replacement.

The advantage of using the replacement, in general, is that some relatively less informative units can be discarded in favour of some more informative ones that can be used multiple times. This should lower the overall bias, but at the same time, it reduces the sample size, and hence it may reduce the precision of the estimate. It is also possible to use oversampling (3 control units for 1 treated and weighting), then the variance should be decreased, but at the same time, the bias may increase (Caliendo & Kopeinig 2005). Hence we do not use oversampling in our main model.

Before moving on to the presentation of the results, we will briefly evaluate the matching procedure. Generally, allowing for replacement leads to better achieved balance, however, when considering the pooled data, matching worked reasonably well even without replacement. For boys, both matching algorithms managed to bring the standardized mean difference to less than 0.05 for all matching variables and for the propensity score. As mentioned above, when allowing for replacement the standardized mean difference was slightly lower. When considering girls there are more significant differences between the two mechanisms. When replacement is not allowed, the matching quality is a little worse but still, the standardized mean difference is lower than 0.1. When we allow for replacement, the standardized mean difference between control and treatment group is practically zero for the propensity score. Another effect of allowing for replacement is a significant increase in the number of units we can work with. This is especially visible for girls, as there are many more girls treated than girls in the control group. When we can use girls from the control group more times we do not have to discard treated units simply because there are not enough girls in the control group.

After evaluating the matching procedure we can now present the results of

interest. The results for both boys and girls are reported in Table 5.1. TIMSS provide five plausible values for the math score. We report the results for the first plausible value but the results are similar across the five plausible values. For the two mechanisms, we report the estimate of the effect of having a same-sex teacher along with a cluster robust standard error. Then, the total number of treatment units is reported. We also report the weighted ATT which takes into account the sampling weights provided in the TIMSS dataset. We use the weights for treated units as we are interested in the ATT. Using the weights should allow us to generalize the results to the population.

Table 5.1: The effect of having a same-sex teacher

	Boys		Girls	
Replacement	No	Yes	No	Yes
ATT	-5.03***	-5.08***	7.05***	-0.73
S.E.	1.33	1.46	1.36	0.89
Weighted ATT	-9.99**	-17.10**	9.86**	1.48
S.E.	4.76	6.83	4.81	2.80
N	20,167	20,946	15,680	70,886

Note: Estimate represents the average treatment effect on boys with a male teacher obtained by 1:1 NN matching without and with replacement. Sampling weights are used for the weighted ATT. The dependent variable is the math test score (1^{st} of the 5 plausible values). N shows the total number of matched individuals from the treatment group. Matching variables include: Math Skills before School, Math Skills Starting School, Pre-primary Education, Parents' Education, Parents' Occupation, Student's Age. Standard errors are cluster-robust. Matching is done within countries.

As for some specific results, we can see that the effect is generally negative for boys. The boys who have a male teacher score 5 points lower on a standardized test. The result is similar for both matching mechanisms. When we account for the weights, the effect is even larger in absolute terms but differs for the two mechanisms: -9.99 without replacement and -17.1 with replacement. The difference is striking and a possible explanation for it is the following: when allowing for replacement we were able to match more units (20,167 vs. 20,946) and apparently these units had large weights and hence were important in the weighted average but not when computing the mean without weights (there the estimate is similar for both matching mechanisms). Another possibility is that the matched pairs differ for the two mechanisms, this is likely to happen as we can use some more informative observations from the control group multiple times, plus we are able to disregard some less informative units from the control group instead of forcing them to match because there are no better

alternatives. Nevertheless, as already mentioned, the downside of allowing for replacement is larger standard errors.

Concerning girls, there is a significant difference between the two matching mechanisms. Mainly, because there is a larger number of treatment than control units. So, when we do not allow for replacement we have to disregard a large number of treatment units that can otherwise be matched when replacement is allowed. Specifically, the difference between the number of units used for analysis is 55,206 (15,680 vs. 70,886). This difference is reflected in the results: while the mechanism without replacement suggests a positive effect for girls of having a female teacher (7.05 not weighted, and 9.86 weighted). When replacement is allowed the effect becomes insignificant: -0.73 without weights, and 1.48 with weights. Since we use more units to estimate the effect and since the matching process performed better in terms of mean standardized difference, the estimates suggesting no effect of gender matching are more trustworthy.

To sum up, if we want to generalize the effects to the population we should consider the effects estimated using the weights. Also, since matching with replacement performed better in terms of standardized bias, it is the preferred mechanism. So, the effect of having a same-sex teacher is negative for boys: -17.10, and positive, yet statistically not different from zero for girls.

5.1.2 Country specific results

As mentioned above, the 1:1 matching with replacement performs better as a matching mechanism and so is used to determine the effect of having a same-sex teacher for boys and girls. We only report the weighted means for each country using the weights provided by TIMSS. Therefore, the results should be generalizable for each country's population. As for the standard errors, we use the cluster robust standard errors as in the pooled model. Otherwise, the procedure remains unchanged — the covariates that we are matching on are: Math skills before the start of school (ASBHENA), Math skills at the start of school (ASBHENT), Pre-primary education (ASDHAPS), Parents' education (ASDHEDUP), Parents' occupation (ASDHOCCP), and Student's age (ASBH05).

The results for boys are reported in Table 5.2, out of the 36 countries, we see a significant effect in 6. Specifically, out of the 28 countries that report a negative effect, only in Korea, Kosovo, Pakistan, Poland, and Spain is this

estimate statistically significant. On the other hand, boys in Montenegro seem to benefit from having a same-sex teacher. The remaining 7 countries report a positive, yet, statistically insignificant effect.

For all the countries the matching procedure worked well in general however in some countries there was some remaining disbalance on individual matching covariates. For example, in Kosovo and Montenegro the standardized mean difference was slightly over 0.1 for specific matching variables but the propensity score was matched almost perfectly. Hence the results should not be compromised. Moreover, to address this potential issue we run a model that accounts for the remaining disbalances in specific matching variables in the next section (Dual modelling).

Turning now to the results for girls. These are presented in Table 5.3. The variation in the results is higher than for boys. In over half of the countries the effect of having a female teacher is statistically significant. The effect is positive in 17 countries (in 10 countries this effect is significant). On the other hand, the effect is negative in 19 countries (of which in 9 this effect is significant). As for the matching procedure, like for boys, although the match on propensity score was almost perfect, there were some remaining disbalances on individual matching variables. We will run a model that accounts for this to check the robustness of our results.

To sum up, the results presented in this section suggest that there are some important differences in the effect for boys and girls. Moreover, we see that the effect varies quite substantially across the countries which demonstrates the importance of this kind of analysis in addition to the analysis of the pooled data. Only in 2 countries, there is a significant effect for both boys and girls. The achievement of students in Spain is hampered by a same-sex teacher i.e. having a male teacher is detrimental for boys while having a female teacher leads to lower math achievement among girls. This corresponds to the results shown by Antecol *et al.* (2015) and Beilock *et al.* (2010) who show that girls' achievement is hampered by female teachers however they do not find a similar effect for boys. In Montenegro, the effect of having a same-sex teacher is positive for boys but negative for girls. One possible explanation is the heterogeneity of the effect between boys and girls (for example, Hermann & Diallo 2017, Hwang & Fitzpatrick 2021, and Antecol *et al.* 2015). Another possibility is that male teachers in Montenegro are systematically better than female teachers (Krieg (2005) found the opposite on data from Washington: male teachers hamper achievement for both boys and girls), to explore this option more we run models

Table 5.2: The effects of having a male teacher for boys across countries

Country	ATT	S.E.	N
Canada	-3.27	5.14	557
Chile	-4.86	4.57	445
Chinese Taipei	-2.14	3.11	867
Cyprus	1.47	3.85	632
Czech Republic	7.65	8.26	159
Denmark	-6.07	6.00	395
Finland	-5.21	4.39	567
France	-5.16	8.24	268
Germany	-3.93	5.49	117
Hong Kong SAR	-0.58	3.87	877
Iran, Islamic Rep. of	-8.22	9.44	925
Ireland	-7.14	6.63	427
Korea,Rep.of	-17.71***	5.41	402
Kosovo	-25.04**	11.52	151
Kuwait	-9.98	28.22	426
Malta	-7.16	9.20	174
Bahrain	7.22	6.76	1,528
Montenegro	36.92**	16.48	100
Morocco	1.47	7.89	1,202
New Zealand	8.30	9.55	224
Pakistan	-31.63*	17.98	1,166
Philippines	-6.46	8.77	443
Poland	-9.73***	3.35	294
Portugal	-5.86	8.09	227
Qatar	-8.68	11.00	656
Serbia	-5.79	10.85	221
Bosnia and Herzegovina	-0.93	10.71	124
Singapore	6.89	5.02	1,220
Slovak Republic	-0.65	4.90	170
South Africa	-5.51	3.72	1,857
Spain	-8.11**	3.53	1,312
Sweden	-3.06	6.98	318
United Arab Emirates	-10.84	8.08	1,114
Turkey	-9.41	6.77	1,218
Albania	13.40	21.64	88
Macedonia, Rep. of	-2.57	20.98	75

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Note: Estimate represents the average treatment effect on boys with a male teacher obtained by 1:1 NN matching with replacement. Sampling weights are used. The dependent variable is the math test score (1st of the 5 plausible values). N shows the total number of matched individuals from the treatment group. Matching variables include: Math Skills before School, Math Skills Starting School, Pre-primary Education, Parents' Education, Parents' Occupation, Student's Age. Standard errors are cluster-robust.

Table 5.3: The effects of having a female teacher for girls across countries

Country	ATT	S.E.	N
Canada	15.18***	4.32	3,474
Chile	11.15***	3.85	1,882
Chinese Taipei	1.71	1.71	2,087
Cyprus	-1.64	3.42	1,904
Czech Republic	0.60	3.70	2,066
Denmark	2.32	4.81	627
Finland	-6.73**	3.22	1,531
France	-6.73*	3.98	1,086
Germany	-13.16***	4.43	1,000
Hong Kong SAR	8.19	5.11	1,187
Iran, Islamic Rep. of	19.05***	5.04	2,333
Ireland	-5.32	4.24	1,266
Korea,Rep.of	2.24	2.78	1,831
Kosovo	9.39	10.86	313
Kuwait	-60.73***	9.74	2,018
Malta	12.97***	4.28	1,214
Bahrain	9.48*	5.02	3,131
Montenegro	-7.36*	4.30	1,099
Morocco	7.94	8.80	1,332
New Zealand	-5.35	5.32	806
Pakistan	-3.39	15.94	953
Philippines	-12.34	8.50	1,866
Poland	-2.21	2.88	3,590
Portugal	-2.39	3.50	1,511
Qatar	-39.61***	6.00	2,565
Serbia	-5.34	3.84	1,647
Bosnia and Herzegovina	-3.83	3.74	1,681
Singapore	24.94***	3.31	3,669
Slovak Republic	8.92***	3.09	2,292
South Africa	28.33***	6.13	2,941
Spain	-8.45***	2.71	3,415
Sweden	-9.59***	4.64	1,139
United Arab Emirates	-16.25***	4.41	7,386
Turkey	-4.13	5.64	2,252
Albania	17.98***	4.84	1,171
Macedonia, Rep. of	30.99***	8.01	621

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Note: Estimate represents the average treatment effect on girls with a female teacher obtained by 1:1 NN matching with replacement. Sampling weights are used. The dependent variable is the math test score (1st of the 5 plausible values). N shows the total number of matched individuals from the treatment group. Matching variables include: Math Skills before School, Math Skills Starting School, Pre-primary Education, Parents' Education, Parents' Occupation, Student's Age. Standard errors are cluster-robust.

that account for teacher quality in the following section on dual modelling and in Chapter 6 (Robustness Check). Similarly, we observe that in South Africa girls prosper with a female teacher, and boys' achievement is hampered by a male teacher, yet, the estimate for boys is not significant at conventional levels with a t-value of 1.48. Again, this could either point to the heterogeneity of the effect across genders or could be caused by the fact that female teachers in South Africa are systematically better than their male counterparts. In 21 countries the effect is only significant for either boys or girls which would again support the hypothesis that the effect differs across the two genders. Finally, in 12 countries there is no significant effect for both boys and girls. So, the results point to a double heterogeneity: effect for boys versus effect for girls, and heterogeneity across countries.

5.2 Dual modelling

Another way to obtain an estimate of the effect of having a same-sex teacher is a method reviewed by Zhao *et al.* (2021) and used, for example, by Hogrebe & Strietholt (2016). The procedure is mostly similar to the matching. First, the propensity score for the selection to treatment is estimated. Then, observations are matched using some mechanism — in our case 1:1 NN matching without and with replacement (as in the previous section). The difference only appears in the final stage. In the previous section, we simply compared the outcomes within the matched pairs and then obtained a weighted mean of this difference to obtain an estimate of the effect. With dual modelling, once we obtain the dataset with matched pairs we run a regression on the matched data including the matching variables to account for any remaining disbalance in the sample after matching, and additional variables that were not used for matching.

The advantages of this method are the following: first, we can account for the remaining disbalance after matching. Generally, matching was not a problem in our analysis when considering the pooled data. Still, the matching was not perfect on all individual matching variables. The imperfections of matching were more pronounced when looking at the analysis for each country separately. Second, since matching is combined with a regression analysis it is possible to include variables that were not used for matching and hence control, for example, as in our case, for the quality of the teacher. Moreover, this approach is less sensitive to misspecification in both the first and the second stage, and thus should lead to less biased estimates Schafer & Kang (2008). On

the other hand, since we bundle the matched pairs together and run a regression on the resulting dataset, we lose the explicit pairing of the observations which may be problematic since we are using special weights for each pair.

For the regression, we use a function from the *intsvy* package which takes into account the special design of TIMSS: the weights, the dependency of observations within one classroom, and the imputation of plausible values. As mentioned when presenting the results for regular matching, there are 5 plausible values provided in the TIMSS dataset. In the previous analysis, we chose to work with one of them and then checked that the results did not vary significantly for the others however the function from *intsvy* package allow us to work with all of them. This also means that the standard errors are more conservative than the cluster robust standard errors used in the previous section since now the standard errors also account for the imputation error.

As mentioned above there is no need to explicitly evaluate the matching procedure since it is the same as in the regular matching approach and it was done in the previous section. The results for both boys and girls are reported in Table 5.4. For both mechanisms, we report the estimate of the effect of having a same-sex teacher along with the standard error. Moreover, the total number of treatment and control units is reported as well as the number of unmatched control units that are hence discarded from the analysis. The controls included in the regression on the matched samples are all the variables used for matching to account for the remaining differences after the matching procedure, but also school and teacher characteristics (School socio-economic composition — ACDGSBC, Teacher's math education — ATDMMEM, and Teacher's age — ATBG03), and country dummies.

As can be seen from Table 5.4 the effect of having a male teacher for boys is negative for the two mechanisms. In 1:1 matching without replacement boys who have a male teacher score 11.26 points lower on the standardized test. This result is statistically significant at the conventional levels and corresponds to the estimate obtained from regular matching (-9.99). Similarly, when 1:1 matching with replacement is considered there is a significant negative effect of treatment: -16.84 which is again similar to the matching estimate: -17.1 but now the standard error is much larger and hence renders the estimate insignificant.

The two mechanisms also vary in the number of unmatched control observations. In 1:1 matching without replacement, naturally, the number of matched control units is the same as the number of treatment units, which means that

Table 5.4: The effects of having a same-sex teacher — dual modelling

	1:1 NN without replacement		1:1 NN with replacement	
	Boys	Girls	Boys	Girls
Estimate	-11.26**	4.96	-16.84	3.28
S.E.	5.22	6.93	11.62	9.12
Control	65,398	15,680	65,398	15,680
Treatment	20,946	70,886	20,946	70,886
Unmatched	45,231	0	53,185	5,044

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Note: Estimate represents the average treatment effect on boys with a male teacher and girls with a female teacher obtained by regression done on matched samples attained by 1:1 NN matching without and with replacement. The dependent variable is math test scores with 5 plausible values. Control and Treatment show the total number of students with female and male teachers. Unmatched is the number of control units disregarded from the analysis. Controls include matching variables (Math Skills before School, Math Skills Starting School, Pre-primary Education, Parents' Education, Parents' Occupation, Student's Age), country dummies, School Socio-economic Composition, Teacher's age, Teacher's Math Education.

a lot of information can be lost since we disregard 45,231 of the 65,398 control units. When we allow for replacement the situation gets even worse in the sense that even more control units remain unused: 53,185. Hence, we also see much larger standard errors when we allow for replacement because we are using fewer observations.

Turning now to the results for girls, they are also reported in Table 5.4. The results for 1:1 matching without replacement vary from the original matching, the estimate of the effect is much smaller in absolute size (4.96 compared to 9.86) this in combination with the fact that the standard errors are much larger leads to the estimate being insignificant. As argued above, the replacement is key for the girls' dataset since there are not enough control units, some of them have to be matched to multiple treatment units.

The results for 1:1 matching with replacement suggest a positive effect for girls of having a female teacher. Specifically, girls with a female teacher score 3.28 points more on a standardized math test. Although this estimate is larger in absolute terms than the one obtained by regular matching (1.48), the large standard errors still mean this estimate is statistically insignificant (like for regular matching).

To sum up, although the estimates are similar to the ones obtained from regular matching, there is a difference in the standard errors which in some cases affect the statistical significance.

Let us now turn to the results across countries. As with the regular matching, we use the 1:1 NN matching with replacement since it outperformed match-

Table 5.5: The effects of having a male teacher for boys across countries — dual modelling

Country	Estimate	S.E.	Control	Treatment
Canada	-10.66	8.36	3,478	557
Chile	0.54	4.94	1,885	445
Chinese Taipei	-1.03	4.08	2,247	867
Cyprus	-2	4.21	1,728	632
Czech Republic	0.97	9.63	2,175	159
Denmark	-3.10	7.36	603	395
Finland	-0.74	6.15	1,584	567
France	-5.01	8.32	1,098	268
Germany	2.56	6.99	987	117
Hong Kong SAR	-3.53	5.36	1,235	877
Iran, Islamic Rep. of	-4.73	11.61	1,362	925
Ireland	-0.13	7.95	1,156	427
Korea,Rep.of	-8.33	7.56	1,936	402
Kosovo	-14.55	19.30	298	151
Kuwait	-13.56	22.01	1,198	426
Malta	-14.13	9.31	1,263	174
Bahrain	13.73*	7.65	2,052	1,528
Montenegro	30.62**	13.95	1,237	100
Morocco	0.92	12.14	1,423	1,202
New Zealand	11.74	9.10	784	224
Pakistan	-30.89	33.88	387	1,166
Philippines	-9.99	10.66	2,038	443
Poland	-9.05	5.90	3,640	294
Portugal	1.16	7.93	1,574	227
Qatar	-4.02	9.81	1,743	656
Serbia	13.94	12.18	1,603	221
Bosnia and Herzegovina	2.99	12.12	1,757	124
Singapore	5.37	5.67	3,397	1,220
Slovak Republic	-2.02	6.63	2,339	170
South Africa	-9.20*	5.39	2,673	1,857
Spain	3.05	4.77	3,540	1,312
Sweden	0.82	9.46	1,067	318
United Arab Emirates	-8.27	9.12	6,144	1,114
Turkey	-6.09	6.39	1,886	1,218
Albania	26.33	21.74	1,199	88
Macedonia, Rep. of	-29.37	21.29	682	75

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Note: Estimate represents the average treatment effect on boys with a male teacher obtained by regression run on matched samples attained by 1:1 NN matching with replacement. The dependent variable is math test scores with 5 plausible values. Control and Treatment show the total number of students with female and male teachers. Controls include matching variables (Math Skills before School, Math Skills Starting School, Pre-primary Education, Parents' Education, Parents' Occupation, Student's Age), School Socio-economic Composition, Teacher's age, Teacher's Math Education.

ing without replacement and the results are easily comparable to regular matching. Again, the matching procedure has already been evaluated in the previous section, so, we will mainly focus on describing the results. For boys, they are reported in Table 5.5. Again, the effect varies across countries not only in size but also in sign. However, only in 3 of the 11 countries is the effect significant. As in the previous section, we find a significant positive effect for boys in Montenegro. Additionally, we find a borderline significant positive effect for Bahrain in contrast to results for matching which suggested an insignificant effect. Similarly, there is a negative effect for South Africa with a t-value of 1.71 which makes the estimate significant at the 10% significance level. This result corresponds to the estimate attained by regular matching where however the t-values was 1.48 and so it was just below the 10% significance level.

Korea, Kosovo, and Spain reported a significant effect using regular matching, but the combination of increased standard errors with a decrease in the absolute size of the effect leads to the estimates becoming insignificant. In the case of Pakistan and Poland despite the estimate remaining similar, the increased standard error makes the estimates insignificant. Yet, in the case of Poland, the estimate is close to the 10% significance level with a t-value of 1.53. For the rest of the countries, the estimate remains statistically indistinguishable from zero.

As for the girls, the results are reported in Table 5.6. Like in the results for boys, the standard errors are larger in comparison to regular matching which in combination with a decrease in the absolute value of the effect for a majority of countries leads to a lot of estimates becoming statistically insignificant. Actually, the majority of countries report no significant effect of having a female teacher for girls. The effect remains significant in Singapore (22.04), South Africa (23.16), and Macedonia (25.02). In all three cases, the suggested effect is positive.

To conclude this section, the estimate of the effect remains relatively similar for both regular matching and dual modelling, especially for the pooled data. However, we can see differences in the size of the standard errors. The standard errors obtained via dual modelling are more conservative, the difference is given by two factors: first, the standard errors take into account the imputation error — we are working with all five plausible values in comparison to just one in the regular matching. Second, the dual modelling procedure itself produces more conservative standard errors than regular matching. Hence, we consider results from both procedures to be valid input in deciding whether to accept or reject

Table 5.6: The effects of having a female teacher for girls across countries — dual modelling

Country	Estimate	S.E.	Control	Treatment
Canada	8.87	7.34	498	3,474
Chile	3.47	6.45	432	1,882
Chinese Taipei	1.40	4.89	796	2,087
Cyprus	-1.31	5.51	669	1,904
Czech Republic	-0.06	9.15	159	2,066
Denmark	-2.39	6.47	392	627
Finland	-6.15	5.83	562	1,531
France	1.64	7.15	286	1,086
Germany	-10.21	7.04	110	1,000
Hong Kong SAR	8.69	6.11	686	1,187
Iran, Islamic Rep. of	8.48	13.05	98	2,333
Ireland	-2.53	7.20	339	1,266
Korea,Rep.of	4.94	5.75	355	1,831
Kosovo	11.11	14.82	123	313
Kuwait	-57.28	39.22	43	2,018
Malta	10.97	7.38	163	1,214
Bahrain	6.12	27.52	124	3,131
Montenegro	-2.76	14.90	87	1,099
Morocco	-9.88	12.33	1,147	1,332
New Zealand	-5.02	7.84	189	806
Pakistan	3.87	38.61	293	953
Philippines	-11.36	10.90	424	1,866
Poland	-2.62	7.22	299	3,590
Portugal	7.92	7.47	204	1,511
Qatar	-19.11	16.18	271	2,565
Serbia	6.67	9.37	180	1,647
Bosnia and Herzegovina	-2.20	10.57	121	1,681
Singapore	22.04***	5.74	990	3,669
Slovak Republic	6.76	6.50	186	2,292
South Africa	23.16***	7.71	1,946	2,941
Spain	-1.85	6.21	1,119	3,415
Sweden	-3.31	8.23	313	1,139
United Arab Emirates	-15.01	9.72	576	7,386
Turkey	-1.51	5.78	1,333	2,252
Albania	1.83	24.92	77	1,171
Macedonia, Rep. of	25.02***	12.30	90	621

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Note: Estimate represents the average treatment effect on girls with a female teacher obtained by regression run on matched samples attained by 1:1 NN matching with replacement. The dependent variable is math test scores with 5 plausible values. Control and Treatment show the total number of students with male and female teachers. Controls include matching variables (Math Skills before School, Math Skills Starting School, Pre-primary Education, Parents' Education, Parents' Occupation, Student's Age), School Socio-economic Composition, Teacher's age, Teacher's Math Education.

the hypothesis that pupils perform better when having a same-sex teacher.

Combining the results from the two approaches, there is sound evidence for a negative effect for boys when considering the pooled data. Conversely, the results from both approaches suggest no significant effect for girls for the pooled data. For the country by country analysis, Montenegro shows a consistent positive effect of having a same-sex teacher for boys. There is also persuasive evidence of a negative effect in Poland and South Africa. For girls, Singapore, South Africa, and Macedonia show a consistent positive effect of having a same-sex teacher for girls but the evidence is also persuasive for Malta. On the contrary, in Germany, Kuwait, and the United Arab Emirates there is a reasonable evidence for a negative effect for girls. To further explore the robustness of these results, we run several alternative models in the next chapter.

Chapter 6

Robustness Check

In this section, we will explore several alternative models to check for the robustness of the results presented in the previous chapter. Our main focus is on evaluating the results obtained from regular matching. First, we will try an alternative model specification, specifically, we will try to run a restricted model where matching is done on fewer variables and an extended one where matching is done on more variables than in the original model. Also, we will explore an alternative matching algorithm. Second, several OLS models will be presented and discussed. Finally, we will repeat the original analysis on a dataset with imputed missing values. For brevity reasons, the robustness check is only presented for boys.

6.1 Matching specifications

As mentioned above, this subsection explores various specifications of the main model to check the robustness of the estimates. First of all, we trim down the model and use only a few key variables. Specifically, we will use these variables in our restricted model: Math skills before the start of school (ASBHENA), Pre-primary education (ASDHAPS), Parents' education (ASDHEDUP), and Parents' occupation (ASDHOCCP). The following variables were dropped from the original model: Math skills starting school (ASBHENT) — this variable was dropped mainly because of the unclear timing formulation (if a survey question is misunderstood by the parents it is possible that this variable is already affected by the treatment and hence its' inclusion in the model would be inappropriate). Age of the student (ASBH05) — this variable was the only one that was insignificant in the determination of the propensity score.

The exclusion of the variables and downsizing of the model should have two general consequences: a higher possibility of satisfaction of the common support assumption, but also a lower probability of satisfying the conditional independence assumption (CIA) (Black & Smith 2004). In our case, the common support was not a major problem when considering the pooled data, but there were some countries where matching was imprecise. With the reduced model it should be easier to find a proper counterfactual for each observation. Moreover, as fewer variables are included in the model, fewer observations are lost due to missing values, with a larger dataset, the estimates should be more precise. Specifically, the original pooled dataset had 86,344 observations, while the new dataset includes 104,241 students — this is a significant increase. Nevertheless, to make the model directly comparable to the original we will perform the analysis on the same dataset as the original model. Still, we remain cautious about the risk that the CIA will not hold, and hence the whole matching procedure could be jeopardized.

The second model is a bit expanded compared to the original model. In addition to the six matching variables, we also add a Teachers' age (ATBG03), Math education of the teacher (ATDMMEM), and School's socioeconomic composition (ACDGSBC). The rationale behind this model is the following: besides the possibility that students are sorted to teachers based on their initial skills and that the sorting can be influenced by the parents, we also account for the possibility that more experienced and better trained teachers may have a say in choosing the students they will teach. Moreover, the sort can be affected by the socioeconomic composition of the school in a way that it may be easier to choose a teacher depending on the affluence of the school's pupils. For both models, we employ the exact match on the country variable so that only observations from the same country can be matched together.

The results are reported in Table 6.1. We only report the weighted estimate that accounts for the sampling weights provided by TIMSS. We use the cluster robust standard errors and we also report the number of treated units. As in the original model, the first plausible value of the math score is considered as the outcome variable.

Generally, the results for the restricted model closely resemble the results of the original model suggesting a significant negative effect of having a same-sex teacher for the boys: -12.2 without replacement and -15.5 with replacement. When looking at the extended specification of the model we see that while the estimate for matching without replacement (-9.29) is similar to the origi-

Table 6.1: The effects of having a male teacher for boys — restricted and extended model

	Without replacement		With replacement	
ATT	-12.20**	-9.29	-15.50**	-10.48
S.E.	5.52	5.85	7.27	9.28
N	20,167	20,167	20,946	20,946
Specification	Restricted	Extended	Restricted	Extended

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Note: Estimate represents the average treatment effect on boys with a male teacher obtained by 1:1 NN matching without and with replacement. Sampling weights are used. The dependent variable is math test scores (1^{st} of the 5 plausible values). N shows the total number of matched individuals from the treatment group. Matching variables include: Math Skills before School, Pre-primary Education, Parents' Education, Parents' Occupation for the restricted model. For the extended model matching variables moreover include: Math Skills Starting School, Student's Age, School Socio-economic Composition, Teacher's age, and Teacher's Math Education. Standard errors are cluster-robust. Matching is done within countries.

nal (-9.99) the estimate for matching with replacement (-10.48) is smaller in absolute terms compared to the original model (-17.1). Also, the standard errors for the extended model are larger compared to the restricted and to the original model. This renders the estimates statistically insignificant. As with the original model, there was no major issue when considering the quality of matching for the pooled data.

The results for the two alternative specifications country by country are reported in Table 6.2, the algorithm used for matching is the same as in the original specification — 1:1 NN matching with replacement. This makes the results easily comparable, again we use the weights provided in the TIMSS dataset to estimate the ATT and the standard errors are cluster robust. It is immediately noticeable that the results vary quite significantly between the two models but also compared to the original model. Only in Korea and Montenegro is the effect significant in both the reduced and the extended model. Since the effect for the two countries is also significant in the original settings we can regard the estimates as relatively robust. On the contrary, the effect for Poland and Spain, is insignificant in both alternative specifications. For Kosovo and Pakistan that also reported a significant effect in the original settings, the effect is confirmed in the restricted model while in the expanded model the effect becomes insignificant.

Generally, the pattern from the pooled analysis is repeated across the countries. There are fewer countries reporting a statistically significant effect under the extended specification (5) compared to the reduced model (13). In both

Table 6.2: The effects of having a male teacher for boys across countries — restricted and extended model

	Restricted		Extended		
Country	ATT	S.E.	ATT	S.E.	N
Canada	-2.69	5.15	-7.50	7.93	557
Chile	-7.57*	4.60	-1.40	6.68	445
Chinese Taipei	-4.21	4.34	-4.12	4.28	867
Cyprus	-13.48**	5.77	1.27	5.10	632
Czech Republic	32.19**	13.05	20.51	17.10	159
Denmark	0.38	6.59	-0.25	6.16	395
Finland	9.84**	4.73	1.33	6.00	567
France	-16.53**	8.16	-3.82	14.55	268
Germany	-14.35**	7.10	-8.09	10.45	117
Hong Kong SAR	-0.21	3.81	-4.44	8.30	877
Iran, Islamic Rep. of	-15.38*	9.02	-10.30	8.40	925
Ireland	-0.78	5.83	-4.70	6.71	427
Korea,Rep.of	-14.27***	3.90	-19.34***	6.43	402
Kosovo	-29.90**	12.92	0.32	19.48	151
Kuwait	-17.30	24.13	11.27	29.99	426
Malta	-1.44	9.29	-4.69	12.42	174
Bahrain	6.84	6.59	14.84**	6.03	1,528
Montenegro	45.49***	13.90	28.04*	14.52	100
Morocco	-9.00	7.88	-12.55	8.90	1,202
New Zealand	-22.46**	11.22	14.79	13.53	224
Pakistan	-31.61***	11.73	-29.04	27.94	1,166
Philippines	-4.95	7.24	16.02	16.50	443
Poland	-7.33	6.46	-3.75	10.29	294
Portugal	10.30	10.09	-8.92	8.79	227
Qatar	-6.31	11.27	-1.42	10.82	656
Serbia	-10.49	10.42	-13.21	11.12	221
Bosnia and Herzegovina	-5.18	11.46	5.50	13.39	124
Singapore	6.09	5.83	-7.50	6.59	1,220
Slovak Republic	-8.22	7.84	-8.26	12.76	170
South Africa	-7.85	5.78	-15.85***	5.44	1,857
Spain	-0.09	4.74	-1.98	6.30	1,312
Sweden	2.67	6.26	-2.60	6.35	318
United Arab Emirates	-14.74*	8.52	-9.03	9.02	1,114
Turkey	-9.23	6.11	-4.06	9.33	1,218
Albania	-4.78	17.58	55.70**	26.28	88
Macedonia, Rep. of	-12.70	25.58	0.40	19.95	75

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Note: Estimate represents the average treatment effect on boys with a male teacher obtained by 1:1 NN matching with replacement. Sampling weights are used. The dependent variable is math test scores (1^{st} of the 5 plausible values). N shows the total number of matched individuals from the treatment group. Matching variables include: Math Skills before School, Pre-primary Education, Parents' Education, Parents' Occupation for the restricted model. For the extended model matching variables moreover include: Math Skills Starting School, Student's Age, School Socio-economic Composition, Teacher's age, and Teacher's Math Education. Standard errors are cluster-robust.

cases however, there is no significant effect of the treatment for the majority of countries, as in the original model.

Regarding the quality of the matching procedure, as expected, the matching is easier when it is done on fewer variables so the matches in the restricted model are good for all the countries. The downside of this approach is the possible violation of the Conditional Independence Assumption (CIA) which would make the estimates attained by this model unreliable. When considering the quality of the matching for the extended model, it generally worked satisfactorily although the quality is worse than for both the reduced and the original model. Again, this was to be expected simply because we are matching on a larger number of variables.

In summary, the matching results are quite sensitive to the specification. This fact reduces the confidence in the results obtained by the original model. For Korea and Montenegro, the results are similar across the original and the alternative specifications. Still, the original model based on the economic theory is the most reliable. As mentioned above, the reduced model is at risk of violating the CIA and the extended model is also at risk of misspecification.

Another way to inspect the robustness of the original results is to try a different matching algorithm. With the original specification, we saw that using the pool data the results for the two matching algorithms (1:1 without replacement and 1:1 with replacement) were qualitatively similar for boys. We will check the country by country results using a different matching algorithm. Specifically, we use the original specification with optimal full matching.

Optimal full matching is a form of sub-classification that assigns each unit into a subclass where the units receive a match. The advantage of weighting observations in subclasses is that observations do not need to be discarded. The procedure is optimal because the distance between treatment and controls is as small as possible within each subclass (Hansen & Klopfer 2006). After the matching procedure, the outcome of the treated observation is compared to the weighted outcome of the control observations as there are multiple control observations in one subclass with one treatment unit.

The results are reported in Table 6.3. One apparent advantage of this matching procedure is the usage of more information as none of the observations (not even those from the control group) is discarded, but rather reweighted. However, the estimates are not much more precise compared to the original specification — the standard errors are similar in the majority of countries. In Korea, Kosovo, and Montenegro the effect remains significant and

Table 6.3: The effects of having a male teacher for boys across countries — Optimal Full Matching

Country	ATT	S.E.	N
Canada	-4.19	6.16	557
Chile	-5.57	6.36	443
Chinese Taipei	-3.79	4.55	867
Cyprus	2.82	5.42	628
Czech Republic	14.45	10.25	159
Denmark	-2.94	6.61	376
Finland	1.62	4.83	567
France	-5.04	10.42	266
Germany	4.13	8.38	117
Hong Kong SAR	-10.03*	5.64	874
Iran, Islamic Rep. of	-9.15	6.43	842
Ireland	7.72	6.85	427
Korea,Rep.of	-19.52***	5.30	402
Kosovo	-27.13***	9.98	144
Kuwait	-6.42	26.33	390
Malta	-15.46	10.00	171
Bahrain	7.96	5.68	1,310
Montenegro	40.48***	11.59	100
Morocco	-8.12	8.09	1,102
New Zealand	16.64	11.29	224
Pakistan	-19.99	29.95	262
Philippines	6.66	13.72	443
Poland	1.22	6.21	294
Portugal	2.01	11.59	227
Qatar	4.88	10.81	656
Serbia	-8.23	12.49	221
Bosnia and Herzegovina	5.06	10.72	124
Singapore	-3.74	6.65	1,218
Slovak Republic	-7.33	11.56	170
South Africa	-22.67***	5.62	1,844
Spain	-12.69	7.91	1,312
Sweden	2.64	9.54	316
United Arab Emirates	-13.45*	8.02	1,101
Turkey	-2.42	9.15	1,156
Albania	9.24	21.08	88
Macedonia, Rep. of	-3.18	21.02	75

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Note: Estimate represents the average treatment effect on boys with a male teacher obtained by Optimal Full Matching. Sampling weights are used. The dependent variable is math test scores (1st of the 5 plausible values). N shows the total number of matched individuals from the treatment group. Matching variables include: Math Skills before School, Math Skills Starting School, Pre-primary Education, Parents' Education, Parents' Occupation, and Student's Age. Standard errors are cluster-robust.

quantitatively similar. For Pakistan and Poland, the smaller absolute size in combination with increased standard errors render the estimates insignificant as opposed to the original model. In Spain, despite a slight increase in the absolute size of the effect the estimate is not statistically significant although the t-value of 1.61 is close to the critical value of 1.65 which would make the estimate significant at the 10% significance level. In South Africa, the effect becomes significant due to a much larger size in absolute terms (-5.51 in the original compared to -22.67 in the model using Optimal Full Matching). Hong Kong and the United Arab Emirates, also report a negative effect significant at the 10% level under this mechanism. Generally, the results remain similar to the original results, with most countries reporting no significant effect of treatment.

6.2 OLS

Generally, we assumed in line with the literature (for example, Cho 2012; Krieg 2005; Dee 2007; and Ammermüller & Dolton 2006) that the assignment of pupils to teachers is non-random and this was partly confirmed by the results of the first stage of the matching procedure. Most selected variables in the original model were significant predictors in the treatment assignment. Yet, this evidence is mostly observing the correlations and so cannot be taken as causal. It may be that the correlation is spurious and that indeed the assignment of pupils to teachers is random. Inclusion of the controls into the OLS model may be enough to reduce the bias in the estimate of the effect of the student-teacher gender match. Moreover, the usage of OLS compared to simple matching allows the inclusion of additional variables as controls since we do not have to restrict the choice only to variables known/collected at the time of the selection to treatment. We again run two analyses, the pooled one and a country by country analysis.

We use 4 different specifications for pooled data: Original includes variables from the original matching specification — included independent variables: Math skills before the start of school (ASBHENA), Math skills at the start of school (ASBHENT), Pre-primary education (ASDHAPS), Parents' education (ASDHEDUP), Parents' occupation (ASDHOCCP), Student's age (ASBH05), and country dummies. Reduced is similar to the reduced specification used in the robustness check (Math skills before the start of school (ASBHENA), Pre-primary education (ASDHAPS), Parents' education (ASD-

HEDUP), Parents' occupation (ASDHOCCP), and country dummies). The extended model also corresponds to the model used in the robustness check above (Math skills before the start of school (ASBHENA), Math skills at the start of school (ASBHENT), Pre-primary education (ASDHAPS), Parents' education (ASDHEDUP), Parents' occupation (ASDHOCCP), Student's age (ASBH05), School socio-economic composition (ACDGSBC), Teacher's math education (ATDMMEM), Teacher's age (ATBG03), and country dummies). Finally, the Extra specification includes a bunch of extra variables that account also for the student's interests (this is suggested to be a significant predictor of the results on standardized tests by Nye *et al.* 2018), further information on students' background, more controls for teacher quality, and some more information on school overall quality. The full list of variables is available in Appendix C.

Table 6.4: The effects of having a male teacher for boys — OLS

	Original	Reduced	Extended	Extra
Estimate	-10.44*	-9.98*	-9.12	3.06
S.E.	6.3	5.42	6.47	4.33
N	86344	108703	86344	30813
R-squared	0.6	0.6	0.61	0.68
Controls	Original	Reduced	Extended	Extra

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Note: Dependent variable is math test scores with 5 plausible values. N is the number of observations. Original model controls: Math Skills before School, Math Skills Starting School, Pre-primary Education, Parents' Education, Parents' Occupation, Student's Age. Reduced model controls: Math Skills before School, Pre-primary Education, Parents' Education, Parents' Occupation, and country dummies. Extended model controls: Math Skills before School, Math Skills Starting School, Pre-primary Education, Parents' Education, Parents' Occupation, Student's Age, School Socio-economic Composition, Teacher's age, Teacher's Math Education, and country dummies. Extra model controls: see Appendix C

The results for the OLS pool are reported in Table 6.4. Apart from the estimate of the effect of student-teacher gender match. We also report the standard error, number of observations used for each analysis, R-squared, and indicator of which controls were used. We use the package *intsvy* to estimate the effect. This package accounts for the specific design of TIMSS data: weights and imputations. The outcome variable is all five plausible values of the results of a math test. Moreover, jackknife replication is used to produce the standard errors which account for both the sampling and imputation error. Generally, the standard errors produced by this package are more conservative than cluster robust standard errors used in the main specification as they also account for the imputation error.

The three specifications — Original, Reduced, and Expanded suggest a negative effect of having a same-sex teacher for boys. The estimates are borderline statistically significant with a t-value around 1.5. While the size of the estimate is similar to the 1:1 matching without replacement (-9.99) the standard errors are slightly larger and are more similar to those reported for matching with replacement. This confirms that the functions within the *intsvy* package produce conservative standard errors. For the Extra specification, the estimated effect is positive but insignificant. Still, this would lead us to believe that the other three models are lacking some variables that cause the estimate to be biased. Further inspection of the results shows this is merely a property of the sample — a higher number of variables means a higher probability of some missing values, hence more observations are discarded. In the end, only 30,813 observations are used for this model. When other models are run on this reduced sample the estimates are similar to the model with extra variables. So, the pooled results of OLS are in line with the results we got from matching: negative effect, yet the significance levels vary across specifications which we also observed in the previous part of the robustness check.

When considering the country by country results we only use the Extended specification as the other OLS specifications are relatively similar and the Extended specifications includes also a measure of teachers' quality compared to just the matching variables. The results are reported in Table 6.5, we again see some heterogeneity across the countries. In 13 countries the effect is positive, in 23 countries the effect is negative. However, only in 7 countries, the estimated effect is statistically different from zero. Specifically, we find a positive effect in Bahrain (15.62) and Montenegro (29.59) — this finding supports the results from dual modelling and in the case of Montenegro also from matching. On the other hand, we find a negative effect in South Africa (-23.27) — as in dual modelling, and in Korea (-10) — as when using matching. But also in Canada (-13), the United Arab Emirates (-12.74), and Malta (-11.53) where the absolute value of the estimate increased compared to the regular matching. Again, in the majority of the countries, the effect is statistically insignificant.

To sum up, the results from OLS mostly support the results from matching and dual modelling. Especially when considering the pooled data. The main difference from the original matching lies in the size of standard errors. When considering the results across countries we can again see that the results are quite sensitive. Only in 2 out of the 6 countries where a significant effect was found using matching did we find a significant effect when using OLS.

Table 6.5: The effects of having a male teacher for boys across countries — OLS

Country	Estimate	S.E.	R^2	n
Canada	-13**	7.33	0.22	4,035
Chile	-0.38	7.79	0.28	2,330
Chinese Taipei	-3.70	5.66	0.16	3,114
Cyprus	-6.90	7.20	0.17	2,360
Czech Republic	5.64	7.28	0.22	2,334
Denmark	-6.85	10.83	0.18	998
Finland	6.57	8.78	0.20	2,151
France	-8.56	9.45	0.31	1,366
Germany	4.33	8.68	0.26	1,104
Hong Kong SAR	-6.71	6.75	0.09	2,112
Iran, Islamic Rep. of	-4.71	16.35	0.18	2,287
Ireland	-0.30	7.68	0.17	1,583
Korea,Rep.of	-10*	7.79	0.14	2,338
Kosovo	-18.91	16.92	0.24	449
Kuwait	-4.45	23.75	0.13	1,624
Malta	-11.53*	6.70	0.22	1,437
Bahrain	15.62**	7.41	0.06	3,580
Montenegro	29.59***	9.19	0.24	1,337
Morocco	-1.91	12.81	0.19	2,625
New Zealand	11.23	13.04	0.29	1,008
Pakistan	-30.54	23.32	0.15	1,553
Philippines	5.15	11.30	0.24	2,481
Poland	4.95	7.32	0.22	3,934
Portugal	-2.69	8.29	0.22	1,801
Qatar	1.89	7.62	0.22	2,399
Serbia	5.05	11.14	0.33	1,824
Bosnia and Herzegovina	2.60	8.33	0.17	1,881
Singapore	-6.80	6.80	0.22	4,617
Slovak Republic	-6.60	10.03	0.33	2,509
South Africa	-23.27***	11.20	0.31	4,530
Spain	-4.16	6.57	0.22	4,852
Sweden	2.22	10.75	0.23	1,385
United Arab Emirates	-12.47*	5.73	0.12	7,258
Turkey	-0.76	8.53	0.34	3,104
Albania	15.97	11.29	0.23	1,287
Macedonia, Rep. of	-20.45	13.48	0.29	757

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Note: Dependent variable is math test scores with 5 plausible values. N is the number of observations. Model controls: Math Skills before School, Math Skills Starting School, Pre-primary Education, Parents' Education, Parents' Occupation, Student's Age, School Socio-economic Composition, Teacher's age, Teacher's Math Education.

On the other hand, in all 3 countries that reported a significant effect using dual modelling did we find a significant effect using the OLS. Still, the OLS estimates should be interpreted with caution. As mentioned before they are at risk of the extrapolation error — especially in some specific countries the control and treatment samples are not well balanced and the estimates differed significantly from estimates obtained by matching.

6.3 MICE data

As is apparent from the pooled analyses of the OLS data, the sample selection can play an important role. The missing values in the data are problematic in two ways. Firstly, the analysis loses power as the size of the standard errors increases (Bouhlila & Sellaouti 2013). This may be a problem in our analysis, as we saw in the analysis across countries for boys, most of the estimates were not statistically significant. In some countries, a sizeable effect was statistically insignificant due to a large standard error. Secondly, there is a risk that the missing values will bias the estimate of interest. Disregarding incomplete cases should not be a problem when the data are Missing Completely at Random (MCAR) — in this case, the estimates are more imprecise but not biased (Bouhlila & Sellaouti 2013). However, this is a pretty strong assumption and as seen in the OLS robustness check it may not hold in our data.

Hence, we will impute the missing values using the Multiple Imputation by Chained Equations (MICE) approach. This method imputes missing datasets based on a set of imputation models. For each variable with missing values, there is a model (Bouhlila & Sellaouti 2013). This method has been shown to work in imputing values to large, national survey datasets. An advantage of this method is that it works for various types of variables — categorical, binary, and continuous (Bouhlila & Sellaouti 2013). This makes it particularly suitable for imputing data in TIMSS, this is also documented by Bouhlila & Sellaouti (2013). Hogrebe & Strietholt (2016) also use MICE in their analysis of large international survey data on educational outcomes.

Once the dataset with imputed values is constructed we run the analysis using regular matching as in the original model. As for the pooled results, they are reported in Table 6.6. We again report the estimate taking into account the weights provided by TIMSS and use the cluster robust standard errors.

Compared to the original model, the standard errors are indeed smaller, so the estimates are more precise when using the MICE dataset. This is given by

Table 6.6: The effects of having a male teacher for boys — MICE data

	1:1 without replacement	1:1 with replacement
ATT	-9.21***	-11.76**
S.E.	3.19	5.13
N	36,416	37,964

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Note: Estimate represents the average treatment effect on boys with a male teacher obtained by 1:1 NN matching without and with replacement. Sampling weights are used. The dependent variable is math test scores (1^{st} of the 5 plausible values). N shows the total number of matched individuals from the treatment group. Matching variables include: Math Skills before School, Math Skills Starting School, Pre-primary Education, Parents' Education, Parents' Occupation for the restricted model, and Student's Age. Standard errors are cluster-robust. Matching is done within countries.

a larger number of both control and treatment units available for matching. Similar to the original results, the estimate is significant and negative for the two matching algorithms. As in the original results, both algorithms worked well in removing the disbalance between the treatment and the control group. The estimate for the 1:1 matching without replacement is almost identical to the original results (-9.99 original dataset, -9.21 MICE dataset). When we allow for replacement the estimate differs more from the original (-17.1 original dataset, -11.76 MICE dataset) suggesting that the original estimates may have been biased downwards due to non-random missing values in the sample. On the other hand, the estimates from the two mechanisms are converging towards each other.

Turning now to the country by country results, these are reported in Table 6.7. First of all, as for the pool of countries, the MICE imputed data manage to decrease the standard errors in 24 out of the 36 countries. For the 12 countries where standard errors were not decreased they remained close to the original, i.e., in no country did the standard errors increase significantly. In Korea, Montenegro, Poland, and Spain the significant effects found in the original data are confirmed on the MICE dataset. In Kosovo and Pakistan, although the standard errors are much smaller compared to the original data there is also a significant difference in the size of the estimate, so the effect in both of these countries becomes insignificant. This suggests that the data are not missing completely at random in these countries and the original estimates may be biased. Similarly for Chile, the estimate is only significant in the analysis of MICE data due to an increase in the absolute size of the effect. This is also the case for South Africa, where the original estimate was close to being significant

Table 6.7: The effects of having a male teacher for boys across countries — MICE data

Country	ATT	S.E.	N
Canada	-7.36	4.56	1,115
Chile	-9.24*	5.11	580
Chinese Taipei	-4.75	3.09	1,019
Cyprus	2.26	4.81	893
Czech Republic	8.54	7.87	209
Denmark	-4.87	4.89	944
Finland	1.94	4.48	743
France	-11.01	7.41	438
Germany	-1.59	6.13	264
Hong Kong SAR	-4.84	3.32	1,279
Iran, Islamic Rep. of	-10.50	7.61	1,198
Ireland	-1.76	5.94	576
Korea,Rep.of	-16.28***	5.66	450
Kosovo	-6.81	7.84	497
Kuwait	-25.13	17.78	1,079
Malta	-4.90	9.73	300
Bahrain	3.84	5.48	2,571
Montenegro	27.46**	11.53	248
Morocco	2.32	6.31	3,090
New Zealand	0.72	5.84	684
Pakistan	-16.80	11.72	2,560
Philippines	-6.99	8.68	724
Poland	-9.08***	2.98	357
Portugal	-2.42	8.21	277
Qatar	7.28	10.15	1,120
Serbia	4.47	12.51	256
Bosnia and Herzegovina	-11.51	10.44	208
Singapore	2.26	5.06	1,411
Slovak Republic	-0.01	4.94	191
South Africa	-10.35**	4.59	4,918
Spain	-8.94**	3.54	1,973
Sweden	-2.10	3.84	629
United Arab Emirates	-5.29	6.36	3,307
Turkey	-2.21	5.62	1,489
Albania	-0.001	16.17	184
Macedonia, Rep. of	12.81	16.14	183

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Note: Estimate represents the average treatment effect on boys with a male teacher obtained by 1:1 NN matching with replacement. Sampling weights are used. The dependent variable is math test scores (1st of the 5 plausible values). N shows the total number of matched individuals from the treatment group. Matching variables include: Math Skills before School, Math Skills Starting School, Pre-primary Education, Parents' Education, Parents' Occupation, Student's Age. Standard errors are cluster-robust.

at the 10% level.

To sum up, the results attained from the dataset with imputed values by chained equations are similar to the original results when considering the pool of countries. Especially when we do not allow for replacement. But again, both models suggest a negative effect of having a male teacher for boys. While in the pooled results there seems to not be any serious bias — the missing values are missing at random, when we look at individual countries we see some evidence that this may not be the case. Despite some of these deviations, the effect is confirmed for most of the countries that reported a significant effect in the original settings, and again in most countries, the effect is not statistically different from zero. So, running the analysis on MICE data generally corroborates the original results.

Chapter 7

Conclusion

The purpose of this thesis was to evaluate the effect of a same-sex teacher on students' educational outcomes. Specifically, the results of standardized tests in mathematics of 4^{th} grade students. So far, the evidence in the existing literature is rather inconclusive: a positive effect of gender match was found by Dee (2007) in the US and by Andersen & Reimer (2019) in Denmark but also by Hermann & Diallo (2017) in an international study in Western Europe. On the other hand, Cho (2012) in an international study, and Krieg (2005) and Winters *et al.* (2013) in the US suggest no effect of a same-sex teacher. Antecol *et al.* (2015) and Beilock *et al.* (2010) even indicate a negative effect of the student-teacher gender match. Moreover, Hwang & Fitzpatrick (2021) point to an interesting heterogeneity for boys and girls — while girls prosper from the gender match there is no effect for boys. We analyse the effect using international data from TIMSS for 36 countries.

For the analysis, we use Propensity Score Matching (PSM) to overcome the possible selection bias due to the non-random selection of pupils to teachers. We match students based on their skills and characteristics before starting school as well as family background. This strategy allows us to estimate the effect on 4^{th} graders in international settings. To our knowledge, this has not been done before. We run the analysis on pooled data but also for each country separately.

As for the pooled results, we estimate the Propensity Score in the first stage of the analysis using a logit model. The results of the first stage indeed suggest a non-random selection to treatment (having a same-sex teacher) for both boys and girls. Students' skills and characteristics as well as the parents' characteristics appear to be decisive for the selection to treatment. Data are

matched using two variants of the Nearest Neighbour mechanism: without replacement and with replacement. To obtain the estimate of interest we use two techniques: First, regular matching where we take the difference in outcome between treatment and control unit within each pair and then take a weighted average across the matched pairs to obtain the Average Treatment Effect on the Treated (ATT). Second, instead of simply comparing the outcomes for each pair we run a regression on the balanced dataset (this should eliminate the main flaw of regression approaches — extrapolation error) to account for any remaining disbalances after matching but also to control for additional variables not used for matching.

While we find no significant effect of having a same-sex teacher for girls — this result is in line with the literature that does not indicate significant effects of teachers' gender (Cho 2012; Krieg 2005; Winters *et al.* 2013), estimates across the pooled models suggest a negative effect of having a same-sex teacher for boys — Antecol *et al.* (2015) and Beilock *et al.* (2010) found negative effects of student-teacher gender match but mainly for girls.

When we turn to the analysis for each country separately, we discover important heterogeneities in the effect. For boys, we find a significant positive effect of student-teacher gender match across all models (both main models and models in robustness check) for Montenegro. On the other hand, in most models we also find a negative significant effect for South Korea, Poland, and South Africa. However, in the majority of countries, there seems to be no significant effect of the student-teacher gender match.

Interestingly, while we found no effect in the pooled data for girls the results are different when considering each country separately. Specifically, we find a significant positive effect of the student-teacher gender match in Malta, Singapore, South Africa, and Macedonia. Conversely, we find a negative effect in Germany, Kuwait, and the United Arab Emirates. Generally, regular matching suggested a significant effect for 19 out of the 36 countries but in the majority of the cases, the effect is not robust enough to be significant for both of the modelling approaches.

The main finding of our thesis is that there is some effect of the student-teacher gender interactions on students' performance in 4^{th} grade. In line with the existing literature (e.g. Hwang & Fitzpatrick 2021, Hermann & Diallo 2017), we show the importance of studying the effect separately for boys and girls as it may vary for the two genders. Moreover, the results are not universal across countries. Hence any policy recommendations should not be based on

pooled results or results from other countries.

This thesis contributes to the existing literature in several ways. Firstly, it brings further evidence on the effect of student-teacher gender match on educational outcomes. Secondly, it uses a novel identification strategy that allows us to estimate the effect for datasets for which the effect could not be reliably estimated using standard methodologies in the existing literature. Although there is no clear policy recommendation as a result of this work, it may serve as a starting point for further research. Whether in exploring causes other than teachers' gender of pertaining gender gaps in math results for countries where no significant effect was found or to further investigate the transmission mechanism of how teachers' gender affects students in the countries where significant effects were found for boys and girls — once a specific transmission mechanism is identified, a policy recommendation can be formulated. Lastly, an interesting follow-up study to this thesis could use the new wave of TIMSS 2023 to determine how the pandemic affected the gender interactions between students and teachers.

Bibliography

Ammermüller, A. & P. J. Dolton (2006): "Pupil-teacher gender interaction effects on scholastic outcomes in england and the usa." *ZEW Discussion Papers 06-060*, Mannheim.

Andersen, I. G. & D. Reimer (2019): "Same-gender teacher assignment, instructional strategies, and student achievement: New evidence on the mechanisms generating same-gender teacher effects." *Research in Social Stratification and Mobility* **62**: p. 100406.

Anghel, B., N. Rodríguez-Planas, & A. S. de Galdeano (2020): "Is the math gender gap associated with gender equality? only in low-income countries." *Economics of Education Review* **79**: p. 102064.

Antecol, H., O. Eren, & S. Ozbeklik (2015): "The effect of teacher gender on student achievement in primary school." *Journal of Labor Economics* **33(1)**: pp. 63–89.

Bedard, K. & I. Cho (2010): "Early gender test score gaps across OECD countries." *Economics of Education Review* **29(3)**: pp. 348–363.

Beilock, S. L., E. A. Gunderson, G. Ramirez, & S. C. Levine (2010): "Female teachers' math anxiety affects girls' math achievement." *Proceedings of the National Academy of Sciences* **107(5)**: pp. 1860–1863.

Black, D. A. & J. A. Smith (2004): "How robust is the evidence on the effects of college quality? evidence from matching." *Journal of Econometrics* **121(1-2)**: pp. 99–124.

Bouhlila, D. S. & F. Sellaouti (2013): "Multiple imputation using chained equations for missing data in TIMSS: a case study." *Large-scale Assessments in Education* **1(1)**.

Breda, T., E. Jouini, & C. Napp (2018): "Societal inequalities amplify gender gaps in math." *Science* **359(6381)**: pp. 1219–1220.

Caliendo, M. & S. Kopeinig (2005): "Some practical guidance for the implementation of propensity score matching." *DIW Discussion Papers 485*, Berlin.

Caro, D. H. & P. Biecek (2017): "intsvy: An r package for analyzing international large-scale assessment data." *Journal of Statistical Software* **81(7)**.

Carrell, S. E., M. E. Page, & J. E. West (2010): "Sex and science: How professor gender perpetuates the gender gap." *The Quarterly Journal of Economics* **125(3)**: pp. 1101–1144.

Cho, I. (2012): "The effect of teacher–student gender matching: Evidence from OECD countries." *Economics of Education Review* **31(3)**: pp. 54–67.

Cordero, J. M., V. Cristóbal, & D. Santín (2017): "Causal inference on education policies: a survey of empirical studies using PISA, TIMSS and PIRLS." *Journal of Economic Surveys* **32(3)**: pp. 878–915.

Dee, T. S. (2007): "Teachers and the gender gaps in student achievement." *The Journal of Human Resources* **42(3)**: pp. 528–554.

Dossi, G., D. Figlio, P. Giuliano, & P. Sapienza (2021): "Born in the family: Preferences for boys and the gender gap in math." *Journal of Economic Behavior Organization* **183**: pp. 175–188.

Einarsson, C. & K. Granström (2002): "Gender-biased interaction in the classroom: The influence of gender and age in the relationship between teacher and pupil." *Scandinavian Journal of Educational Research* **46(2)**: pp. 117–127.

Etaugh, C. & V. Hughes (1975): "Teachers' evaluations of sex-typed behaviors in children: The role of teacher sex and school setting." *Developmental Psychology* **11(3)**: pp. 394–395.

European Commission, 2022 (2022): "Gender pay gaps in the european union: a statistical analysis: 2022 edition." `https://ec.europa.eu/eurostat/en/web/products-statistical-working-papers/-/ks-tc-22-002`.

Fryer, R. G. & S. D. Levitt (2010): "An empirical analysis of the gender gap in mathematics." *American Economic Journal: Applied Economics* **2(2)**: pp. 210–240.

Guiso, L., F. Monte, P. Sapienza, & L. Zingales (2008): "Culture, Gender, and Math." *Science* **320**.

Hansen, B. B. & S. O. Klopfer (2006): "Optimal full matching and related designs via network flows." *Journal of Computational and Graphical Statistics* **15(3)**: pp. 609–627.

Hermann, Z. & A. Diallo (2017): "Does teacher gender matter in europe? evidence from timss data." *Budapest Working Papers on the Labour Market BWP - 2017/2*, Budapest.

Hogrebe, N. & R. Strietholt (2016): "Does non-participation in preschool affect children's reading achievement? international evidence from propensity score analyses." *Large-scale Assessments in Education* **4(1)**.

Hwang, N. & B. Fitzpatrick (2021): "Student–teacher gender matching and academic achievement." *AERA Open* **7**: p. 233285842110400.

Kim, D. H. & H. Law (2012): "Gender gap in maths test scores in south korea and hong kong: Role of family background and single-sex schooling." *International Journal of Educational Development* **32(1)**: pp. 92–103.

Krieg, J. M. (2005): "Student gender and teacher gender: What is the impact on high stakes test scores?." *Current Issues in Education* **8**.

Lavy, V. & R. Megalokonomou (2019): "Persistency in teachers' grading bias and effects on longer-term outcomes: University admissions exams and choice of field of study." *Technical report.*

Meece, J. (1987): "The influence of school experiences on the development of gender schemata." *In L. S. Liben M. L. Signorella (Eds.), New directions for child development, No. 38: Children's gender schemata (pp. 57â€“73). Jossey-Bass.* .

Meinck, S. & F. Brese (2019): "Trends in gender gaps: using 20 years of evidence from TIMSS." *Large-scale Assessments in Education* **7(1)**.

Mullis, I. V. S., M. O. Martin, P. Foy, D. L. Kelly, & B. Fishbein (2020): "Timss 2019 international results in mathematics and science." https://timssandpirls.bc.edu/timss2019/international-results/.

Nye, C. D., S. M. Butt, J. Bradburn, & J. Prasad (2018): "Interests as predictors of performance: An omitted and underappreciated variable." *Journal of Vocational Behavior* **108**: pp. 178–189.

OECD, 2017 (2017): "The under-representation of women in STEM fields." In "The Pursuit of Gender Equality," pp. 105–112. OECD.

Parker-Price, S. & A. F. Claxton (1996): "Teachers' perceptions of gender differences in students." .

Pope, D. G. & J. R. Sydnor (2010): "Geographic variation in the gender differences in test scores." *Journal of Economic Perspectives* **24(2)**: pp. 95–108.

Schafer, J. L. & J. Kang (2008): "Average causal effects from nonrandomized studies: A practical guide and simulated example." *Psychological Methods* **13(4)**: pp. 279–313.

T19 UG Supplement1 (2019): "Timss 2019 user guide for the international database. supplement 1." `https://timss2019.org/international-database/downloads/T19_UG_Supp1-international-context-questionnaires.pdf`.

T19 UG Supplement3 (2019): "Timss 2019 user guide for the international database. supplement 3." `https://timss2019.org/international-database/downloads/T19_UG_Supp3-derived-context-variables.pdf`.

Terrier, C. (2020): "Boys lag behind: How teachers' gender biases affect student achievement." *Economics of Education Review* **77**: p. 101981.

TIMSS 2019 Context Questionnaires (2019): "Timss 2019 context questionnaires." 2021-12-18.

UN SDGs, 2022 (2022): "Un: The 17 sustainable development goals." `https://sdgs.un.org/goals`.

Winters, M. A., R. C. Haight, T. T. Swaim, & K. A. Pickering (2013): "The effect of same-gender teacher assignment on student achievement in the elementary and secondary grades: Evidence from panel data." *Economics of Education Review* **34**: pp. 69–75.

Zhao, Q.-Y., J.-C. Luo, Y. Su, Y.-J. Zhang, G.-W. Tu, & Z. Luo (2021): "Propensity score matching with r: conventional methods and new features." *Annals of Translational Medicine* **9(9)**: pp. 812–812.

Appendix A

Appendix A

Appendix A shows the number of girls and boys taught by a male/female teacher for the 36 countries in scope. For all student-teacher gender combinations, average achievement in math standardized tests is reported.

		Teacher gender					
		N (of students)		Mean (SE)		SD (SE)	
Country	Student sex	Female	Male	Female	Male	Female	Male
Albania	Girls	2086	158	496.7 (3.99)	467.21 (9.76)	83.75 (2.69)	88.59 (9.97)
	Boys	2178	184	499.32 (4.17)	482.98 (13.77)	87.35 (3.27)	79.47 (11.14)
Bahrain	Girls	5344	201	481.71 (3.31)	502.35 (10.58)	85.49 (2.26)	91.03 (6.24)
	Boys	3261	2571	475.58 (3.84)	481.46 (5.3)	87.04 (2.73)	89.11 (2.67)
Bosnia and Herzegovina	Girls	2528	200	449.01 (2.68)	428.8 (10.42)	72.74 (1.48)	75.31 (5.26)
	Boys	2647	208	456.44 (2.97)	450.85 (7.4)	77.09 (1.73)	70.47 (4.5)
Canada	Girls	6104	1021	507.15 (4.22)	495.84 (4.88)	75.22 (2.16)	71.54 (2.9)
	Boys	6206	1115	527.64 (2.59)	509.93 (5.26)	76.59 (1.87)	74.94 (2.48)
Cyprus	Girls	2669	942	525.48 (3.65)	523.73 (6.07)	75.64 (2.27)	77.33 (2.63)
	Boys	2423	893	545.98 (3.66)	537.79 (6.32)	78.29 (1.94)	82.62 (4.06)
Czech Republic	Girls	2667	204	527.78 (3.23)	527.62 (10.29)	73.43 (2.36)	77.28 (6.71)
	Boys	2804	209	537.77 (3.31)	552.43 (9.44)	75.59 (2.56)	73.91 (4.15)
	Girls	1246	980	522.63 (2.73)	517.09 (3.81)	68.6 (1.97)	71.6 (2.5)

(continued)

		Teacher gender					
		N (of students)		Mean (SE)		SD (SE)	
Country	Student sex	Female	Male	Female	Male	Female	Male
Denmark	Boys	1241	944	529.65 (3.33)	528.75 (4.49)	75.04 (1.88)	76.88 (2.28)
Finland	Girls	1881	675	528.92 (3.38)	535.42 (4.58)	75.18 (1.66)	73.91 (2.98)
	Boys	1960	743	531.8 (3.45)	536.49 (4.27)	78.1 (2.2)	75.56 (3.28)
France	Girls	1988	448	478.37 (3.87)	473.29 (8.48)	77.86 (2.62)	80.48 (3.55)
	Boys	2054	438	493.13 (3.84)	481.48 (10.13)	80.04 (2.35)	87.47 (5.92)
Germany	Girls	1968	233	513.06 (3.25)	523.5 (7.3)	69.26 (1.99)	66.3 (5.63)
	Boys	1987	264	525.8 (2.91)	531.18 (6.7)	72.49 (2.3)	73.04 (5.79)
Hong Kong SAR	Girls	1528	990	602.21 (5.37)	597.17 (5.35)	67.7 (2.6)	66.74 (2.7)
	Boys	1607	1279	610.87 (6.39)	602.21 (5.06)	69.88 (3.12)	71.21 (2.25)
Chile	Girls	2459	555	442.01 (4.02)	424.71 (5.63)	74 (2.27)	66.88 (2.96)
	Boys	2410	580	450.01 (3.58)	443.54 (6.87)	77.28 (2.31)	71.94 (3.42)
Chinese Taipei	Girls	2639	940	598.02 (2.37)	593.47 (4.68)	61 (1.42)	67.09 (4.22)
	Boys	2854	1019	602.58 (2.55)	594.99 (3.6)	68.68 (1.91)	69.4 (2.18)

(continued)

		Teacher gender					
		N (of students)		Mean (SE)		SD (SE)	
Country	Student sex	Female	Male	Female	Male	Female	Male
Iran, Islamic Rep. of	Girls	2860	124	443.88 (6.35)	372.06 (24.15)	91.16 (3.1)	100.68 (12.97)
	Boys	1828	1198	453.12 (6.64)	437.69 (9.32)	93.85 (2.64)	100.96 (4.59)
Ireland	Girls	1840	471	541.89 (3.42)	556.16 (5.94)	74.24 (1.71)	75.78 (3.68)
	Boys	1654	576	550.21 (3.43)	555.15 (5.58)	76.73 (2.27)	75.8 (2.48)
Korea,Rep.of	Girls	1974	389	598.81 (2.43)	585.84 (6.18)	66.99 (1.94)	74.6 (4.06)
	Boys	2131	450	604.51 (3.28)	591.92 (5.66)	73.59 (1.97)	71.75 (3.9)
Kosovo	Girls	1717	467	445.67 (3.5)	429.86 (6.08)	79.19 (1.91)	76.69 (3.52)
	Boys	1780	497	448.32 (3.88)	440.62 (7.36)	81.72 (2.34)	81.18 (3.86)
Kuwait	Girls	4134	93	385.25 (6.3)	431.81 (45.2)	105.7 (2.69)	100.9 (17.37)
	Boys	3387	1079	387.83 (7.8)	356.14 (15.91)	113.12 (3.73)	122.42 (8.12)
Macedonia, Rep. of	Girls	1402	177	474.55 (6.7)	458.33 (13.98)	100.83 (3.73)	92.02 (8.18)
	Boys	1550	183	474.49 (6.23)	456.89 (13.94)	96.49 (3.35)	95.64 (5.34)
	Girls	1804	284	510.16 (2.14)	489.46 (4.3)	72.37 (1.74)	69.87 (3.11)

(continued)

		Teacher gender					
		N (of students)		Mean (SE)		SD (SE)	
Country	Student sex	Female	Male	Female	Male	Female	Male
Malta	Boys	1926	300	515.04 (1.94)	498.93 (4.88)	76.14 (1.43)	85.25 (3.47)
Montenegro	Girls	2086	224	450.43 (3.07)	446.04 (10.36)	83.35 (2.12)	89.78 (7.87)
	Boys	2397	248	453.91 (2.68)	469 (8.99)	86.14 (2.01)	92.06 (4.64)
Morocco	Girls	3503	2813	391.89 (5.9)	364.12 (6.37)	102.29 (4.14)	91.77 (3.73)
	Boys	3704	3090	387.29 (5.94)	363.72 (6.06)	105.36 (4.05)	94.74 (3.59)
New Zealand	Girls	2221	594	483.97 (4.32)	488.8 (4.99)	86.9 (2.19)	78.62 (3.77)
	Boys	2336	684	491.04 (4.39)	504.38 (8.1)	93.73 (2.13)	92.54 (3.45)
Pakistan	Girls	2214	525	345.54 (21.71)	329.61 (28.53)	101.41 (5.23)	103.38 (7.61)
	Boys	1012	2560	331.13 (19.46)	310.87 (14.34)	98.66 (5.17)	108.81 (5.68)
Philippines	Girls	3873	669	314.82 (7.53)	328.45 (11.53)	105.48 (2.7)	108.07 (4.22)
	Boys	4121	724	280.89 (7.43)	292.48 (12.82)	110.98 (2.93)	114.26 (5.75)
Poland	Girls	4309	357	516.78 (3.01)	517.15 (7.67)	73.41 (1.87)	74.4 (5.19)
	Boys	4534	357	524.67 (2.99)	523.81 (7.12)	79.72 (1.94)	79.77 (4.57)

(continued)

		Teacher gender					
		N (of students)		Mean (SE)		SD (SE)	
Country	Student sex	Female	Male	Female	Male	Female	Male
Portugal	Girls	1819	238	517.37 (3.25)	512.56 (6.28)	73.93 (1.52)	73 (4.42)
	Boys	1899	277	533.27 (3.19)	535.71 (8.09)	76.69 (1.87)	75.53 (5.41)
Qatar	Girls	4077	433	439.12 (5.07)	489.65 (13.21)	83.8 (2.34)	87.2 (6.32)
	Boys	3191	1120	440.08 (3.55)	458.1 (10.07)	92.57 (2.95)	95.32 (4.78)
Serbia	Girls	1941	214	509.87 (3.7)	497.09 (9.73)	79.61 (2.44)	75.38 (5.23)
	Boys	1947	256	506.46 (4.26)	509.32 (12.62)	89.56 (3.19)	89.23 (5.15)
Singapore	Girls	4035	1133	627.78 (4.29)	601.04 (5.89)	75.22 (2.08)	76.68 (4.2)
	Boys	3756	1411	633.06 (4.73)	620.64 (6.24)	82.57 (3.42)	80.71 (4.26)
Slovak Republic	Girls	2655	203	503.79 (3.6)	497.94 (15.74)	75.06 (3.23)	76.91 (4.91)
	Boys	2740	191	516.61 (4.25)	517.57 (14.38)	79.35 (2.9)	80.64 (8.7)
South Africa	Girls	6893	4677	392.58 (5.03)	365.96 (4.91)	102.59 (2.66)	90.91 (2.7)
	Boys	6900	4918	370.26 (4.78)	349.8 (5.28)	102.03 (2.39)	94.51 (3.62)
	Girls	4907	1703	494.6 (2.48)	496.44 (4.7)	69.56 (1.86)	67.4 (2.44)

(continued)

		Teacher gender					
		N (of students)		Mean (SE)		SD (SE)	
Country	Student sex	Female	Male	Female	Male	Female	Male
Spain	Boys	5074	1973	509.02 (3.13)	511.29 (5.62)	74.41 (2.13)	74.33 (3.45)
Sweden	Girls	1761	624	516.69 (3.99)	517.9 (5.04)	73.54 (2.01)	67.92 (3.07)
	Boys	1794	629	524.92 (3.58)	531.16 (5.09)	75.31 (2.7)	72.9 (2.99)
Turkey	Girls	2626	1554	526.27 (5.59)	514.42 (6.47)	96.15 (3.09)	92.94 (2.99)
	Boys	2334	1489	526.13 (6.26)	524 (7.45)	103.06 (3.04)	104.09 (3.44)
United Arab Emirates	Girls	18107	1620	473.45 (3.28)	499.29 (5.72)	94.36 (1.37)	91.69 (3.3)
	Boys	16234	3307	485.13 (2.89)	468.52 (6.23)	99.81 (1.56)	105.47 (3.66)

Appendix B

Appendix B

In Appendix B the summary statistics for matching variables for individual countries are shown.

B.1 Early numeracy activities before school

Early Numeracy Activities Before School (ASBHENA) derived from Home Questionnaire completed by parents. This variable sums up information from a series of questions that ask about pre-school activities. For example: "Before your child began primary/elementary school, how often did you or someone else in your home do the following activities with him or her? Count different things" The answers are mapped into a scale where a higher number means better skills.(T19 UG Supplement1, T19 UG Supplement3)

Country	Gender	Freq	Mean	s.e.
Albania	Girls	2186	11.03	0.09
	Boys	2277	10.90	0.10
Bahrain	Girls	5279	10.65	0.05
	Boys	5260	10.45	0.05
Bosnia and Herzegovina	Girls	2657	10.78	0.05
	Boys	2776	10.70	0.05
Canada	Girls	5101	11.17	0.05
	Boys	5107	10.97	0.06
	Girls	3464	10.78	0.05

Cyprus	Boys	3143	10.68	0.06
Czech Republic	Girls	2399	10.89	0.05
	Boys	2522	10.68	0.07
Denmark	Girls	1353	10.19	0.08
	Boys	1336	10.06	0.06
Finland	Girls	2325	9.94	0.04
	Boys	2352	9.85	0.06
France	Girls	2301	10.59	0.05
	Boys	2270	10.50	0.05
Germany	Girls	1479	10.55	0.07
	Boys	1480	10.28	0.08
Hong Kong SAR	Girls	2367	9.62	0.07
	Boys	2679	9.64	0.10
Chile	Girls	2801	10.07	0.07
	Boys	2754	10.13	0.05
Chinese Taipei	Girls	3520	9.75	0.05
	Boys	3767	9.68	0.06
Iran, Islamic Rep. of	Girls	2935	9.64	0.09
	Boys	2949	9.40	0.11
Ireland	Girls	2194	11.24	0.07
	Boys	2078	11.08	0.06
Korea,Rep.of	Girls	2339	10.67	0.06
	Boys	2532	10.72	0.07
Kosovo	Girls	2110	10.75	0.06
	Boys	2186	10.61	0.06
Kuwait	Girls	3638	10.42	0.06
	Boys	3345	10.13	0.06

Macedonia, Rep. of	Girls	1416	11.05	0.09
	Boys	1533	10.88	0.09
Malta	Girls	1524	11.27	0.05
	Boys	1609	11.28	0.06
Montenegro	Girls	2257	10.96	0.05
	Boys	2545	10.76	0.04
Morocco	Girls	6168	8.19	0.13
	Boys	6606	8.10	0.13
New Zealand	Girls	1183	11.55	0.10
	Boys	1248	11.12	0.09
Pakistan	Girls	2323	9.08	0.25
	Boys	2924	8.60	0.25
Philippines	Girls	4350	10.06	0.06
	Boys	4663	9.89	0.06
Poland	Girls	4478	11.25	0.04
	Boys	4609	11.12	0.05
Portugal	Girls	1983	10.36	0.05
	Boys	2055	10.35	0.06
Qatar	Girls	3936	10.38	0.06
	Boys	3385	10.01	0.07
Serbia	Girls	2108	11.08	0.07
	Boys	2151	10.97	0.06
Singapore	Girls	5071	10.22	0.05
	Boys	5000	10.15	0.04
Slovak Republic	Girls	2754	11.08	0.08
	Boys	2793	10.85	0.08
South Africa	Girls	10176	9.90	0.05
	Boys	9786	9.72	0.06

Spain	Girls	6041	10.41	0.06
	Boys	6252	10.20	0.04
Sweden	Girls	2024	9.82	0.06
	Boys	1921	9.55	0.05
Turkey	Girls	3946	9.33	0.14
	Boys	3539	9.21	0.15
United Arab Emirates	Girls	10689	10.96	0.04
	Boys	9758	10.85	0.05

B.2 Early numeracy tasks starting school

Early Numeracy Tasks Beginning School (ASBHENT) derived from Home Questionnaire completed by parents. This variable intends to quantify information at the start of the school by compounding several questions like: "Could your child do the following when he/she began the <first grade> of primary/elementary school? Count by himself/herself" This variable is also a scale where a higher number means better skills. (T19 UG Supplement1, T19 UG Supplement3)

Country	Gender	Freq	Mean	s.e.
Albania	Girls	2203	10.65	0.08
	Boys	2280	10.67	0.07
Bahrain	Girls	5295	10.87	0.04
	Boys	5288	10.84	0.06
Bosnia and Herzegovina	Girls	2650	9.60	0.05
	Boys	2774	9.77	0.05
Canada	Girls	5093	10.11	0.07
	Boys	5102	10.34	0.06
Cyprus	Girls	3475	9.83	0.05
	Boys	3154	10.19	0.05
Czech Republic	Girls	2396	9.34	0.05
	Boys	2517	9.64	0.05
Denmark	Girls	1340	9.21	0.05
	Boys	1339	9.46	0.07
Finland	Girls	2328	10.30	0.05
	Boys	2350	10.54	0.05
France	Girls	2283	9.20	0.05
	Boys	2264	9.35	0.05
Germany	Girls	1481	9.29	0.07
	Boys	1473	9.40	0.07

Hong Kong SAR	Girls	2363	11.33	0.08
	Boys	2662	11.39	0.06
Chile	Girls	2804	9.64	0.06
	Boys	2758	9.77	0.05
Chinese Taipei	Girls	3560	11.60	0.05
	Boys	3826	11.69	0.05
Iran, Islamic Rep. of	Girls	2950	9.36	0.08
	Boys	2981	9.35	0.08
Ireland	Girls	2191	11.08	0.05
	Boys	2066	11.00	0.06
Korea,Rep.of	Girls	2342	11.48	0.06
	Boys	2532	11.45	0.06
Kosovo	Girls	2139	10.33	0.06
	Boys	2227	10.52	0.06
Kuwait	Girls	3618	9.84	0.06
	Boys	3349	9.91	0.06
Macedonia, Rep. of	Girls	1413	10.27	0.07
	Boys	1532	10.12	0.08
Malta	Girls	1521	9.30	0.05
	Boys	1610	9.47	0.06
Montenegro	Girls	2256	9.61	0.04
	Boys	2542	9.64	0.04
Morocco	Girls	6189	9.06	0.09
	Boys	6617	9.03	0.07
New Zealand	Girls	1185	8.84	0.06
	Boys	1252	8.82	0.07
	Girls	2334	8.75	0.42

Pakistan	Boys	2955	8.46	0.22
Philippines	Girls	4352	10.44	0.08
	Boys	4668	10.25	0.07
Poland	Girls	4481	10.29	0.05
	Boys	4603	10.60	0.04
Portugal	Girls	1980	9.42	0.04
	Boys	2060	9.58	0.05
Qatar	Girls	3963	10.30	0.08
	Boys	3415	10.37	0.08
Serbia	Girls	2113	9.96	0.06
	Boys	2147	10.10	0.06
Singapore	Girls	5068	11.16	0.04
	Boys	4999	11.27	0.05
Slovak Republic	Girls	2750	8.81	0.07
	Boys	2797	9.08	0.08
South Africa	Girls	10280	9.94	0.04
	Boys	9898	9.76	0.04
Spain	Girls	6039	10.32	0.04
	Boys	6228	10.37	0.06
Sweden	Girls	2030	10.42	0.06
	Boys	1923	10.55	0.06
Turkey	Girls	3928	9.58	0.11
	Boys	3545	9.45	0.11
United Arab Emirates	Girls	10637	11.11	0.03
	Boys	9725	11.14	0.04

B.3 Pre-primary education

Student Attended Preprimary Education (ASDHAPS) is also derived from Home Questionnaire but intends to provide us with details about students' pre-school education (kindergarten etc.). It is a factor variable with four levels: 0: Did Not Attend; 1: 1 Year or Less; 2: 2 Years; 3: 3 Years or More. (T19 UG Supplement1, T19 UG Supplement3)

Country	Gender	Freq	Mean	s.e.
Albania	Girls	2090	2.34	0.04
	Boys	2169	2.33	0.03
Bahrain	Girls	5093	2.00	0.03
	Boys	5061	1.95	0.04
Bosnia and Herzegovina	Girls	2501	1.34	0.04
	Boys	2626	1.34	0.04
Canada	Girls	4356	1.96	0.04
	Boys	4436	2.06	0.07
Cyprus	Girls	3369	2.27	0.02
	Boys	3055	2.25	0.03
Czech Republic	Girls	2385	2.77	0.02
	Boys	2495	2.77	0.02
Denmark	Girls	1350	2.93	0.01
	Boys	1336	2.94	0.01
Finland	Girls	2304	2.58	0.02
	Boys	2341	2.59	0.02
France	Girls	2225	2.77	0.02
	Boys	2224	2.81	0.02
Germany	Girls	1438	2.23	0.04
	Boys	1453	2.29	0.04
	Girls	2304	2.35	0.03

Hong Kong SAR	Boys	2621	2.36	0.03
Chile	Girls	2688	2.32	0.03
	Boys	2673	2.37	0.02
Chinese Taipei	Girls	3544	2.52	0.02
	Boys	3809	2.54	0.02
Iran, Islamic Rep. of	Girls	2748	1.30	0.05
	Boys	2752	1.20	0.05
Ireland	Girls	2117	2.35	0.02
	Boys	2009	2.37	0.02
Korea,Rep.of	Girls	2319	2.89	0.01
	Boys	2509	2.86	0.01
Kosovo	Girls	1912	0.97	0.04
	Boys	1991	0.98	0.04
Kuwait	Girls	3391	1.89	0.04
	Boys	3084	1.95	0.04
Macedonia, Rep. of	Girls	1343	1.46	0.07
	Boys	1474	1.35	0.07
Malta	Girls	1478	2.19	0.02
	Boys	1574	2.26	0.02
Montenegro	Girls	2052	1.86	0.03
	Boys	2328	1.86	0.03
Morocco	Girls	5660	1.57	0.05
	Boys	5999	1.58	0.04
New Zealand	Girls	1182	2.47	0.03
	Boys	1249	2.44	0.03
Pakistan	Girls	1776	0.96	0.20
	Boys	2304	0.79	0.16
	Girls	3994	1.93	0.03

Philippines	Boys	4235	1.90	0.03
Poland	Girls	4371	2.55	0.03
	Boys	4459	2.54	0.03
Portugal	Girls	1915	2.74	0.02
	Boys	2010	2.70	0.02
Qatar	Girls	3657	1.68	0.03
	Boys	3133	1.66	0.04
Serbia	Girls	2078	2.23	0.04
	Boys	2108	2.20	0.03
Singapore	Girls	4986	2.73	0.01
	Boys	4930	2.75	0.01
Slovak Republic	Girls	2732	2.59	0.04
	Boys	2776	2.58	0.04
South Africa	Girls	8894	2.10	0.03
	Boys	8584	2.12	0.03
Spain	Girls	5795	2.37	0.03
	Boys	6056	2.41	0.02
Sweden	Girls	1995	2.82	0.02
	Boys	1898	2.81	0.03
Turkey	Girls	3854	1.01	0.03
	Boys	3453	1.08	0.04
United Arab Emirates	Girls	9951	1.54	0.02
	Boys	9119	1.53	0.02

B.4 Parents' education

Parents' Highest Education Level (ASDHEDUP) is a factor variable with 6 levels: 1: University or Higher; 2: Post-secondary but not University; 3: Upper Secondary; 4: Lower Secondary; 5: Some Primary, Lower Secondary or No School; 6: Not Applicable. (T19 UG Supplement1, T19 UG Supplement3)

Country	Gender	Freq	Mean	s.e.
Albania	Girls	1999	3.15	0.05
	Boys	2105	3.12	0.05
Bahrain	Girls	5007	1.97	0.04
	Boys	5051	1.95	0.04
Bosnia and Herzegovina	Girls	2584	2.56	0.04
	Boys	2708	2.57	0.03
Canada	Girls	5091	1.63	0.03
	Boys	5093	1.56	0.03
Cyprus	Girls	3316	1.83	0.04
	Boys	3005	1.77	0.04
Czech Republic	Girls	2386	2.18	0.03
	Boys	2506	2.21	0.03
Denmark	Girls	1329	1.42	0.04
	Boys	1323	1.44	0.03
Finland	Girls	2313	1.74	0.04
	Boys	2357	1.69	0.03
France	Girls	2264	2.15	0.05
	Boys	2241	2.11	0.04
Germany	Girls	1469	2.24	0.04
	Boys	1449	2.16	0.04
Hong Kong SAR	Girls	2368	2.28	0.07
	Boys	2666	2.17	0.06

Chile	Girls	2799	2.46	0.04
	Boys	2753	2.39	0.04
Chinese Taipei	Girls	3545	1.91	0.03
	Boys	3820	1.88	0.03
Iran, Islamic Rep. of	Girls	2923	2.95	0.09
	Boys	2929	2.95	0.08
Ireland	Girls	2193	1.87	0.04
	Boys	2067	1.82	0.04
Korea,Rep.of	Girls	2340	1.63	0.04
	Boys	2530	1.62	0.03
Kosovo	Girls	2093	2.86	0.04
	Boys	2177	2.84	0.04
Kuwait	Girls	3400	1.78	0.04
	Boys	3110	1.75	0.05
Macedonia, Rep. of	Girls	1337	2.68	0.08
	Boys	1462	2.76	0.08
Malta	Girls	1505	2.39	0.04
	Boys	1592	2.43	0.04
Montenegro	Girls	1993	2.06	0.03
	Boys	2225	2.05	0.03
Morocco	Girls	5527	4.25	0.06
	Boys	5794	4.28	0.05
New Zealand	Girls	1178	1.75	0.05
	Boys	1249	1.72	0.05
Pakistan	Girls	2364	3.43	0.20
	Boys	2952	3.69	0.09
Philippines	Girls	4122	3.00	0.06
	Boys	4505	3.00	0.06

Poland	Girls	4428	1.97	0.03
	Boys	4526	1.90	0.04
Portugal	Girls	1947	2.43	0.05
	Boys	2027	2.46	0.05
Qatar	Girls	3753	1.64	0.04
	Boys	3302	1.58	0.05
Serbia	Girls	2076	2.44	0.04
	Boys	2115	2.48	0.04
Singapore	Girls	4993	1.68	0.02
	Boys	4918	1.66	0.03
Slovak Republic	Girls	2743	2.23	0.05
	Boys	2792	2.25	0.05
South Africa	Girls	8699	3.06	0.04
	Boys	8297	3.07	0.04
Spain	Girls	5967	2.33	0.06
	Boys	6198	2.24	0.06
Sweden	Girls	1908	1.78	0.06
	Boys	1811	1.72	0.04
Turkey	Girls	3948	3.47	0.06
	Boys	3543	3.45	0.07
United Arab Emirates	Girls	10544	1.60	0.03
	Boys	9570	1.64	0.03

B.5 Parents' occupation

Parents' Highest Occupation Level (ASDHOCCP) is a factor variable with 7 levels: 1: Professional; 2: Small Business Owner; 3: Clerical; 4: Skilled Worker; 5: General Laborer; 6: Never Worked for Pay; 7: Not Applicable. (T19 UG Supplement1, T19 UG Supplement3)

Country	Gender	Freq	Mean	s.e.
Albania	Girls	1922	3.42	0.08
	Boys	1993	3.35	0.08
Bahrain	Girls	4520	2.70	0.06
	Boys	4510	2.65	0.06
Bosnia and Herzegovina	Girls	2394	2.78	0.05
	Boys	2503	2.86	0.05
Canada	Girls	5063	1.88	0.04
	Boys	5079	1.81	0.08
Cyprus	Girls	3341	2.31	0.05
	Boys	2988	2.28	0.04
Czech Republic	Girls	2377	2.19	0.05
	Boys	2493	2.24	0.05
Denmark	Girls	1320	1.72	0.05
	Boys	1325	1.71	0.05
Finland	Girls	2297	1.89	0.05
	Boys	2343	1.88	0.04
France	Girls	2217	2.38	0.05
	Boys	2195	2.36	0.07
Germany	Girls	1442	2.32	0.05
	Boys	1434	2.20	0.05
Hong Kong SAR	Girls	2313	2.39	0.08
	Boys	2617	2.30	0.08

Chile	Girls	2658	3.13	0.07
	Boys	2613	3.06	0.07
Chinese Taipei	Girls	3504	2.16	0.05
	Boys	3788	2.17	0.04
Iran, Islamic Rep. of	Girls	2758	3.50	0.08
	Boys	2699	3.46	0.07
Ireland	Girls	2128	2.04	0.05
	Boys	2001	2.08	0.05
Korea,Rep.of	Girls	2332	2.19	0.04
	Boys	2520	2.15	0.05
Kosovo	Girls	2026	3.18	0.06
	Boys	2101	3.14	0.06
Kuwait	Girls	3170	2.39	0.08
	Boys	2865	2.39	0.10
Macedonia, Rep. of	Girls	1280	3.10	0.09
	Boys	1375	3.23	0.09
Malta	Girls	1495	2.18	0.05
	Boys	1586	2.21	0.05
Montenegro	Girls	1715	2.93	0.04
	Boys	1918	2.87	0.04
Morocco	Girls	5232	4.01	0.09
	Boys	5578	4.09	0.07
New Zealand	Girls	1178	1.74	0.05
	Boys	1245	1.77	0.06
Pakistan	Girls	2334	3.06	0.22
	Boys	2928	3.11	0.13
Philippines	Girls	3843	3.39	0.06
	Boys	4171	3.42	0.06

Poland	Girls	4398	2.37	0.04
	Boys	4487	2.28	0.05
Portugal	Girls	1868	2.47	0.05
	Boys	1955	2.46	0.06
Qatar	Girls	3448	2.23	0.06
	Boys	3071	2.16	0.07
Serbia	Girls	2001	2.66	0.05
	Boys	2022	2.67	0.05
Singapore	Girls	4964	1.67	0.03
	Boys	4885	1.69	0.04
Slovak Republic	Girls	2672	2.69	0.07
	Boys	2734	2.60	0.09
South Africa	Girls	7757	3.39	0.05
	Boys	7147	3.49	0.06
Spain	Girls	5570	2.26	0.04
	Boys	5858	2.24	0.05
Sweden	Girls	1962	1.85	0.08
	Boys	1865	1.75	0.06
Turkey	Girls	3940	3.71	0.05
	Boys	3551	3.57	0.05
United Arab Emirates	Girls	10366	2.17	0.05
	Boys	9460	2.16	0.04

B.6 Student's age starting school

Students' Age when Starting School (ASBH05) is a factor variable with four levels: 1: 5 years old or younger; 2: 6 years old; 3: 7 years old; 4: 8 years old or older. (T19 UG Supplement1)

Country	Gender	Freq	Mean	s.e.
Albania	Girls	2192	2.25	0.02
	Boys	2274	2.29	0.01
Bahrain	Girls	5283	2.00	0.02
	Boys	5278	1.98	0.02
Bosnia and Herzegovina	Girls	2665	2.10	0.01
	Boys	2783	2.13	0.01
Canada	Girls	5115	1.71	0.01
	Boys	5128	1.71	0.01
Cyprus	Girls	3469	1.92	0.01
	Boys	3146	1.97	0.01
Czech Republic	Girls	2407	2.20	0.01
	Boys	2527	2.34	0.02
Denmark	Girls	1351	1.70	0.02
	Boys	1334	1.75	0.02
Finland	Girls	2331	2.72	0.02
	Boys	2360	2.70	0.02
France	Girls	2294	1.82	0.01
	Boys	2255	1.84	0.01
Germany	Girls	1484	2.09	0.02
	Boys	1478	2.14	0.02
Hong Kong SAR	Girls	2359	2.00	0.01
	Boys	2660	1.99	0.02
	Girls	2808	1.90	0.01

Chile	Boys	2760	1.91	0.01
Chinese Taipei	Girls	3552	2.90	0.02
	Boys	3819	2.86	0.02
Iran, Islamic Rep. of	Girls	2946	2.75	0.02
	Boys	2974	2.69	0.02
Ireland	Girls	2190	2.22	0.02
	Boys	2062	2.21	0.02
Korea,Rep.of	Girls	2338	2.85	0.02
	Boys	2521	2.85	0.01
Kosovo	Girls	2146	2.03	0.01
	Boys	2228	2.04	0.01
Kuwait	Girls	3656	1.84	0.02
	Boys	3360	1.84	0.02
Macedonia, Rep. of	Girls	1428	1.93	0.02
	Boys	1538	1.94	0.01
Malta	Girls	1524	1.12	0.01
	Boys	1615	1.11	0.01
Montenegro	Girls	2260	1.92	0.01
	Boys	2548	1.94	0.01
Morocco	Girls	6132	2.01	0.01
	Boys	6531	2.02	0.01
New Zealand	Girls	1189	1.05	0.01
	Boys	1255	1.05	0.01
Pakistan	Girls	2328	1.34	0.06
	Boys	2892	1.36	0.06
Philippines	Girls	4393	1.91	0.02
	Boys	4717	1.95	0.02

Poland	Girls	4486	2.36	0.02
	Boys	4593	2.38	0.02
Portugal	Girls	1984	1.84	0.01
	Boys	2062	1.84	0.01
Qatar	Girls	3957	1.80	0.03
	Boys	3417	1.74	0.03
Serbia	Girls	2104	2.71	0.01
	Boys	2149	2.70	0.01
Singapore	Girls	5064	2.64	0.01
	Boys	4990	2.64	0.01
Slovak Republic	Girls	2757	2.23	0.01
	Boys	2806	2.29	0.02
South Africa	Girls	10347	2.10	0.02
	Boys	9977	2.09	0.02
Spain	Girls	6031	1.59	0.01
	Boys	6205	1.61	0.01
Sweden	Girls	2024	2.68	0.02
	Boys	1926	2.65	0.03
Turkey	Girls	3954	2.24	0.02
	Boys	3565	2.27	0.02
United Arab Emirates	Girls	10684	1.73	0.01
	Boys	9800	1.70	0.01

Appendix C

Appendix C

In Appendix D we report the complete list of variables used in the OLS model Extra which is a part of the robustness check.

- ASBH02A — Was your child born in <country>?
- ASBH03A — What language did your child speak before he/she began school? <language of test>
- ASBH09E — What do you think of your child's school? My child's school promotes high academic standards
- ASBH09G — What do you think of your child's school? My child's school does a good job in helping him/her become better in mathematics
- ASBH12A — Were the child's <parents/guardians> born in <country>? <Parent/Guardian A>
- ASBH12B — Were the child's <parents/guardians> born in <country>? <Parent/Guardian B>
- ASBH16 — How far in his/her education do you expect your child to go?
- ASDG05S — Number of Home Study Supports (derived)
- ASBG08 — About how often are you absent from school?
- ASBG09A — How often do you feel this way when you arrive at school? I feel tired
- ASBG09B — How often do you feel this way when you arrive at school? I feel hungry

- ASBG10A — What do you think about your school? Tell how much you agree with these statements. I like being in school
- ASBM01 — In mathematics lessons, how often do you work problems on your own?
- ASBM02A — How much do you agree with these statements about learning mathematics? I enjoy learning mathematics
- ASBM02E — How much do you agree with these statements about learning mathematics? I like mathematics
- ASBM03D — How much do you agree with these statements about your mathematics lessons? My teacher is good at explaining mathematics
- ASBM05D — How much do you agree with these statements about mathematics? I learn things quickly in mathematics
- ATBG01 — By the end of this school year, how many years will you have been teaching altogether?
- ATDMNUM — Percent of Students Taught Number Topics (derived)
- ATDMGEO — Percent of Students Taught Measurement and Geometry Topics (derived)
- ATDMDAT — Percent of Students Taught Data Topics (derived)
- ATBG06C — How would you characterize each of the following within your school? Teachers' expectations for student achievement
- ATBG06D — How would you characterize each of the following within your school? Teachers' ability to inspire students
- ATBG06H — How would you characterize each of the following within your school? Parental support for student achievement
- ATBG06K — How would you characterize each of the following within your school? Students' respect for classmates who excel academically
- ATBG08C — How often do you feel the following way about being a teacher? I am enthusiastic about my job

- ATBG09A — Indicate the extent to which you agree or disagree with each of the following statements. There are too many students in the classes
- ATBG09D — Indicate the extent to which you agree or disagree with each of the following statements. I need more time to prepare for class
- ATBG10A — How many students are in this class?
- ATBG12G — How often do you do the following in teaching this class? Ask students to decide their own problem solving procedures
- ATBG13E — In your view, to what extent do the following limit how you teach this class? Disruptive students
- ATBM01 — In a typical week, how much time do you spend teaching mathematics to the students in this class? (minutes)
- ATBM06A — How often do you usually assign mathematics homework to the students in this class?
- ATBM09AA — In the past two years, have you participated in professional development in any of the following? Mathematics content
- ATBM09AB — In the past two years, have you participated in professional development in any of the following? Mathematics pedagogy/instruction
- ACDGTIHY — Total Instructional Hours per Year (derived)
- ACBG13AA — How much is your school's capacity to provide instruction affected by a shortage or inadequacy of the following? General School Resources: Instructional materials
- ACBG13BA — How much is your school's capacity to provide instruction affected by a shortage or inadequacy of the following? Resources for Mathematics Instruction: Teachers with a specialization in mathematics
- ACBG16B — To what degree is each of the following a problem among teachers in your school? Absenteeism
- ACBGLNS — Students Enter with Literacy and Numeracy Skills/SCL
- ACBGMRS — Instruction Affected by Math Resource Shortage/SCL