Gender Agenda in Classroom and Student Educational Outcomes

Inaugural dissertation to obtain a doctoral degree from the

University of Bamberg, Graduate School of Social Sciences

Zahra Kamal, M. Sc.

Bamberg, 2022

This manuscript has been submitted to the Otto-Friedrich University of Bamberg, Faculty of Social Sciences, Economics, and Business Administration as a dissertation for doctoral degree.

First Supervisor: Pro. Dr. Guido Heineck

Second Supervisor: Prof. Dr. Michael Gebel

Third Supervisor: Prof. Dr. Silke Anger

Date of Oral Examination: December 22nd, 2021

URN: urn:nbn:de:bvb:473-irb-564101
DOI: https://doi.org/10.20378/irb-56410

Acknowledgements

Like at the beginning and the end of all my journeys in life, upon completion of my dissertation I am deeply grateful to GOD ALMIGHTY for His graces and blessings, for all inspirations He provided to me through the ups and downs in my life, or by making my path to encounter, get to know, and learn from such inspiring people and great minds.

First and foremost, I would like to thank my first supervisor, Guido Heineck, for his initial trust in my work followed by his ongoing and kind support. I have always enjoyed his admirable sense of humor that made most of his valuable feedback stick in my mind, sometimes performing rather than just stating his comment on my work. I will never for example start my "introduction" with something like, "according to the definition by…" because I would then imagine Guido falling asleep with such a boring start to my paper. Thank you Guido for all I learned from you, your knowledge, expertise and your great personality. I also want to deeply thank my second supervisor, Michael Gebel, for his instrumental contribution in guiding my projects. Despite his busy schedule, he was so kind and supportive to closely consider my work. Although we did not meet on a regular basis, every single session was an intensive progression to my work, where I learned from both his direct comments and sometimes his brief opposition and leaving me to find my own way out of the challenges, a crucial skill I certainly need in my career path. Thank you Michael for the latter even more than the former. Moreover, I would like to thank Silke Anger who agreed to be my third supervisor. I am also indebted to the Bamberg Graduate School of Social Sciences (BAGSS) and the Women's office at the University of Bamberg who provided me with funding and a smooth and fruitful environment for research. Furthermore, I sincerely thank two of the greatest experts in the field, Jeffrey M. Wooldridge and Donald B. Rubin, who upon their workshops in Germany kindly considered my work and provided me with valuable general comments on the methodology and robustness checks.

My special appreciation also goes to Prof. Dr. Ali Naghi Mashayekhi, whose deep concerns about the education system in Iran along with his resilience and positivity to plan and implement educational reforms in spite of all the stumbling blocks on the way have always been inspiring and motivating to me. In collaborating with his initiative the "Aseman Educational and Research Group" at Sharif University of Technology in Tehran, I learned how to think big and aim high, and at the same time plan for incremental and continuing improvements. I will always remain thankful to him for my

academic path from engineering to this more social field of study where I can pursue my values and social concerns.

I am also grateful to Prof. Dr. Reyn van Ewijk, for his inspiring lectures on “economics of education” at Goethe University of Frankfurt, for his valuable comments on my master and PhD research projects, and his kind support beyond the call of duty as a master thesis supervisor. Furthermore, I express my gratitude to my colleagues from the Chair for Empirical Microeconomics at BAGSS and the LifBi institute, specifically Stefanie Herber, Johanna Sophie Quis, Anica Kramer, Jacqueline Lettau, and Susanne Elsas for their useful comments on my presentations. In addition, I owe many thanks to authorities in the education system in Iran who paved my way to collect the data I needed for my research, to Dr. Khodaei, Mr. Hashemkhani, Dr. Raeis Jafari, Dr. Gharibi, Dr. Tofighi, Dr. Mehran, and the authorities at Sanjesh Organization and the University of Allameh Tabatabaei in Tehran.

Last but certainly not least, I thank my family, my parents, and friends. Despite their initial opposition to my decision to leave the country and study abroad, my parents have always backed me up and lifted my mood when I talked to them via phone, like absorbing all my frustrations and disappointments and cheering me up from thousands of miles away. My father-in-law who just passed away due to Covid-19 used to prevent me from helping him out for housework while we visited them for vacations in Tehran. With his great kindness and modesty, he always insisted that I focus on my work rather than “unimportant housework that he could also do”. I thank my friend Daniela who remained a good friend even after planned (child birth) and unplanned (Covid-19 lockdowns) life events prevented us from meeting more frequently. Finally, I am deeply thankful to my husband, Mohsen, who encouraged me to pursue my dreams and patiently coped with the stresses and difficulties of moving to a foreign country, and to my two lovely children, Amir and Selma, who also took some unavoidable parts of the burden of my efforts to complete my dissertation.

Thank you all!

Bamberg, 22. June 2021

Zahra Kamal

Contents

List of Tables

List of Figures

List of Abbreviations

COED	*Coeducation (mixed-gender educational environment)*
DiD	*Difference-in-difference*
FD estimator	*First-difference estimator*
FE estimator	*Fixed-effect estimator*
GPA	*Grade Point Average*
HLM	*Hierarchical Linear Modeling*
IEA	*International Association for the Evaluation of Educational Achievement*
JRR	*Jackknife Repeated Replication*
K-12 education	*Kindergarten to 12^{th}-grade education*
MLE	*Maximum Likelihood Estimation*
MLR	*Multiple Linear Regression*
MM countries	*Muslim-majority countries*
OECD	*Organization for Economic Co-operation and Development*
OLS	*Ordinary Least Square*
PIRLS	*Progress in International Reading Literacy Study*
PISA	*Program for International Student Assessment*
SES	*Socio-economic Status*
SLR	*Simple Linear Regression*
STEM	*Science, Technology, Engineering, Mathematics*
TIMSS	*Trends in International Mathematics and Science Study*
UAT	*University of Allameh Tabatabaei*

Chapter 1

Introduction

My research projects in this dissertation lie under the umbrella theme of *gender interactions in classroom environment and the impact on student achievement*. The relevance of this topic to the economic literature could be established from two different but related aspects:

Broadly speaking, education is the most fundamental tool for improving human capital as a central element of economic growth (Hanushek, 2013). For much of the last 100 years, a large and expanding body of research considers the growth of economies, with most studies remaining as theory and much less as empirical work (Hanushek & Woessmann, 2010). More recently however, studies particularly focused on empirical observations and the role of human capital in their growth modelling (Hanushek & Woessmann, 2010). Given the linkages between the basic area of education and labor market outcomes at the individual and national levels, quantitative measures of education have received the most attention among the predictors of economic growth (Hanushek & Kimko, 2000; Hanushek & Woessmann, 2008). Hanushek and Woessmann (2012, 2015) showed that three-quarters of the variations in country growth rates could be explained by a simple growth model that focuses on

cognitive skills measured by student test scores[1]. Assuming education as a production function with test scores as the output, much attention has been directed at the inputs[2], particularly those perceived to be relevant for policymaking such as school resources, classroom environment, and teacher characteristics (Hanushek, 2020).

More specifically, the gender dynamics in classroom has been regarded as a setting for policymaking to promote gender equality in countries. Gender equality has been at the core of many national attempts for economic prosperity in different countries (Smithers & Robinson, 2006; Novotney, 2011). Regarding the relationship between gender equality and economic growth, empirical evidence for both positive and negative links exists[3]. However, according to a recent review by Klasen (2018), most theoretical and empirical literature shows that gender gaps, particularly in the basic area of education, hinders economic growth, with the impact being sizable and robust. Given that educational inequalities and gender differences in academic outcomes most likely lead to labor market inequalities through differentiated signaling or occupational opportunities (Lim & Meer, 2015), identifying the root causes of gender achievement gaps appears to be a fundamental economic issue. A broad range of policies and interventions has been proposed to tackle the existing gaps in education. As the social rather than biological factors come into primary focus for explaining these gaps (Klein, 2004; Eisenkopf et al., 2015; Salikutluk & Heyne, 2017), a large and growing body of research has examined the effect of gender interactions in classroom either among the students or between the students and the teacher.

[1] According to Hanushek (2013), the focus on human capital rather than merely improving school attainment underscores the importance of improved cognitive skills as the main driver of economic growth. The cognitive skills better reflect the education quality than measures of school attainment and are best measured by students' test scores especially in the international standard tests.

[2] For a broad review on determinants of student performance, see for example, Woessmann (2005), Hanushek and Woessmann (2011), or Caponera and Losito (2016).

[3] See for example Klasen (2002), Klasen and Lamanna (2009), Duflo (2012), World Bank (2014), and Hakura et al. (2016), for the positive link, and some earlier studies such as Barro (1991), and Barro and Lee (1994) which found a negative link between gender equality and economic growth. To explain the conflicting empirical evidence and the complex relationship between inequality and growth at the national level, Turnovsky (2015) emphasizes on the role of public investment which might depend on several factors i.a. the selected analytical framework, the relative magnitude of externalities, underlying financing policies, time period and factor sustainability.

Summing up, from either of the aforementioned perspectives, the impact of gender interactions in classroom on student achievement is highly relevant in education economics as it provides policy implications for promoting gender equality, increased human capital, and economic growth.

Regarding the gender agenda in classroom, two policy levers have been proposed and widely discussed in the economic literature, namely the use of single-sex education[4] and student-teacher gender-matching (Winters et al., 2013). It has been suggested that the gender composition of peers in classroom could affect students' performance mainly through its impact on the activation or reinforcement of gender stereotypes and altering students' self-concept, or via potential sexual attractions among genders and the related distractions that hinder the focus on academic tasks (Hoxby, 2000; Lavy & Schlosser , 2011; Jackson, 2012; Oosterbeek & van Ewijk, 2014; Pahlke & Hyde, 2016; Booth et al., 2018). On the other hand, teachers might attribute and display higher academic expectations to the students of the same gender, leading to their higher motivation and better self-concept. Teachers who share gender with their students' might act as better role models[5] and be less likely to activate negative stereotypes against their gender group in class, resulting in improved identification and self-concept of the same-sex students (Ammermueller & Dolton, 2006; Dee, 2005; Paredes, 2014; Lim & Meer, 2015; Gershenson et al., 2016).

Despite the large number of empirical studies on both topics, the results are quite mixed (Pahlke et al., 2014; Cho, 2012). Regarding the gender peer effect, empirical evidence exists for nearly all possible directions. Some studies found that single-sex education is only beneficial for girls (Laster, 2004; Adkinson, 2008; Santos et al., 2013), some found it beneficial for boys (Riordan, 1994; Brathwaite, 2010; Sullivan et al., 2010), for both genders (Riordan, 1985; Doris et al., 2013), or for neither of the two (Baker et al., 1995; Edwards,

[4] The U.S. Department of Education defines single-sex education as "education at the elementary, secondary, or postsecondary level in which males and females attend school exclusively with members of their own sex". In contrast, coeducation is provided when students attend in a mixed-gender setting (U.S. Department of Education, 2004). A related though different phenomenon is single-sex classroom, offered by institutions that enroll both genders while offering separate classes for each gender (Mael et al., 2005).

[5] For a more detailed explanation about "role-model effect" please see chapter 4, section 4.2 about the theories behind teacher gender effect.

2002)[6]. Also, regarding the teacher gender effect, empirical studies offer almost all possible results, ranging from null or negligible effect of teacher gender (Ammermueller & Dolton, 2006; Holmlund & Sund, 2008; Cho, 2012; Sansone, 2017), to positive impact of a certain teacher gender for all students (Klein, 2004; Lim & Meer, 2015), or positive effect of student-teacher gender-match (Dee, 2007; Paredes, 2014).

The inconsistencies of the results are mostly attributed to the fact that such investigations are mostly afflicted by methodological problems, particularly due to the selection of students into schools of either type (single-sex versus coeducational) or the nonrandom assignment of students and teachers to classrooms (Dee, 2005; Cho, 2012; Jackson, 2012; Park et al. 2018). On the one hand, studies on gender peer effect predominantly compared the student performance in single-sex versus coeducational schools/classrooms, and thereby did not disentangle the effect of self-selection and institutional factors from the impact of the gender composition of learning environments per se (Park et al., 2018). On the other hand, the contradictory results of the studies on teacher gender impact might partly reflect the possible correlation between students' unobserved traits and the teacher sex that leads to biased estimates of the reduced-form equations (Dee, 2007).

Moreover, given the social essence of the impact[7], the inconsistencies could also pinpoint the key role of cultural background. The students' and teachers' perceptions about "gender" and "gender-match", which stem from their cultural background, most likely influence how they react to a certain gender agenda in classroom. Therefore, the mechanisms of the impact of gender interactions might largely differ in various contexts. For example, not for all peers in all classrooms stereotypes about the inferiority of females in math skills are equally invoked

[6] Despite the numerous but inconclusive studies on single-sex vs. coeducation effect, only few studies examined the impact of varying gender proportions in learning environment, mostly pointing to the positive link between the share of females and both girls' and boys' performances (Oosterbeek & van Ewijk, 2014; Lavy & Schlosser, 2011).

[7] WHO (World Health Organization) distinguishes between the two related concepts "gender" and "sex". The former is regarded as a "social construct" and refers to "the characteristics of women, men, girls and boys that are socially constructed, including norms, behaviors and roles associated with being a woman, man, girl or boy, as well as relationships with each other, and might vary from society to society and over time". The latter however, refers to "the biological and physiological characteristics of females and males such as chromosomes, hormones and reproductive organs" (World Health Organization, 2021).

or applied. Thus, the "stereotype threat" (Steele & Aronson, 1995) mechanism might be highly relevant in some contexts while it induces trivial effect on students from a different cultural background[8].

That said, one could hardly extrapolate the existing results to new contexts. This dissertation builds on the literature on *gender peer effect* and *teacher gender effect*, with particular focus on some of the unexamined contexts. The thesis consists of three research projects presented in the following three chapters (Chapters 2, 3, and 4).

My research project presented in chapter 2 examines the impact of participation in single-sex versus mixed-gender classrooms on student performance in higher education. As the major problem with most previous work on this topic is the challenge to address potential threats to causality (like self-selection), the project was motivated by a perfect research opportunity to address the causal link: In 2011, one of the largest universities in Iran launched a policy of gender separation[9] at classroom level without publicly announcing it beforehand. As a result, the cohort admitted at the university in that specific academic year had not selected to but participated in separated classrooms, in most cases with the same professors and curriculum as the previous cohort who entered the university in 2010. Via a long and time-consuming process to get access and combine the administrative data collected by two governmental organizations from the cohorts of pre- and post-policy implementation, I managed to utilize this natural experiment to identify the causal impact of participation in single-sex versus mixed classrooms on student achievement. Using the difference-in-difference (DiD) approach, I found that when students' characteristics and educational competencies were taken into account, participation in single-sex classrooms improved both males' and females' average performances in exams. While the academic benefit of the gender separation policy for females did not depend on their ability level, the effect was

[8] For more detailed explanation of the "stereotype threat" theory, please refer to section 2.2 in chapter 2, on the theoretical arguments.

[9] I use the term *gender separation* as in the literature *gender segregation* in education is mainly used to address the policy of *imposing gender-based restrictions on enrollment for certain fields of study*, leading to differentiated educational choices, limited access to specific job sectors, and gender inequality in working life. [e.g. Wilson & Boldizar, 1990; Epstein, 1997; Mehran, 2003; Shirazi, 2014; Barone, 2011; Vuorinen-Lampila, 2016].

considerably heterogeneous among males with different initial ability. Nearly all positive effect for males was driven by upper-medium-ability male students performing remarkably better in all-male classrooms. As the extensive research on single-sex education has mostly focused on the context of K-12 education and the context of western or east-Asian countries (eg. Eisenkopf et al., 2015; Booth et al., 2018; Park et al., 2018), my research addressed the gap in the literature for investigations in higher education and on the different cultural context of a Muslim-majority country.

The analyses in chapter 3 investigate how student achievement in mathematics in different countries is affected by the gender composition in classroom via a cross-country analysis of TIMSS[10] data. Using a hierarchical linear modelling approach (random intercept models), I conducted two separate but related analyses, the first assessing non-extreme cases of varying gender proportions in mixed settings, and the second evaluating binary cases of single-sex versus coeducation. The first analysis for 37 TIMSS-participating countries showed that both male and female students gain academic benefit from a higher proportion of female classmates. The academic improvement was however nonlinear throughout the range of female proportions. The second analysis for the 17 participating countries with sizable single-sex and coeducation, revealed that students mostly benefit from the presence of the opposite gender in classroom environment. In both analyses, the impacts were heterogeneous across the countries. While I abstain from causal interpretation of such cross-sectional analyses, I relied on the rich multivariate quality data collected by TIMSS 2015 from nearly 40 countries and addressed the dearth of such investigations in some of the unexamined contexts.

My third research project, presented in chapter 4, is an attempt to examine the gender agenda in classroom from a different aspect, namely the student-teacher gender match. The study investigates how the teacher gender influences the academic outcomes of male or female students. Exploiting the data structure of TIMSS that provides two test scores for each student (in math and science subjects) and the fact that in secondary education most students have different math and science teachers, I used student-fixed-effect approach- proposed by Dee (2005)- to deal with unobserved student-level variables (eg. ability) and address the threats

[10] Trends in International Mathematics and Science Study

to causality in the associations. My particular focus in this project was the context of Muslim-majority countries with gender issues being highly relevant but mostly untouched in the economic literature. Findings about the eight Muslim-majority country that participated in TIMSS2015 showed that girls and boys in these countries generally perform better when assigned to male rather than female teachers. The impact was specifically large for single-sex classrooms and heterogeneous across individual countries.

To the best of my knowledge, no empirical study so far has investigated the causal impact of single-sex versus coeducational classrooms at higher educational levels in Muslim-majority countries. Nor am I aware of international comparisons of student performance, which have emphasized the role of varying gender shares in classroom. Also, no empirical study in the economic literature so far has investigated the causal impact of teacher gender in Muslim-majority countries. As mentioned above, the existing evidence is not generalizable specifically to the countries with distinctive cultural norms and values such as the Muslim-majority countries. While gender-related concepts have nearly been "undone" among most western countries (Riegle-Crumb & Humphries, 2012), they are differently perceived and highly endorsed in gender-segregated societies. Girls and boys in these countries grow up mostly in separated environments with gendered social norms and values transmitted to them by the media and their significant others. As a result, they naturally develop differentiated motivation, self-concept, and educational aspirations (Maher & Al-Malki, 2014). In such context, the gender composition of peers and the gender of the teacher most likely operate differently. To illustrate, with gender stereotypes being more reinforced in the media and social scenes, the stereotype threat is probably much more relevant for a girl from a Muslim-majority country rather than a girl from a western society. Similarly, while the gender of the teacher is not an issue for girls who have had several male teachers from childhood, it is probably the most salient attribute of the teacher for girls who firstly encounter a male teacher in secondary education, inducing entirely different studying behaviors. As another illustration, having a female math teacher with profound math background could induce a stronger role-model effect for a girl from a traditional Muslim family who has not typically interacted with highly-educated women outside the classroom than for a girl from a western and gender-equal cultural background. Thus, a separate investigation in the context of

Muslim-majority countries could fill the gap in the literature for a solid inference about the impact of gender interactions in classroom.

My research findings provide important implications for the policy-making not only in Muslim countries but also for western societies with large inflows from the middle-eastern countries. Destination countries mostly face serious challenges of integrating their new members into the mainstream society. Despite their well-designed policies to address these challenges, they sometimes end up with unintended negative outcomes or simply do not reach their integration goals. This has particularly happened for the educational assimilation of the newcomers from Muslim societies in Germany, with post-evaluations revealing that the reform mostly became futile on the ground of cultural ignorance (Dahl et al., 2020). In fact, certain distinct aspects of Muslim culture that naturally induced contradictory mechanisms had been initially overlooked by the policymakers. Similarly, the so-called cultural ignorance might cause policymakers to pass over some of the potentially strong and effective policy levers to address integration issues. A closer look on the gender dynamics in classroom could provide insights for accelerating the reforms with integration goals and promoting educational equality in destination countries.

Chapter 2

Gender Separation and Academic Achievement in Higher Education; Evidence from a Natural Experiment in Iran

2.1. Introduction

Does it make a difference who you sit next to in class? Does being surrounded by classmates of your own or the opposite sex affect how much you learn and how you perform in exams?

Previous research on single-sex education has produced inconsistent results, mainly due to methodological issues and selection biases (Pahlke et al. 2014). According to Jackson (2012), most of these studies suffer from two major limitations: first, because students who *decide* to participate in single-sex education are likely to differ in important unobserved characteristics from those who opt for attending coeducation, comparison between the two groups' outcomes is potentially subject to severe *self-selection bias*. Second, since single-sex institutions often differ systematically from mixed institutions (eg. in terms of curriculum, selectivity, teachers' motivation and compensation, extracurricular activities, and so forth), the comparisons confound single-sex education effects with other institutional differences.

Moreover, while many rigorous studies examined the impact of single-sex schooling, research on the effect of such policies at higher educational level is rare (Pahlke et. al. 2014). Due to the potentially different underlying mechanisms for different age groups, results from research on different schooling levels are not applicable to other levels of study. Therefore, a separate investigation of the consequences and impacts at each educational level is crucial (Pahlke & Hyde, 2016).

Furthermore, single-sex education has been under scrutiny in several western countries such as Britain, the United States and Canada where single-sex schools make up a small and selective group, or in New Zealand, Australia and Ireland, countries with a sizable number of single-sex schools (Smyth, 2010). However, studies on the effect of the policy in Muslim-majority countries are scarce although single-sex education is even more prevalent in such societies, and many students spend all their school years in separated environments. As the mechanisms and thereby the size and direction of the policy effect heavily depend on the context (Baker et al., 1995), the results from western countries are hardly applicable to societies with different cultural norms and values.

In order to contribute to the literature by particularly addressing the research gaps mentioned above, in this chapter, I evaluate the impact of gender separation policy *at higher educational level* in the *context of a Muslim-majority country*. As the design in this study exploits a unique natural experiment setting - an *abrupt* change of policy in *one* university - no selection bias based on students' choice or the institution's characteristics is expected to influence the results. The chapter also reports on the moderating effect of initial ability on the impact for different subgroups of students.

2.2. Theoretical Arguments

The recent resurgence of single-sex education is mainly associated with rising concerns about gender equality (Hannan et al., 1996). Many scholars, policymakers and authorities in education have debated the merits of single-sex education as a tool to address existing gender gaps in academic performances, decisions to study certain fields and degrees, and occupations and wages (eg. Salomone, 2006; Billger, 2009; Booth et el., 2018). In this

section, I present the main rationales for the arguments of the supporters and the counterarguments of the critics of single-sex educational environment.

2.2.1 Arguments for separated education

Biological and behavioral differences

Sax(2005) represents the essential-difference view asserting that substantial biological differences between girls and boys lead to different learning processes, and thus, are educationally relevant (Sax, 2005). He argues that that by failing to recognize these differences between girls and boys, teachers and schools are unable to support students to reach their full potential, and that students perform optimally if instruction targets these learning-related differences in single-sex classrooms (Sax, 2005). In addition, some researchers imply that certain behavioral differences between girls and boys such as boys' tendency to call out answers or more hands-on activities in class may lead to one gender (mostly boys) receiving most of the teacher's attention (Smyth, 2010).

Nevertheless, gender differences are generally addressed as a ground for separating boys and girls at primary educational level. According to Raznahan et al. (2010), sex differences in brain-related behavior and cognition diminish as a function of age. As children enter adolescence, they develop stereotype consciousness and awareness of others' stereotypes (Pahlke & Hyde, 2016), and interact quite differently (Oosterbeek & van Ewijk, 2014). Therefore, for adolescents the following rationales are more relevant.

Sexism and gender biases

Several supporters of single-sex education focus on sexism and biases particularly aimed at female students in coeducational setting. In this respect, three theories are mostly emphasized to illustrate the mechanism for the impact (Pahlke & Hyde, 2016):

Firstly, *stereotype threat* has been defined by Steele and Aronson (1995) as the risk of negative evaluation and rejection by others. When a person feels at risk of confirming negative stereotypes about her group, she is likely to underperform due to the perceived anxiety about being judged based on those stereotypes rather than personal merit (Steele et al., 2002). Thus, the theory implicitly posits that the elimination of stereotype threat could result in better performance of students who otherwise feel at risk of being stereotyped.

Advocates of single-sex education argue that while coeducation reinforces and activates commonly held stereotypes against females' abilities, in all-girl classrooms leaders and top-performers in all subjects are female students. Therefore, by having good same-sex role models, females are unlikely to hold these stereotypes in a single-sex environment (Park et al., 2018). Additionally, in all-female classrooms females feel no pressure to conform to negative stereotypes, leading to better performance and higher scores (Jackson, 2012).

Secondly, *expectancy-value theory* posits that a student's perception of others' endorsement of traditional gender stereotypes may result in less self-confidence and interest for pursuing gender-atypical fields (Lee & Bryk, 1986). If negative stereotypes about females' abilities are activated in mixed classrooms, females become aware that others expect low performance. This perception might negatively affect females' academic goals and performance in traditionally masculine fields such as STEM[11] (Sadker et al., 2009).

Thirdly, according to *identity theory*, perceived group status differences, perceived legitimacy and stability of the status differences, and perceived ability to move from one group to another affect one's behavior and performance (Tajfel & Turner, 1979; Turner et al., 1999). Supporters of single-sex education suggest that, in coeducational contexts, status differences are probably endorsed, for example, by males making negative comments on females' abilities and competencies in specific subjects (Pahlke & Hyde, 2016).

Adolescent culture and sexual attraction

To justify their support for single-sex education, a number of advocates refer to adolescent culture based on sexual attraction among genders which distracts student's attention away from academic tasks in coeducational contexts. In an early study, Coleman (1961) drew attentions to "rating and dating culture"- i.e. students' obsession about appearance and attractiveness and peer pressure for prioritizing relations with the opposite sex over schoolwork- as a main reason for the low achievement of girls in coeducational American high schools (Smyth, 2010). Several later studies such as Dyer and Tiggemann's (1996)

[11] STEM (Science, Technology, Engineering, and Mathematics) subjects are recognized in the literature as male-dominated subjects, in which females are underrepresented or typically underperform (See for example, Park et al., 2018).

endorsed his findings. Similarly, Riordan (1985) points out that high-ability girls intentionally avoid competing with boys because excelling academically might make them unattractive as potential sexual partners for boys. Consequently, proponents of the policy argue that in the absence of the opposite sex students could better concentrate on their learning tasks and academic activities.

2.2.2 Arguments against separated education

Insufficient evidence for relevant gender differences

The line of reasoning against single-sex education is primarily based on the insufficiency of scientific evidence for essential learning-related differences among genders and implications for single-sex education (see for example Halpern et. al, 2011). For instance, in his argument in support of single-sex education, Sax refers to a number of studies showing distinctive learning-related processes and behaviors among genders[12]. Several scholars criticize his views arguing that some of these studies have used inadequate and non-representative samples or find only small differences between males and females[13] (Bracey, 2006; Liberman, 2008). In contrast, Hyde (2005) emphasizes *similar* psychological traits among males and females and demonstrates that gender differences can vary substantially in magnitude at different ages and in various measurement contexts.

Reinforcement of gender biases and stereotypes

Despite the perspectives on more sexist attitudes in coeducational settings, opponents argue that dividing students by gender can reinforce gender biases and entrenched stereotypes. They refer to *development intergroup theory* which assumes that increased psychological salience of gender leads to higher levels of essentialist thought, in-group favoritism, and out-group bias (Pahlke & Hyde, 2016). Epstein (1997) expresses worries that by denying the diversity within educational institutions, stereotypes are perpetuated. Halpern et al. (2011) assert that the relative presence, intensity, and activation of stereotype threat in single-sex

[12] For example, Sax has referred to i.a. Corso's (1959) study about sex differences in hearing, Lenroot et. al's (2007) paper on sexual dimorphism of brain developmental trajectories, and Raznahan et. al's (2010) research on sex differences in brain-related behavior and cognition.

[13] In particular, regarding Sax's assertions about sex differences in hearing based on Corso's (1959) study, Liberman (2008) says that the study found only between one-quarter and one-half of a standard deviation in male and female hearing thresholds.

versus mixed environment is not clear-cut, and present evidences that separating genders in educational contexts gives rise to gender stereotyping.

Beyond educational achievements

In addition to all other counterarguments, many opponents imply that regardless of the underlying rationales, separating genders in education is problematic for the same reason that segregation by race and social class is, that the diverse environment in education promotes tolerance and cooperation (Rustad & Woods, 2004). They worry that by reduced cross-group communication in single-sex classrooms, students are less likely to learn from and cooperate with one another (Jackson & Smith, 2000). Hyde (2005) expresses concerns about potential harm in numerous realms beyond educational outcomes including women's opportunities in the workplace, couple conflict and communication, and self-esteem problems among adolescents.

2.3. Literature Review and Knowledge Gap

Numerous empirical studies evaluated the effect of single-sex versus mixed schooling on various outcomes of students either at primary or secondary level (eg. Riordan, 1994; Campbell & Evans, 1997; Hoffman et al. 2008). Most studies examined the impact on students' academic performance in certain subjects such as mathematics, science, and verbal/English (eg. Baker et al., 1995; Jackson, 2012; Eisenkopf et al., 2015). Other outcomes most frequently addressed in the literature are students' tracking, course-taking choices, and subject preferences (Billger, 2009; Jackson, 2012; Schneeweis & Zweimüller, 2012).

Despite the vast literature on single-sex schooling, the overall picture of the impact is still ambiguous as the findings are inconsistent and in many cases contradictory. Most studies found a positive effect on females' performances and a negative or statistically insignificant impact on males' (eg. Adkinson, 2008; Lee & Bryk, 1986, Laster, 2004; Santos et al., 2013; Sax et al., 2009). However, there also exist studies implying that the policy is merely beneficial to males (eg. Brathwaite, 2010; Riordan, 1994; Roth, 2009; Spielhofer et al., 2004; Sullivan et al., 2010). Furthermore, while some studies strictly favored single-sex education both for male and female students (eg. Riordan, 1985; Stephens, 2009; Doris et al., 2013), several found null effect on either group (Baker et al., 1995; Edwards, 2002), and some

reported mixed evidence both in support of and against single-sex schooling (Stotsky et al., 2010; Vrooman, 2010).

Many scholars regarded research design issues as the primary reason for inconsistent results in this field (Jackson, 2012; Pahlke et al., 2014; Park et al., 2018). According to Park et al. (2018), findings from most of the previous literature do not disentangle the effect of self-selection and institutional factors from the impact of the gender composition of learning environments per se. Some scholars attempted to overcome methodological issues and selection biases by conducting randomized experiments, controlling for confounding factors, or exploiting a natural experiment setting. Nevertheless, their results were also mixed: For example, Park et al. (2013) used the exceptional feature of the current educational system in Korea that randomly assigns students to high schools, and found that both boys and girls outperform in a single-sex environment. However, they did not disentangle the impact from the effect of school factors such as the degree of autonomy in the teacher hiring process and teacher tenure policies that were mostly associated with private single-sex schools in the South Korean educational system (Eisenkopf et al., 2015). In another study, Eisenkopf et al. (2015) addressed the issue of institutional factors using a natural experiment performed at a single high school in Switzerland where the same teachers at the same school taught all-female and mixed classes. Their findings showed a positive impact of single-sex education on females' proficiency in mathematics but not in native language skills.

Scholars have also emphasized the role of several moderators as sources of variation in the size and direction of the impact found by distinct studies (Pahlke et al., 2014). In their review, Pahlke et al. (2014) identified three main moderators (besides age). 1) Dosage or level of exposure (class- or school- level separation): most findings indicated larger effects among girls when single-sex versus coeducation occurred in classes rather than in schools. 2) Socioeconomic status: the policy has been recognized to be more beneficial for students of lower social class. 3) Race/ethnicity: the impact on various racial groups received the most attention in the American studies that mainly reported an educational benefit for minorities (see for example Riordan, 1994; Gordon et al., 2009). Additionally, some studies demonstrated the role of *ability level* on the impact (Oosterbeek & van Ewijk, 2014; Eisenkopf et al., 2015). However, empirical findings on the role of innate ability are mixed.

For example, while Oosterbeek and van Ewijk (2014) found no evidence for heterogeneous gender peer effect based on students' ability level, Eisenkopf et al. (2015) found a larger impact on students with higher ex-ante ability.

Several scholars attempted to integrate previous research in the field and conclude on the size or at least the direction of the impact (Mael et al., 2005; Morse, 1998; Pahlke et al., 2014). In the most recent review, Pahlke et al. (2014) conducted a meta-analysis and assigned weights to the measured effects by past studies according to their sample size. Distinguishing between descriptive studies (with no control for confounding factors) and controlled studies (with appropriate controls or randomized experiment design), the researchers concluded that single-sex education was mainly supported by uncontrolled studies, and the results from controlled studies or random trials only showed trivial differences between students' performance in single-sex versus mixed schools, in some cases favoring coeducation. However, studies with experimental and controlled designs continue to produce inconsistent results. For instance, in a more recent study, Park et al. (2018) examined the impact of single-sex environment on students' performance in STEM subjects using a natural experiment approach. They found a statistically significant positive effect of all-boy schools on students' achievements in all STEM subjects. Interestingly, their findings revealed no statistically significant effect for females' performances in STEM. The authors attributed the contrast in their results with major previous related work to "no contamination by upward bias caused by positive selection into single-sex schools", and that "probably girls nowadays are less affected by different types of schools than in the past". Likewise, other scholars have stressed the influence of context on the impact. In their cross-country analysis, Baker et al. (1995) addressed the national contexts and cultural background as a reason for different estimated effects among various countries. According to Park et al. (2018), for a better assessment of potential costs and benefits of single-sex education more evidence on relevant outcomes under various contexts is needed.

Whereas most of the previous research examined the primary and secondary schooling context, very few studies focused on the impact of single-sex education at tertiary or higher educational levels (Pahlke et. al. 2014). Due to more freedom of choice for adults and their higher tendency to participate in mixed education, conducting a field experiment to evaluate

single-sex higher education is often prohibitively expensive. Few such studies tried to lower the costs by limiting their sample size to students in one major, or confining the exposure to treatment (single-sex education) to merely a small proportion of instruction hours. Oosterbeek and van Ewijk (2014) concentrated on gender peer effects and used a less extreme form of gender variation by exogenously manipulating the share of females in workgroups of first-year students majoring in economics and business at a Dutch university. The authors found only little evidence for academic success of female students that could be attributed to the increase in the proportion of women in workgroups. Interestingly, they found a negative impact of a higher share of females on males' performance in courses with a high math component. To explain this result, Oosterbeek and van Ewijk referred to their focus on university students rather than younger children at primary or secondary educational level arguing that male and female students might interact differently at various ages. This idea was reinforced by their supporting survey which showed that in tertiary education, the presence of males did not work disruptively in a traditional sense and did not cause reduced attention during class activities (Oosterbeek and van Ewijk, 2014). In a more recent study, Booth et al. (2018) conducted a field experiment at a high-ranked university in the UK to examine the effect of participation in single-sex versus mixed classrooms on students' first-year grades and their course choices in the second year. Their research design was restricted to participation in single-sex classrooms for one out of twelve instruction hours per week and to students majored in the field of economics. Thus, the authors did not claim to generalize the positive effect that they found to higher exposure to treatment or to students in other subject areas (Booth et al., 2018).

Among the countries and contexts under study, Muslim-majority and MENA countries have received the least attention in the literature. Nevertheless, the practice of single-sex education is even more prevalent in such cultures. In their meta-analysis, Pahlke et al. (2014) referred to only one study in the context of Iran as a Muslim-majority country[14]; Esfandiari & Jahromi (1989) compared the achievements and aspirations of students from a single-sex monolingual

[14] In their meta-analysis, Pahlke et al. (2014) also included some studies from Nigeria, a country which is sometimes counted as a part of MENA (eg. Banu, 1986; Egbochuku & Aihie, 2009; Lee & Lockheed, 1990; Mallam, 1993).

high school and a bilingual mixed high school in Tehran. However, as the two schools differed in various systematic ways, the measured effect was not plausibly attributable to the gender composition of the educational environment as the authors concluded on the effect of bilingualism versus monolingualism rather than single-sex versus coeducation.

The current study adds to the literature by providing evidence for the impact of gender separation policies at the higher educational level. Additionally, as the data are from administrative sources of a large university in Iran, the results provide implications for other Muslim-majority countries with a dominant culture of religious norms and Islamic values. I also investigate the heterogeneity of the impact by student's initial ability.

While in line with the more recent stream of empirical research focusing on adult interaction in higher education (eg. Oosterbeek and van Ewijk, 2014; Booth et al., 2018), I expect a positive effect of a single-sex environment on students' academic outcomes, the overall blurred picture provided by previous empirical findings does not allow for precise expectations.

2.4. Context Overview

2.4.1 Iran education system

Iran's education system was modeled on the French Education structure in the 19th century. Formal education is highly centralized and divided into K-12 education plus higher (tertiary) education supervised by the Ministry of Education and the Ministry of Science, Research and Technology respectively. There are both public and private institutions at all educational levels from elementary to university levels. Individual schools have the authority to take their exams at the end of each academic year. However, in the last years of both elementary and secondary levels, all students participate in the same final exams held at the national level.

2.4.2 Higher education and university admission process

Iran has a large network of private and public or state-affiliated universities offering degrees in all levels of higher education. According to the last report of the Institute for Research and Planning in Higher Education (2017), among the 2569 higher educational institutions in Iran, 141 public universities -the most competitive and selective institutions- have capacity for

only less than 20% of Iranian university students. To let the most talented students enter public universities, Iranian male and female students graduated from high schools have to participate in a National Examination for University Entrance -called Konkour (from the French "Concours"). Seeking an admission to public universities, around one million high school graduates take part in Konkour each year in one of the five disciplines (exam groups): Physics and Mathematics, Natural Sciences, Humanities, Art, and Foreign Languages. Then, having their raw Konkour test scores, the participants can determine their preferences for application to universities in order of priority and submit the selection lists to the Sanjesh Organization, a governmental agency which administers all processes related to Konkour and university admissions under the supervision of the Ministry. Students are assigned accordingly by the organization to universities in successive rounds. In the admission process by the Sanjesh Organization, preferential treatment is considered for Konkour participants from lower social classes and disadvantaged families in order to remove educational gaps among the Iranian population. As an affirmative action policy, the "quota system" has been in practice since 1983. Accordingly, the organization assigns quota (1) to eight highly-developed big cities, quota (2) to 141 medium-developed cities, and quota (3) to the remaining less-developed and small cities and all rural areas[15] and a certain quota for the students from martyrs and veterans families. Thereby, deprivations caused by non-ability related factors are at least partially compensated for.

2.4.3 Single-sex education

K-12 education has always been separated by gender in Iran[16]. Even before the Islamic Revolution in 1979 schools were basically either for girls or boys, reflecting religious norms and the culture of the society. In contrast, higher education was mainly not separated by gender, and only few all-female universities existed. After the Revolution, single-sex schooling was regulated and maintained, and higher education remained primarily as coeducation. Today, among public universities, very few have limited their enrollment by

[15] According to the latest report of the Ministry of Interior about the official administrative subdivisions of Iran, there exist 31 Provinces, 429 counties, 1057 districts, and 1245 cities in Iran (Statistical Center of Iran, 2019).

[16] Coeducational schools merely existed in rural and remote areas due to a limited access to educational institutions. There were also very few mixed international schools in the capital or big cities aimed at the children of foreigners residing in Iran.

gender. Nevertheless, in post-revolutionary Iran, the issue of gender separation in educational environments has always been a controversial debate, which mostly relies on ideological and political ground rather than the expected benefits proved by policy evaluations (Iranian Association for Scientific Development, 2011). The debate stems from the "Ratification of Retaining Islamic Values in Universities and Higher Education Centers" passed by the Supreme Council for Cultural Revolution in 1987. In an attempt for the Islamization of universities' environments, the ratification required that universities with adequate facilities and resources offer separate classrooms for male and female students (Supreme Council for Cultural Revolution, 2011). The ratification had not been enacted until a recent resurgence of the issue among authorities and politicians between 2009 and 2011.

In the academic year 2011/2012, one of the highest ranked and largest public universities in Tehran -The University of Allameh Tabatabaei[17] (UAT) – started to implement the policy of gender separation at classroom level, *without a pre-announcement to the public*. Thus, while undergraduate students who selected and were assigned to the UAT in that academic year expected to attend coeducation like the previous cohorts, they attended classrooms merely with those of their own sex. As the UAT implemented the policy for all students at the same time and continued to offer single-sex classes in subsequent semesters, cohorts 2010/2011 and 2011/2012 studied their first-year courses in classrooms with distinct gender composition (mixed versus single-sex classrooms). The educational experience of the two cohorts in the first year was otherwise the same[18]. The curriculum in the first-year consisted of 18 to 20 compulsory credits, and did not change between the two years. Both cohorts had almost all of their lectures and instructions with the same professors for each course, and same professors instructed all-male and all-female classrooms for the second cohort, except for less than 20% of the credits (0 to 4 out of 20 credits). Other characteristics of the programs

[17] The University of Allameh Tabatabaei is the largest Iranian public university in Humanities and Foreign Languages with around 19000 students majoring in 197 disciplines and subfields at 11 faculties.

[18] To ensure this, I examined some of the university's official documents from the "office for educational planning" and the "office for human resource planning and recruitment" at the UAT. I also conducted several interviews with professors and students working and studying at the UAT faculties at the time of the policy implementation.

such as the assignments, tutorials, exams, and extracurricular activities were also fairly comparable for the two academic years.

2.5. Data

For my analysis, I combined two administrative datasets collected from the UAT administration and the Sanjesh Organization. The UAT data contained information on some of the basic demographic characteristics as well as on the educational program and first-year overall performances of the students who started their undergraduate study at the university either in 2010/2011 or in 2011/2012 and attended mixed or single-sex classrooms respectively. Sanjesh data included the students' high school GPAs (Grade Point Averages), exam groups and Konkour test scores. More specifically, for all 2672 UAT entrants of the two cohorts -1435 of the first and 1237 of the second cohort- the merged and cleaned dataset[19] observes these variables: *age, gender, cohort, Konkour quota, field of study, faculty, exam group in Konkour, Konkour test score, high school GPA, and first-year-university GPA*. In addition, in order to control for potential changes in one of the basic institutional factors between the two years, I used a separate dataset from the "office for human resource planning and recruitment" at the UAT, and calculated *student-to-professor ratio* in each faculty for each academic year. Basic summary statistics for categorical and continuous variables are provided in table 2.1 and 2.2.

[19] Roughly 85 percent of the UAT dataset was linked to Sanjesh data with no contradictory information on similar fields. For the few problematic cases where the information provided by the two organizations differed for the same individuals, I contacted the authorities at both organizations to decide on the correct value for the variables. There were less than 30 individuals in Sanjesh dataset that were not in the UAT's, and less than 25 individuals whose information was among the UAT dataset but not in Sanjesh's. The former students were the Konkour participants who had an admission from the UAT, but did not register as they decided to go to a private university in a different field of study. The latter individuals were students who got an admission from the UAT without being assigned by the Sanjesh organization to the UAT. Most of these individuals were exchange students or foreigners in the field of Persian Language. Although in some cases the inclusion of such individuals was ideal, I left them out from the sample relying on the fact that the registration and dropout of these students were entirely unrelated to the practice of gender separation policy at the UAT and that the number and proportions of each group hardly changed between the two cohorts.

Table 2.1 shows that the two cohorts are comparable in terms of socio-economic status, denoted by the proxy variable "Quota", and Konkour exam groups of the entrants. However, the proportions of female and male students differed between the cohorts.

Table 2.1. Summary statistics for the student and program covariates by cohort.

Variable/Value	1st Cohort (mixed classes)	2nd Cohort (single-sex classes)
Female	81.4	61.4
Age	19.60	19.82
	(3.78)	(3.91)
Quota		
Highly-developed regions	56.4	60.7
Medium-developed regions	24.7	21.0
Less-developed regions	13.9	14.2
Families of Martyrs and Veterans	5.1	4.0
Exam Group in Konkour		
Mathematics and Physics	10.5	8.1
Natural Sciences	12.8	15.8
Humanities	66.6	64.0
Foreign Languages	10.1	12.1
Field of Study at University		
Theology and Islamic Knowledge	2.2	2.4
Statistics and Mathematics	1.9	-
Accounting	2.4	5.5
Laws	2.9	5.1
Guidance and Counseling	7.1	9.0
Public Relations	2.7	4.2
Psychology	5.4	4.9
Journalism	2.6	4.3
Languages and Literature	14.9	19.4
Social Sciences	7.9	8.3
Economics	7.6	6.4
Educational Sciences	11.6	9.6
Political Sciences	2.0	-
Philosophy	2.0	-
Library and Information Science	1.7	-
Social Work	2.7	-
Management	20.0	17.0
ECO College of Insurance	2.4	3.9
Observations	1435	1237

Note: Own calculations based on the UAT datasets for the 2010/2011 and 2011/2012 entering cohorts. All numbers indicate the percentage of the sample group in the respective category, except for the values for the age variable which relate to the average age (in years) and standard deviations (in parentheses) for each cohort.

The reason behind this fact is that while the implementation of the gender separation policy was not announced to the public in advance, the capacity of enrollment for each field was

announced separately for males and females in the admission process of the UAT in 2011/2012 to allow for the offer of single-sex classrooms. Since women tended to be more successful than men at entering the UAT prior to 2011/2012, assigning equal shares limited females' ability to enter and enabled males with relatively lower Konkour test scores gain admission from the UAT in 2011/2012. Thus, controlling for incoming ability is of paramount importance in the analyses. To account for potential differences in the level of difficulty in the Konkour exam between the two years and have more precise sorting of abilities in the sample, I used the population mean and standard deviation of Konkour test scores for each exam group in each year and normalized the scores. Table 2.1 also shows that the university did not admit any student in certain fields of study in the academic year 2011/2012 (statistics and mathematics, political sciences, philosophy, library and information science, and social work). In the section on robustness checks, I discuss how this exclusion might affect the results.

Table 2.2 compares the mean academic performances of the two cohorts in the first-year university and high school final exams and Konkour. Accordingly, females in the second cohort, i.e. who attended single-sex classrooms at university, on average performed 0.51 points better in their first-year university exams than did females in the previous cohort who participated in mixed classrooms. However, males who attended all-male classrooms underperformed those who participated in mixed classes by 0.43 points. Both mean differences are statistically significant at the 1% significance level. Therefore, merely comparing the university achievements might lead one to conclude that single-sex classrooms had a positive effect on females' academic performances and a negative impact on males'. This inference is strengthened when the means at high school level are compared. Particularly for females, those who attended all-female classrooms and performed better in university had on average lower achievement scores at high school level. However, comparing the means for normalized Konkour test scores weakens the argument above because both male and female groups who outperformed at university (males of the first and females of the second cohort) had initially achieved higher test scores in the entrance exam (Konkour). Thus, if one regards Konkour test scores as more precisely reflecting individuals' ability, the mean differences in university performances could plausibly be attributed to

students' higher ability levels rather than the gender composition of their classrooms at university.

Table 2.2. Summary statistics for the student performance by cohort and gender.

Variable	Group	1st Cohort (mixed classes)	2nd Cohort (single-sex classes)	Mean Difference
GPA (University)	All	15.88	15.97	0.10*
		(0.05)	(0.06)	(0.07)
	Females	15.92	16.43	0.51***
		(0.05)	(0.07)	(0.08)
	Males	15.67	15.24	-0.43***
		(0.11)	(0.09)	(0.15)
GPA (High school)	All	16.51	15.68	-0.82***
		(0.05)	(0.07)	(0.09)
	Females	16.79	16.64	-0.15**
		(0.05)	(0.07)	(0.09)
	Males	15.28	14.15	-1.13***
		(0.14)	(0.12)	(0.20)
Konkour test-score	All	1.94	1.97	0.03
		(0.03)	(0.03)	(0.04)
	Females	1.94	2.21	0.26***
		(0.03)	(0.03)	(0.05)
	Males	1.93	1.59	-0.34***
		(0.06)	(0.05)	(0.08)
Observations		1435	1237	-

Note: Own calculations based on the merged dataset (the UAT and Sanjesh data for the 2010/2011 and 2011/2012 cohorts combined). The numbers stated for GPA scores represent the mean for students' overall performances at high school or first-year university level which can vary between 0 (the lowest mark possible) and 20 (the highest score possible). The scale for students' test scores in Konkour is different for each year/cohort. Thus, the Konkour test-scores were normalized according to the mean and standard deviation for the whole population of Konkour participants in each year-exam group. Numbers in parentheses state standard errors. Stars show the level of statistical significance (P-values of the t-test) of mean differences between the two cohorts (*** p<0.01, ** p<0.05, * p<0.1).

In other words, the higher incoming ability of outperformers at university could at least be partially responsible for the ostensibly large effect of single-sex classrooms. This is reinforced by comparing the distributions of the students' achievements at each educational level (first-year university, high school and Konkour exams) plotted separately for individuals participating in either type of the classrooms as illustrated in figure 2.1.

In figure 2.1, the solid lines show the Kernel densities for performances of students participating in single-sex classrooms in the first-year university courses, while the dashed lines relate to students who attended mixed classrooms. In the first two plots on the top row, the solid lines lie mostly above the dashed lines for females and below that for males. More

precisely, Mann-Whitney P-Values in both diagrams indicate that the difference between the distributions in each plot is statistically highly significant.

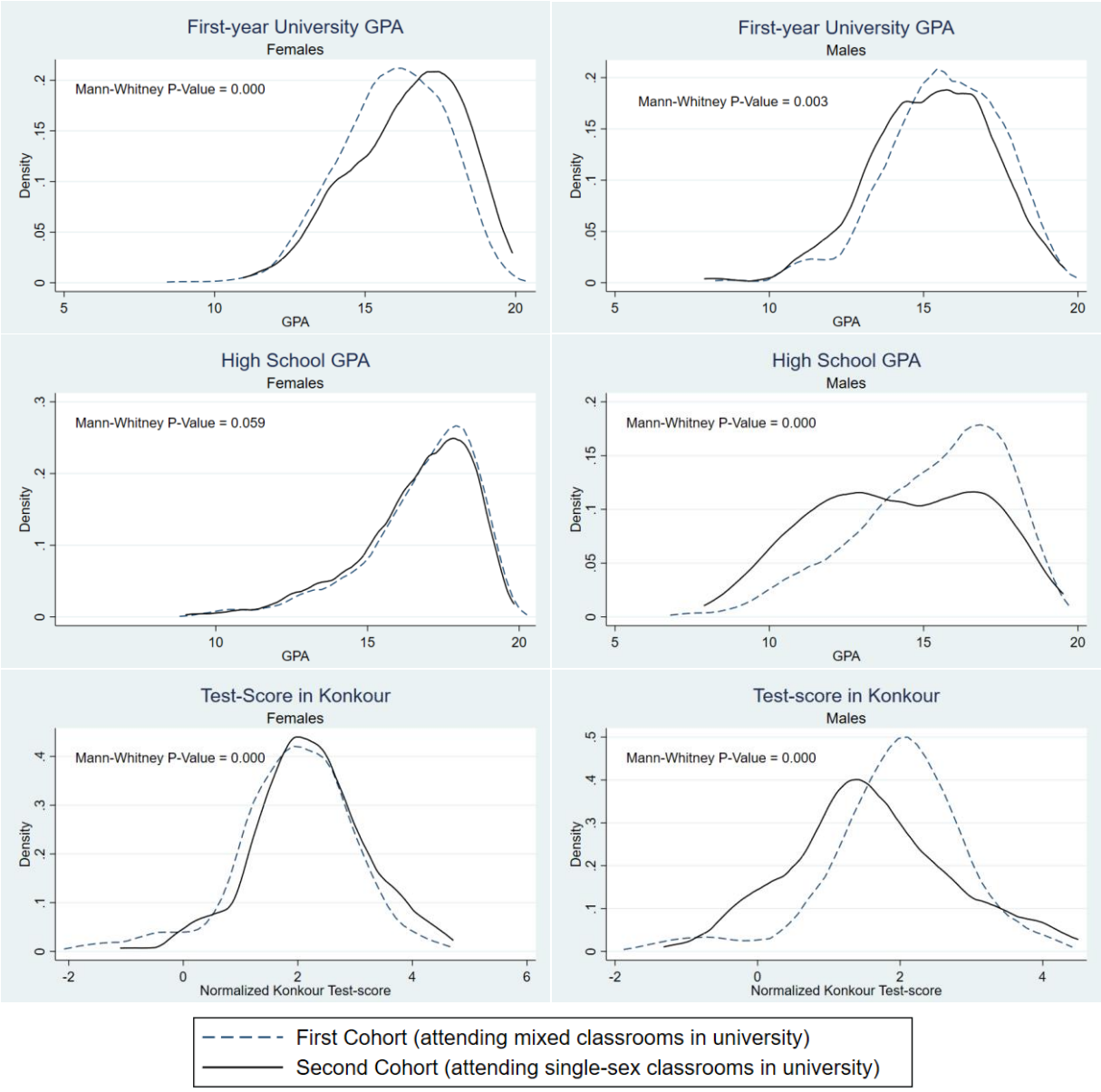

Figure 2.1. Distribution of first (dashed line) and second (solid line) cohort students' achievements at each educational level

According to the distributions, it seems that participation in single-sex classrooms had educational benefit for females while it does more harm than good for male students. It also appears from the females' university GPA distributions that all-female classrooms had more

benefit for females at the upper part of the distribution. However, the relative position of male distributions for university GPAs show that all-male classrooms could help males at the lower part of the distribution, while males at the upper parts do worse in a single-sex environment. Here again, the statistically significant differences in the distributions of pre-university performance in the remaining four plots (second and third rows) prevents a conclusion on the real effect of the policy. The distributions of the students' high school GPAs and Konkour test-scores show approximately the same patterns as shown in table 2.2, strengthening the conjecture that part of the seemingly large effect of single-sex education comes from pre-existing differences between the two cohorts in terms of their ability.

2.6. Method

To examine the effect of participation in single-sex versus mixed classrooms on educational achievement, the simplest approach is to compare the educational outcomes of the same students in both environment, namely the test scores of students in the second cohort at single-sex high school level with their outcomes at university level (mixed classrooms). This simplified view is however, prone to severe bias because besides the gender composition of the learning environment, many other relevant factors change between the two educational levels, thereby the zero conditional mean assumption is most likely violated. The results of this estimation approach is therefore not included in the analysis of this chapter.

Having the data for two subsequent cohorts, if one believes that the two cohorts have on average similar characteristics and similar experience regarding the curriculum and exams in the first year university study, a simple comparison between the first-year GPA of the two cohorts would reveal the impact of participation in single-sex classrooms. The impact is therefore estimated by an ordinary least square (OLS) estimation of the equation 2.1.

$$GPA1 = \beta_0 + \gamma_1 D + u \qquad (2.1)$$

In this equation, $GPA1$ stands for students' GPA at the end of the first year at university. The intercept β_0 shows the average first-year GPA of all students regardless of the year they entered the university. The error term u denotes individual deviations from the average test score estimated for each cohort. The binary variable D equals zero for students of the first

cohort (mixed classrooms) and one for the second cohort (single-sex classrooms)[20]. γ_1 is then the parameter of interest.

Nevertheless, the estimate of the simple linear regression (SLR) model for γ_1 is most likely biased because according to tables 2.1 and 2.2 and the figure 2.1, the two cohorts clearly differ in systematic and relevant ways. To control for these differences, a vector of control variables could be added to the model, capturing the impact of important context factors other than the policy.

$$GPA1 = \beta_0 + \gamma_1 D + \sum_{k=1}^{n} \beta_k X_k + u \qquad (2.2)$$

In the multiple linear regression (MLR) model shown by equation 2.2, the vector X includes variables for students' age, squared age, quota, field of study at university, exam group in Konkour, and Konkour test score, as well as the student-to-professor ratio in the faculty. The coefficient γ_1 estimates the association between the treatment variable (participation in single-sex classrooms) and student achievement at the end of the first-year.

Nonetheless, by looking merely at the students' outcomes at university level, equation 2.2 makes a simple cross-sectional comparison between the students of the two cohorts. From this static point of view, the estimated effect possibly suffers from bias due to the potential unobserved pre-differences between the two groups. In fact, for a consistent estimation of the effect with this approach, one needs to assume that the two groups (cohorts) do not differ in unobservable variables. The estimation is therefore not reliable if for example the average motivation level of the students differs between the two groups.

Therefore, with a dynamic approach, I look at the *transition of students from secondary to higher education.* Figure 2.2 presents a schematic diagram to illustrate the static approach versus the dynamic approach. The latter approach compares the *changes* in the achievements of each group from single-sex high schools to university, where one group attended mixed classrooms (control group) and the other participated in single-sex classes (treatment

[20] As the normal practice in Iranian higher educational institute is coeducation, I regarded the treatment as "gender separation" at classroom level, and defined the control and treatment groups accordingly. However, the control and treatment groups could simply be reversed when one considers the treatment as "mixing genders" or "participation in coeducational classes" at university level education.

group)[21]. This setting provides a classical context for using the difference-in-difference approach.

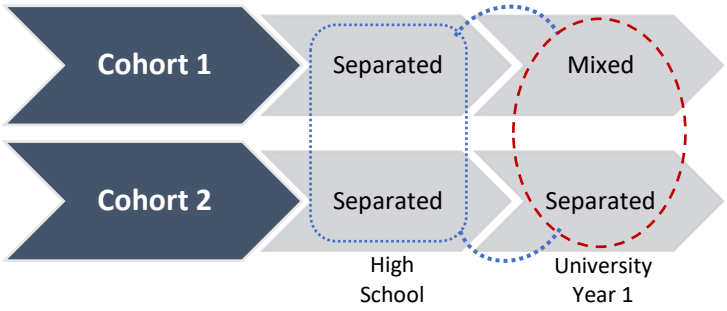

Figure 2.2. Schematic representation of static (dashed line) and dynamic (dotted line) approaches

Using the DiD approach, one would no longer need to assume that the two groups do not differ in unobservable ways. It is sufficient to suppose that the unobservable variables do not change between the two levels of study, say a highly-motivated student at secondary educational level would remain highly-motivated in higher education, which is a more plausible assumption.

The DiD estimate of the policy effect comes from the OLS estimation of the equation 2.3 as follows:

$$Y = \beta_0 + \gamma_1 D + \gamma_2 t_1 + \gamma_3 D * t_1 + \sum_{k=1}^{n} \beta_k X_k + u \qquad (2.3)$$

In this equation, the time indicator t_1 equals zero at the time of graduation from high school, and one at the end of the first year at university. The dependent variable is Y which equals the student's high school GPA in time $t_1 = 0$ and the student's first-year GPA at university in $t_1 = 1$. Again, X controls for the students' characteristics, their educational program and competencies by including the same variables as in equation 2.2. The coefficient γ_3 for the

[21] While having data from the students of a non-separated university in that specific year would have been ideal to be used as a control group for this investigation, such data was unfortunately not available. Alternatively, I chose the pre-policy entering cohort as the control group, who experienced the normal practice of mixed-gender classrooms in higher education in Iran and compared them with the post-policy entering cohort at the same university.

interaction term between the treatment variable *D* and the time indicator t_1 would then gives the DiD estimation of attending single-sex rather than coeducational classrooms in tertiary education. This coefficient allows us to infer the counterfactual test scores for the second cohort, i.e. how would the second-cohort students have performed had they not been separated by gender in their first-year classes:

$$\gamma_3 = (E[GPA \mid cohort2, university\ level] - E[GPA \mid cohort2, high\ school\ level]) \\ -(E[GPA \mid cohort1, university\ level] - E[GPA \mid cohort1, high\ school\ level])$$

Finally, to examine the heterogeneity of the effect by ability level, I define students' ability levels according to their performance percentile in the Konkour exam[22] and run separate regressions using equation 2.3 for the subgroups with different ability levels.

2.7. Results

2.7.1 Effects on student performance

Table 2.3 presents the estimates of the coefficients using OLS and DiD approaches. For the first two models, the coefficient of *D* measures the association between participation in single-sex classrooms and students' achievements. In the naïve model (first two columns), the coefficient of *D* is equivalent to the mean difference in university GPA between the two cohorts for each gender group which has been estimated using equation 2.1. This simple linear regression model gives a statistically significant positive relationship (+0.51) between single-sex education and females' outcomes and a negative association (-0.43) between all-male classrooms and males' achievements. When control variables are added in model 2, they capture part of the variations in outcomes between the two cohorts. Thus, the coefficient of *D* in the MLR model is attenuated for both genders. While females who attended single-sex classrooms on average performed 0.23 points (out of total 20.00 points) better than their counterparts of the first cohort (mixed classes), males in all-male classes underperformed males with equal characteristics but who participated in mixed-gender classrooms by 0.36 points on average.

[22] Students with upper than 75 percentile and lower than 25 percentile performances were classified as high- and low-ability group respectively, and those who performed between 25 and 50 or between 50 and 75 percentiles were categorized as of medium and upper-medium ability levels respectively.

Table 2.3. Estimated coefficients for males and females by different modeling approaches.

	(1) SLR		(2) MLR		(3) DiD without controls		(4) DiD with controls	
	Males	**Females**	**Males**	**Females**	**Males**	**Females**	**Males**	**Females**
D	-0.43***	0.51***	-0.36**	0.23***	-1.13***	-0.14	-0.91***	-0.34***
	(0.14)	(0.08)	(0.15)	(0.07)	(0.19)	(0.09)	(0.16)	(0.07)
t1					0.38**	-0.87***	0.37**	-0.87***
					(0.18)	(0.07)	(0.16)	(0.06)
D*t1					0.71***	0.65***	0.71***	0.65***
					(0.24)	(0.12)	(0.21)	(0.10)
Age			0.23*	0.18***			-0.14	-0.10*
			(0.14)	(0.05)			(0.10)	(0.06)
Age-squared			-0.00	-0.00**			0.00	0.00**
			(0.00)	(0.00)			(0.00)	(0.00)
Quota								
Medium-developed regions			-0.09	0.23***			-0.29**	0.02
			(0.16)	(0.08)			(0.13)	(0.06)
Less-developed regions			-0.06	0.44***			-0.57***	0.03
			(0.16)	(0.10)			(0.13)	(0.08)
Families of Martyrs and Veterans			-0.43	-0.53**			-0.78***	-0.82***
			(0.30)	(0.21)			(0.27)	(0.18)
Exam Group								
Mathematics and Physics			2.58***	3.27***			3.00***	2.70***
			(0.29)	(0.17)			(0.24)	(0.12)
Natural Sciences			2.02***	1.90***			2.12***	1.68***
			(0.27)	(0.16)			(0.23)	(0.12)
Foreign Languages			2.39***	2.92***			1.81***	1.91***
			(0.36)	(0.17)			(0.27)	(0.14)
Konkour Test-score			0.78***	0.93***			1.01***	0.94***
			(0.10)	(0.06)			(0.09)	(0.05)
Student-to-Professor Ratio			-0.04	-0.00			-0.03	0.07***
			(0.04)	(0.02)			1.01***	0.94***
University Field of Study	-	-	✓	✓	-	-	✓	✓
Observations	736	1908	734	1902	1468	3814	1464	3802
R-squared	0.01	0.02	0.29	0.38	0.06	0.04	0.34	0.35

Note: Own calculations based on the merged dataset (the UAT and Sanjesh data for the 2010/2011 and 2011/2012 cohorts combined). The number of observations for the first two models (1 and 2) equals the number of individuals in each group of students excluding the students with missing information. However, for the next two models with a DiD approach, each student's performance was observed twice (high school and first-year university). The number of observations is thus equivalent to student per level of study, i.e. the number of observations in the OLS models were doubled and the three students with missing high school GPAs were excluded. The reference groups for categorical covariates, Quota and Exam Group, are "highly-developed regions" and "humanities" respectively. Standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

In the next two models with DiD approach, the coefficient of D measures the impact of the differences between the two cohorts in pre-treatment period, i.e. not due to their class gender composition. The coefficient of *t1* represents a general time trend without the treatment and captures the average change in students' performances from high school to university. This baseline trend without the treatment could reflect the inherent differences in programs and exams' difficulty at the two educational levels. The coefficient of the interaction term between D and *t1* measures the improvement or decline in students' GPA that is plausibly attributable to the participation in single-sex versus mixed classrooms in first-year university courses. According to the DiD estimations, the absence of males in classrooms increases females' achievements on average by 0.65 points, which is equivalent to nearly 0.35 standard deviation. Interestingly with a DiD approach, the negative impact of all-male classrooms vanishes and turns into a positive and statistically significant effect of 0.71 points, even larger than the size of impact for females. Including additional controls in the DiD approach does not change the size and direction of the estimated effect but the model explains more variations in outcomes (larger R-squared) and only slightly improves in efficiency of the estimated impacts (smaller standard errors). In sum, the analysis in this chapter shows that participation in single-sex classrooms improves both males' and females' achievements by 0.71 and 0.65 points respectively (both different from zero with high statistical significance), which is equivalent to nearly 0.35 standard deviation for both genders.

2.7.2 Heterogeneous effects by ability

The aggregate estimations of the effect for male and female students might mask relatively significant disparities among the effects on various subgroups. As shown by distributional diagrams in figure 2.1, heterogeneity among the subgroups with different ability levels is likely. Therefore, I used a pre-treatment measure of ability level (student's performance percentile in Konkour) to categorize students as low, medium, upper-medium or high ability level, and allow the estimations to vary among the subgroups. Table 2.4 gives DiD estimations of the effects for the students with different ability, using model 4.

Table 2.4. DiD estimations for the effect of participation in single-sex classrooms by ability level.

Ability Level	High		Upper-medium		Medium		Low	
Gender	Males	Females	Males	Females	Males	Females	Males	Females
D*t1	-0.06	0.83***	1.18***	0.78***	0.37	0.63***	0.66	0.71**
	(0.40)	(0.15)	(0.44)	(0.18)	(0.37)	(0.20)	(0.46)	(0.28)
Observations	278	1038	272	1044	374	944	540	776

Note: Own calculations based on the merged dataset (the UAT and Sanjesh data for the 2010/2011 and 2011/2012 cohorts combined). The number of observations is equivalent to students per level of study (two observations for each individual student). Details on the estimated coefficients of control variables are presented in the appendix, table A.2.5. Standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

The results show that the positive impact of attending single-sex classrooms is almost the same in size and direction for females with different levels of ability, only slightly larger for those with upper-medium and high levels of ability. However, the impact is considerably heterogeneous among male groups. While male students with upper-medium ability perform remarkably better in case of participation in all-male classrooms, their counterparts at the lower or top part of the ability distribution are not affected by the gender composition of their classes at any statistically significant level.

2.7.3 Robustness checks

To provide evidence that the effect reported in this study stems from the gender composition of classrooms, I used a related though different variable- female ratio- which varies between 0 and 100 percent across all fields of study. Substituting the policy variable (*D*) with female ratio in equation 2.3 produces consistent results presented in detail in the appendix, table A.2.6. On the one hand, holding all other conditions constant, as the share of females in classroom increases by 10 percent, female students' average achievements improve by 0.4 points (statistically significant at 1% level). On the other hand, each 10 percent increase in the share of females reduces males' achievements by 0.1 points on average (statistically significant at 1% level). These results reinforce the educational benefit of single-sex classrooms for both genders.

Furthermore, the heterogeneity of the effects by students' ability level follows exactly the same pattern which is presented in the appendix, table A.2.7. While the effect of female ratio

for female students with different levels of ability is almost the same as for the whole female group (always positive with high statistical significance), only males with upper-medium ability are affected by changes in female ratio of their classrooms (with high statistical significance and relatively large negative effect for males with upper-medium ability).

Thus, the reported effects for participation in single-sex classrooms in this chapter are attributable to the policy and are unlikely due to the confounding unobserved factors associated with the cohorts (students' and instructors' motivation, class size, etc.).

In addition, as mentioned earlier in the data section (2.5), some fields of studies in the UAT had no entrant from the second cohort. According to table 2.1, the UAT did not admit any student in the academic year 2011/2012 in these fields: Statistics and Mathematics, Political Sciences, Philosophy, Library and Information Science, and Social Work. If the average university GPA of students in these fields are typically lower than the average performance in other majors, exclusion of these fields in 2011/2012 admissions would result in an overestimation of the effect of gender separation policy. To examine this potential bias in the results, I dropped all 146 individuals who were admitted in the first cohort in those fields of study from the sample, and conducted the same model for the remaining 2495 students (reduced sample). The estimated effects for males and females by model 4 then decreased to 0.64 and 0.63 points respectively (still both statistically significant at α = 1%). Therefore, ignoring the exclusion of some fields in the 2011/2012 admission process has caused only small upward bias for the estimated effects on males' and females' outcomes. Details on the estimated effects with the reduced sample are presented in the appendix, table A.2.8.

2.8. Discussion

As an attempt to uncover the causal effect of single-sex education on students' achievements, the study in this chapter benefits from a specific context of a natural experiment in which 1) no selection from the student side actually exists and 2) both treated and untreated groups studied in the same institution with the same curriculum and were taught by the same professors. Two recent studies with randomized experimental designs have also examined

similar policy effects in higher education[23] and have both of the advantages mentioned above. However, several features make the current research distinctive and relevant in our body of knowledge on the effect of single-sex education: While in their randomized control trial at a Dutch university, Oosterbeek and van Ewijk (2014) examined the effect of an increase in female ratio, this research focuses on the extreme level of gender separation at classroom level. Moreover, in the context of my research, the students were exposed to the treatment (single-sex education) for all instruction hours, contrary to the randomized experiment at a British university by Booth et al. (2018) in which the exposure was limited to only a share of tutoring hours. Additionally, the sample of students in this study provides certain benefits in terms of comprehensiveness and generalizability. Firstly, the sample included students from several fields of study, while both studies mentioned above limited their sample to students in one or two field of study (economics and business). Furthermore, most evaluations of single sex education have taken place in western cultures whereas this study examines the effect on a sample of students from a distinct cultural background, and thereby provides the possibility to generalize the results to a nearly intact context. To the best of my knowledge, this study is unique in bringing such features together.

However, application of the findings in this chapter should be made bearing in mind that they are generated from a specific context;

First, K-12 education is completely separated by gender in Iran. Although the results are in line with most previous findings in western countries, the current research does not intend to extrapolate the estimated effects to such different contexts. After all, the mechanisms could vary. For instance, university students who participated in mixed schools and classrooms might not feel uncomfortable in expressing their ideas and getting involved in class discussions in the presence of the opposite sex. Therefore, the results of this research are mainly applicable in countries where single-sex schools are dominant such as Muslim-majority countries.

[23] Oosterbeek and van Ewijk (2014), and Booth et al. (2018).

Second, the context of this research does not allow for separate investigations for the fields of studies with low versus high proportions of females[24]. The University of Allameh is predominantly specialized in humanities, language studies and social sciences, the fields mostly recognized as female-dominated majors. For nearly all fields, the proportion of female students in the dataset were between 60 to 70 percent[25].

Third, high proportions of females in all fields impose another threat to causality in the estimated effect of this study if females and males differ in average unobserved background characteristics. For example, if females on average have higher motivation, the effect reported in this chapter includes *other aspects of peer effect rather than* pure gender peer effect[26].

Fourth, the data in this research contains information for the two adjacent cohorts whose gender composition of classrooms was different only in the first year of study. For the following years of undergraduate study, variation in gender composition of classes disappeared and all students attended single-sex classrooms. Thus, I could only measure the short-term effect of participation in single-sex classrooms on the educational outcomes of students. Further data could help to provide an answer for how the educational outcomes of students who participated in mixed classrooms for the whole course of their study differ from those attending separated classrooms and uncover the effect of the policy on educational outcomes of students in the long term.

Fifth, due to the quasi-experimental design and data limitations in this research, I was not able to examine the role of other moderating factors such as dosage of exposure or socioeconomic status.

---

[24] Lavy and Schlosser (2011) found the largest positive impact of higher female proportions in cases where females constituted more than two-third of the students.

[25] Regarding how different fields are perceived in the culture of society as a male- or female-dominated major and the classic STEM categorization, only the second field "Statistics and Mathematics" could be considered as male-dominated. Unfortunately, the UAT did not admit students in this field for the second academic year. Therefore, the study is unable to distinguish the effects for different subject categories (male- vs. female-dominant) with the available data.

[26] Oosterbeek and van Ewijk (2014) refer to the same limitation in their estimation of "gross effects which also include the effect of females being different from males in other characteristics than just in their gender", an issue that "arises in all other gender peer studies as well."

Last but not least, academic performance is not the only important outcome that could be affected by the gender composition of learning environments. Whether this effect is positive or negative, for a thorough evaluation of single-sex education, policymakers should also take into account developmental and social issues and investigate the specific consequences in each dimension. If single-sex education ends up having a positive impact on academic performance but negatively affects the social and emotional development of students, decision makers who opt to implement the policy should seek additional policies and plans to compensate. Future studies with additional data on various aspects need to evaluate the impact of the policy also on key related social outcomes such as the average age of marriage, rate of divorce, time to find a job, wage, and life satisfaction of separated versus mixed university students. Therefore, although in this study the policy of gender separation at universities turned out to positively affect the educational outcomes of students, the question of whether the government should widen the scope of the policy in terms of the number of public single-sex universities or mixed universities with separated classrooms is still open.

2.9. Chapter Overview and Conclusion

Insights from the previous literature on single-sex education are mostly contaminated with self-selection bias and issues related to institutional characteristics. This chapter provided the first evaluation of single-sex education in the context of higher educational level in a Muslim-majority country, utilizing a unique natural experiment at an Iranian university, University of Allameh Tabatabaei (UAT). Using the DiD approach, I compared the pre-university and first-year-university performances of the two adjacent cohorts, one of which attended mixed classrooms and the other participated in but *had not actually selected* single-sex classrooms. Since the UAT did not pre-announce the implementation of the gender separation policy to the public, the change was unlikely to have been foreseen by the applicants. Moreover, as the two adjacent cohorts were studying in the same university with the same curriculum and taught mostly by the same faculty members, the effect found in this research is unlikely to reflect most of the unobserved differences that usually exist between single-sex and coeducational institutions.

Findings showed that separating classrooms by gender improved both males' and females' average performance by 0.37 and 0.36 standard deviation respectively. While the positive impact on females was not heterogeneous among females with different ability levels, the positive impact of all-male classrooms was mainly driven by male students with upper-medium ability level. Males with lower ability levels and those on top of the ability distribution were not affected by the gender composition of their classrooms at any statistically significant level.

The results presented in this chapter provide certain implications for shaping parallel policies to promote educational equality in Iran. The scope of the applicability of the results is not limited to the Iranian education system though. Other countries in the region which share many cultural and social factors with Iran could also use these findings while devising policies to address inequality issues and gender gaps in education and the labor market.

Furthermore, several western countries with considerable numbers of immigrants from middle-eastern countries nowadays face the problems of integration, particularly in the basic domain of education. The results of this study could also be of use in policymaking to

overcome the issues such as the gender gap in educational achievements among the immigrants from countries with similar cultural norms and values.

Appendix

Table A.2.5. Estimated coefficients of the DiD model for males and females by ability level.

Ability Level	**High**		**Upper-medium**		**Medium**		**Low**	
Gender	Males	Females	Males	Females	Males	Females	Males	Females
D	-0.52	-0.81***	-1.23***	-0.49***	-0.78***	-0.51***	-0.80**	-0.03
	(0.34)	(0.10)	(0.37)	(0.14)	(0.29)	(0.15)	(0.36)	(0.23)
t1	-0.31	-1.71***	0.34	-0.91***	0.43	-0.45***	1.02**	-0.38***
	(0.33)	(0.10)	(0.27)	(0.11)	(0.27)	(0.13)	(0.41)	(0.14)
D*t1	-0.06	0.83***	1.18***	0.78***	0.37	0.63***	0.66	0.71**
	(0.40)	(0.15)	(0.44)	(0.18)	(0.37)	(0.20)	(0.46)	(0.28)
Age	-0.72	-0.85**	-0.18	0.04	-1.00***	-0.50***	0.13	0.03
	(0.96)	(0.36)	(0.30)	(0.12)	(0.31)	(0.19)	(0.14)	(0.07)
Age-squared	0.01	0.02**	0.00	-0.00	0.02***	0.01**	-0.00	0.00
	(0.02)	(0.01)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)
Quota								
Medium-developed regions	-0.16	0.09	-0.36	0.17	-0.10	-0.09	-0.12	-0.12
	(0.25)	(0.10)	(0.32)	(0.11)	(0.28)	(0.14)	(0.26)	(0.17)
Less-developed regions	0.03	0.27**	-0.11	0.14	-0.40	-0.08	-1.05***	-0.17
	(0.29)	(0.13)	(0.29)	(0.14)	(0.36)	(0.19)	(0.23)	(0.18)
Families of Martyrs and Veterans		0.23	-0.04	-0.68	-0.57	-0.15	-0.90**	-1.39***
		(0.41)	(0.46)	(0.72)	(0.82)	(0.48)	(0.36)	(0.29)
University Field of Study								
Theology and Islamic Knowledge			-0.46	0.26	1.69***	0.88**	-1.16*	1.43**
			(0.86)	(0.37)	(0.65)	(0.37)	(0.68)	(0.60)
Statistics and Mathematics					-0.91		-0.48	-1.01***
					(1.40)		(0.79)	(0.36)
Accounting	-0.36	0.39**	-1.76***	0.03	0.54	0.40**	0.26	-1.47
	(0.44)	(0.19)	(0.54)	(0.34)	(0.33)	(0.18)	(0.49)	(1.01)
Laws	0.00	0.21	1.41*	-0.03	1.65*	-1.17		
	(0.58)	(0.19)	(0.79)	(0.81)	(0.86)	(0.86)		
Guidance and Counseling	0.87	1.21***	0.61	-0.69	0.89	0.81**	0.27	-0.57
	(0.66)	(0.17)	(0.59)	(0.58)	(0.82)	(0.35)	(0.54)	(0.45)
Public Relations		0.85***	1.00	-0.19	0.80	0.81	0.51	-1.73
		(0.29)	(0.90)	(0.41)	(1.54)	(0.82)	(1.15)	(1.09)
Psychology	0.92*	0.71***	-0.35	0.22	1.07*	0.50*	-0.62	0.55
	(0.47)	(0.20)	(0.97)	(0.39)	(0.59)	(0.28)	(0.90)	(0.98)
Journalism	0.13	1.18***	0.26	-0.16	1.14	-0.33	0.15	-1.63
	(0.74)	(0.29)	(0.83)	(0.42)	(1.63)	(0.91)	(1.13)	(1.51)
Language and Literature	0.22	0.29	2.11***	-0.71**	0.84	-0.61*	-0.88	0.47
	(0.61)	(0.21)	(0.52)	(0.31)	(0.54)	(0.36)	(0.57)	(0.68)
Social Sciences	1.21*	0.45**	0.32	-0.58**	0.78	-0.15	-1.18*	0.67
	(0.62)	(0.21)	(0.45)	(0.28)	(0.48)	(0.38)	(0.61)	(0.47)
Economics	1.44***	-0.55**	-1.43***	-1.88**	-0.16	-0.13	-0.16	-0.22
	(0.31)	(0.22)	(0.53)	(0.75)	(0.50)	(0.18)	(0.36)	(0.20)
Educational Sciences		1.08***	1.21	0.13	0.66	0.22	0.25	-0.17
		(0.25)	(0.90)	(0.32)	(1.05)	(0.35)	(0.52)	(0.41)
Political Sciences	0.91	1.04*	0.33	-0.13		1.09		
	(0.97)	(0.59)	(0.57)	(0.36)		(0.76)		

Table A.2.5 - continued. Estimated coefficients of the DiD model for males and females by ability level.

Ability Level	**High**		**Upper-medium**		**Medium**		**Low**	
Gender	Males	Females	Males	Females	Males	Females	Males	Females
Philosophy		0.98***	0.16	-0.30	0.93	0.62	-1.11	-0.62
		(0.35)	(1.35)	(0.37)	(0.59)	(0.57)	(0.98)	(1.01)
Library and Information Science				0.98*		0.70		-0.51
				(0.55)		(0.53)		(0.53)
Social Work	0.93**	0.91***		0.06		0.09		2.01***
	(0.39)	(0.24)		(0.37)		(0.55)		(0.49)
ECO College of Insurance	-0.48	-0.64***	-0.26	-0.09	0.21	0.01	-0.03	0.28
	(0.48)	(0.23)	(1.75)	(0.61)	(0.53)	(0.19)	(0.49)	(0.52)
Exam Group								
Mathematics and Physics			3.54***		3.79***	2.90***	1.77***	2.13***
			(0.73)		(0.54)	(0.32)	(0.56)	(0.41)
Natural Sciences			2.76**	1.44***	2.67***	2.05***	1.42***	1.26***
			(1.10)	(0.29)	(0.51)	(0.33)	(0.49)	(0.39)
Foreign Languages	0.54	1.03***	-0.70	1.49***	1.91***	2.13***	2.80***	1.33**
	(0.91)	(0.33)	(0.45)	(0.21)	(0.65)	(0.27)	(0.48)	(0.62)
Konkour Test-score	1.22***	1.19***	1.25*	1.39***	0.88	1.50***	1.20***	0.61***
	(0.31)	(0.12)	(0.72)	(0.30)	(0.67)	(0.36)	(0.23)	(0.16)
Student-to-Professor Ratio	-0.01	-0.03	-0.02	-0.00	0.07	0.02	-0.10	0.17**
	(0.12)	(0.03)	(0.09)	(0.03)	(0.13)	(0.04)	(0.09)	(0.07)
constant	21.05**	23.61***	14.37***	13.07***	23.50***	18.98***	11.23***	13.05***
	(9.98)	(3.81)	(4.21)	(1.76)	(4.15)	(2.20)	(1.96)	(1.13)
Observations	278	1038	272	1044	374	944	540	776
R-squared	0.21	0.41	0.25	0.22	0.42	0.38	0.35	0.37

Note: Own calculations based on the merged dataset (the UAT and Sanjesh data for the 2010/2011 and 2011/2012 cohorts combined). The reference groups for categorical covariates, Quota, Exam Group, and University Field of Study, are "highly-developed regions", "humanities", and "management" respectively. The number of observations is equivalent to student per level of study (two observations for each individual student). Standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A.2.6. Estimated coefficients of the DiD model female-ratio variable.

	DiD with Female-Ratio Variable	
	Males	Females
Female ratio	0.01***	-0.02***
	(0.00)	(0.00)
t1	1.08***	-3.95***
	(0.13)	(0.41)
Female ratio*t1	-0.01***	0.04***
	(0.00)	(0.00)
Age	-0.14	-0.10*
	(0.10)	(0.05)
Age-squared	0.00	0.00**
	(0.00)	(0.00)
Quota		
Medium-developed regions	-0.29**	0.02
	(0.13)	(0.06)
Less-developed regions	-0.57***	0.03
	(0.13)	(0.08)
Families of Martyrs and Veterans	-0.78***	-0.81***
	(0.27)	(0.17)
University Field of Study		
Theology and Islamic Knowledge	0.15	1.02***
	(0.37)	(0.18)
Statistics and Mathematics	-0.76	-1.66***
	(0.61)	(0.29)
Accounting	0.23	0.46***
	(0.22)	(0.12)
Laws	0.54**	0.73***
	(0.27)	(0.16)
Guidance and Counseling	0.88***	0.69***
	(0.27)	(0.13)
Public Relations	1.12**	-0.13
	(0.47)	(0.18)
Psychology	0.79***	0.42***
	(0.27)	(0.12)
Journalism	0.67	-0.07
	(0.44)	(0.20)
Language and Literature	0.12	-0.09
	(0.26)	(0.13)
Social Sciences	0.30	0.05
	(0.23)	(0.12)
Economics	-0.29	-0.31***
	(0.25)	(0.11)
Educational Sciences	0.78**	0.25*
	(0.32)	(0.13)
Political Sciences	0.74*	0.82***
	(0.40)	(0.24)
Philosophy	-0.08	0.44*
	(0.71)	(0.24)

Table A.2.6 – continued. Estimated coefficients of the DiD model female-ratio variable.

	DiD with Female-Ratio Variable	
	Males	Females
Library and Information Science		0.22
		(0.23)
Social Work	1.57***	0.81***
	(0.29)	(0.18)
ECO College of Insurance	0.13	-0.21
	(0.29)	(0.15)
Exam Group		
Mathematics and Physics	3.00***	2.71***
	(0.24)	(0.12)
Natural Sciences	2.12***	1.69***
	(0.23)	(0.12)
Foreign Languages	1.81***	1.91***
	(0.27)	(0.14)
Konkour Test-score	1.01***	0.95***
	(0.09)	(0.05)
Student-to-Professor Ratio	-0.04	0.07***
	(0.04)	(0.01)
constant	13.46***	16.75***
	(1.36)	(0.78)
Observations	1464	3802
R-squared	0.34	0.35

Note: Own calculations based on the merged dataset (the UAT and Sanjesh data for the 2010/2011 and 2011/2012 cohorts combined). The reference groups for categorical covariates, Quota, Exam Group, and University Field of Study, are "highly-developed regions", "humanities", and "management" respectively. The number of observations is equivalent to student per level of study (two observations for each individual student). Standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A.2.7. Estimated coefficients of the DiD model with female-ratio variable by ability level.

Ability Level	**High**		**Upper-medium**		**Medium**		**Low**	
Gender	Males	Females	Males	Females	Males	Females	Males	Females
Female ratio	0.01	-0.04***	0.02***	-0.02***	0.01**	-0.03***	0.01**	-0.01
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.01)	(0.00)	(0.01)
t1	-0.37*	-5.05***	1.53***	-3.40***	0.78***	-3.96***	1.67***	-4.58***
	(0.23)	(0.65)	(0.34)	(0.75)	(0.26)	(0.79)	(0.20)	(1.15)
Female ratio*t1	0.00	0.04***	-0.02***	0.03***	-0.00	0.04***	-0.01	0.05***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)	(0.01)
Age	-0.72	-0.87**	-0.18	0.04	-1.00***	-0.50***	0.13	0.04
	(0.96)	(0.36)	(0.30)	(0.13)	(0.31)	(0.18)	(0.14)	(0.07)
Age-squared	0.01	0.02**	0.00	-0.00	0.02***	0.01**	-0.00	0.00
	(0.02)	(0.01)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)
Quota								
Medium-developed regions	-0.16	0.09	-0.36	0.17	-0.11	-0.09	-0.12	-0.11
	(0.25)	(0.09)	(0.32)	(0.11)	(0.28)	(0.13)	(0.26)	(0.17)
Less-developed regions	0.04	0.28**	-0.11	0.14	-0.41	-0.09	-1.05***	-0.16
	(0.29)	(0.13)	(0.29)	(0.14)	(0.36)	(0.19)	(0.23)	(0.18)
Families of Martyrs and Veterans		0.22	-0.06	-0.68	-0.57	-0.17	-0.91**	-1.37***
		(0.41)	(0.46)	(0.70)	(0.82)	(0.48)	(0.36)	(0.28)
University Field of Study								
Theology and Islamic Knowledge			-0.41	0.23	1.82***	0.80**	-1.15*	1.62***
			(0.86)	(0.38)	(0.64)	(0.37)	(0.68)	(0.61)
Statistics and Mathematics					-0.85		-0.46	-0.89**
					(1.41)		(0.79)	(0.36)
Accounting	-0.33	0.26	-1.71***	-0.02	0.57*	0.35**	0.28	-1.44
	(0.44)	(0.19)	(0.56)	(0.33)	(0.33)	(0.18)	(0.50)	(1.06)
Laws	0.03	0.19	1.41*	-0.05	1.70**	-1.21		
	(0.58)	(0.20)	(0.79)	(0.80)	(0.85)	(0.78)		
Guidance and Counseling	0.84	1.19***	0.63	-0.69	0.89	0.83**	0.27	-0.64
	(0.66)	(0.17)	(0.59)	(0.58)	(0.82)	(0.35)	(0.54)	(0.45)
Public Relations		0.75***	1.06	-0.24	0.77	0.77	0.55	-1.71
		(0.28)	(0.90)	(0.40)	(1.53)	(0.81)	(1.15)	(1.07)
Psychology	0.91*	0.64***	-0.32	0.18	1.07*	0.47*	-0.61	0.57
	(0.47)	(0.19)	(0.97)	(0.39)	(0.59)	(0.27)	(0.90)	(0.99)
Journalism	0.20	0.94***	0.37	-0.25	1.12	-0.45	0.21	-1.43
	(0.76)	(0.28)	(0.82)	(0.40)	(1.62)	(0.89)	(1.14)	(1.72)
Language and Literature	0.23	0.26	2.12***	-0.73**	0.86	-0.63*	-0.88	0.54
	(0.61)	(0.21)	(0.52)	(0.31)	(0.54)	(0.36)	(0.57)	(0.76)
Social Sciences	1.22**	0.42**	0.34	-0.61**	0.80	-0.18	-1.17*	0.72
	(0.62)	(0.21)	(0.45)	(0.28)	(0.48)	(0.39)	(0.61)	(0.46)
Economics	1.47***	-0.57***	-1.43***	-1.89**	-0.15	-0.14	-0.16	-0.19
	(0.31)	(0.22)	(0.53)	(0.75)	(0.49)	(0.18)	(0.36)	(0.20)
Educational Sciences		1.05***	1.19	0.11	0.60	0.23	0.25	-0.24
		(0.24)	(0.89)	(0.32)	(1.05)	(0.35)	(0.52)	(0.40)
Political Sciences	0.95	1.00	0.34	-0.14		1.06		
	(0.97)	(0.61)	(0.57)	(0.37)		(0.67)		
Philosophy		1.09***	0.10	-0.28	0.91	0.67	-1.16	-0.67
		(0.31)	(1.36)	(0.37)	(0.60)	(0.58)	(0.99)	(1.03)

Table A.2.7 - continued. Estimated coefficients of the DiD model with female-ratio variable by ability level.

Ability Level	**High**		**Upper-medium**		**Medium**		**Low**	
Gender	Males	Females	Males	Females	Males	Females	Males	Females
Library and Information Science				1.02*		0.80		-0.78
				(0.59)		(0.53)		(0.60)
Social Work	0.86**	1.11***		0.10		0.17		1.84***
	(0.40)	(0.23)		(0.37)		(0.56)		(0.59)
ECO College of Insurance	-0.49	-0.64***	-0.22	-0.08	0.21	0.02	-0.02	0.24
	(0.48)	(0.24)	(1.75)	(0.65)	(0.53)	(0.18)	(0.49)	(0.54)
Exam Group								
Mathematics and Physics			3.55***		3.79***	2.88***	1.76***	2.15***
			(0.73)		(0.54)	(0.31)	(0.56)	(0.41)
Natural Sciences			2.76**	1.44***	2.68***	2.03***	1.42***	1.28***
			(1.09)	(0.30)	(0.51)	(0.33)	(0.49)	(0.38)
Foreign Languages	0.56	1.02***	-0.70	1.49***	1.90***	2.12***	2.80***	1.32*
	(0.91)	(0.34)	(0.45)	(0.21)	(0.65)	(0.27)	(0.48)	(0.71)
Konkour Test-score	1.21***	1.19***	1.24*	1.39***	0.88	1.48***	1.21***	0.61***
	(0.31)	(0.12)	(0.72)	(0.30)	(0.67)	(0.36)	(0.23)	(0.16)
Student-to-Professor Ratio	-0.01	-0.01	-0.03	0.01	0.07	0.03	-0.10	0.17**
	(0.12)	(0.03)	(0.09)	(0.03)	(0.13)	(0.04)	(0.09)	(0.07)
constant	20.58**	26.95***	13.18***	14.66***	22.68***	21.39***	10.47***	13.82***
	(9.92)	(3.90)	(4.20)	(1.92)	(4.14)	(2.17)	(1.90)	(1.39)
Observations	278	1038	272	1044	374	944	540	776
R-squared	0.21	0.41	0.25	0.22	0.42	0.39	0.35	0.38

Note: Own calculations based on the merged dataset (the UAT and Sanjesh data for the 2010/2011 and 2011/2012 cohorts combined). The reference groups for categorical covariates, Quota, Exam Group, and University Field of Study, are "highly-developed regions", "humanities", and "management" respectively. The number of observations is equivalent to student per level of study (two observations for each individual student). Standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A.2.8. Estimated coefficients of the DiD model for the reduced sample.

	Males	Females
D	-0.87***	-0.33***
	(0.16)	(0.08)
t1	0.45***	-0.85***
	(0.17)	(0.07)
D*t1	0.64***	0.63***
	(0.21)	(0.10)
Age	-0.13	-0.10*
	(0.11)	(0.06)
Age-squared	0.00	0.00**
	(0.00)	(0.00)
Quota		
Medium-developed regions	-0.31**	-0.00
	(0.14)	(0.06)
Less-developed regions	-0.59***	0.04
	(0.14)	(0.09)
Families of Martyrs and Veterans	-0.79***	-0.89***
	(0.27)	(0.18)
University Field of Study		
Theology and Islamic Knowledge	0.10	1.04***
	(0.37)	(0.18)
Accounting	0.21	0.47***
	(0.22)	(0.12)
Laws	0.54**	0.79***
	(0.27)	(0.16)
Guidance and Counseling	0.88***	0.67***
	(0.27)	(0.13)
Public Relations	1.08**	-0.17
	(0.47)	(0.18)
Psychology	0.77***	0.42***
	(0.27)	(0.13)
Journalism	0.61	-0.10
	(0.44)	(0.21)
Language and Literature	0.11	-0.09
	(0.26)	(0.13)
Social Sciences	0.29	0.05
	(0.23)	(0.12)
Economics	-0.29	-0.30***
	(0.25)	(0.11)
Educational Sciences	0.79**	0.22*
	(0.32)	(0.13)
ECO College of Insurance	0.12	-0.22
	(0.29)	(0.15)
Exam Group		
Mathematics and Physics	2.99***	2.70***
	(0.24)	(0.12)
Natural Sciences	2.12***	1.67***
	(0.23)	(0.12)

Table A.2.8 - continued. Estimated coefficients of the DiD model for the reduced sample.

	Males	**Females**
Foreign Languages	1.80***	1.90***
	(0.27)	(0.14)
Konkour Test-score	1.00***	0.92***
	(0.09)	(0.05)
Student-to-Professor Ratio	-0.03	0.07***
	(0.04)	(0.02)
constant	14.22***	14.96***
	(1.38)	(0.75)
Observations	1422	3552
R-squared	0.35	0.36

Note: Own calculations based on the merged dataset (the UAT and Sanjesh data for the 2010/2011 and 2011/2012 cohorts combined). The reference groups for categorical covariates, Quota, Exam Group, and University Field of Study, are "highly-developed regions", "humanities", and "management" respectively. The number of observations is equivalent to student per level of study (two observations for each individual student). Standard errors are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Chapter 3

Class Gender Composition and Student Math Achievement: An International Comparison using TIMSS data

3.1. Introduction

Does it matter who you sit next to in class? Does the gender composition of peers in the classroom affect how much you learn and how you perform in exams? In theory, previous literature on the gender peer effect points out some of the possible channels through which the impact could operate[27]. Empirically however, it has been shown that the answer largely depends on the context.

[27] Different channels of gender peer effect are active through various levels of education. As this study investigates the effect at secondary educational level, the following theory section merely includes some of the mechanisms potentially operating during teen-age and at secondary level of study. Thus, the channels of the impact mentioned here illustrate but certainly do not exhaust all possible mechanisms at work.

3.1.1 Theory and possible mechanisms

Most studies on the gender peer effect attribute the impact to behavioral differences between genders in the learning environment (eg. Lavy & Schlosser, 2011; Hayes et al., 2011). For example, male students tend to behave more disruptively while female students are more supportive and obedient. In this sense, higher proportions of females in classrooms promote a better atmosphere in the learning environment, induce less teacher fatigue and burn-out and thereby improve students' performance (Lavy & Schlosser, 2011; Oosterbeek & van Ewijk, 2014).

The second mechanism frequently addressed in studies at secondary and higher educational levels relates to the *stereotype threat theory*[28] (Steele & Aronson, 1995). By the time of entering secondary education, children start to identify themselves as a member of a group, and develop stereotype consciousness. The stereotypical beliefs about gender-specific competencies shape children's intellectual identity and influence their academic performance (Steele, 1997). While several empirical studies maintain the role of class gender composition in the degree of activation and endorsement of gender stereotypes in classroom (eg. Bigler & Liben, 2006; Huguet & Régner, 2007; Kessels & Hannover, 2008; Sullivan et al., 2010), the extent and direction of the influence is a matter of debate among scholars. On the one hand, proponents of single-sex education argue that in such a learning environment, gendered views about educational competencies are less endorsed, leading to reduced psychological salience of gender (Sullivan et al., 2010; Schneeweis & Zweimuller, 2012). In the presence of same-sex high-performing role models in all subjects (Park et al., 2018) students have less access to gender-related self-knowledge and thereby are less likely to hold stereotypes against their ability[29], and thus more likely to have a better self-concept in gender-atypical subjects (Kessels & Hannover, 2008). On the other hand, opponents of single-sex education argue that sex-segregation could even increase gender salience and reinforce stereotypes (Halpern et al, 2011; Bigler et al, 2014). In their counter-argument, they refer to *development*

[28] The theory posits that a self-evaluative threat can beset the members of any group about whom a negative stereotype exists (Steele & Aronson, 1995).

[29] McCoy et al. (2012) endorses this view empirically by reporting that while in general boys have more positive attitude towards math than girls do, those girls and boys educated in single-sex environment do not differ in terms of math attitude.

intergroup theory (Bigler & Liben, 2006), that emphasizing gender can lead to essentialist thought, in-group favoritism, and out-group bias (Pahlke & Hyde, 2016). Accordingly, they believe that the mere name of a single-sex school or simply calling the student audiences by gendered words in these contexts (like "Good morning, ladies") act as gender labels (Bigler et al., 2014), and thus, single-sex settings actually reinforce gender stereotyping (Halpern et al, 2011; Pahlke & Hyde, 2016). A related but less-frequently addressed mechanism is that teachers who hold stereotypical views against a certain gender's competencies might dumb down their teaching level when students of that gender constitute a higher share in the classroom (van Ewijk & Sleegers, 2010).

Another channel for the gender peer effect refers to the adolescent culture and sexual attraction among genders of the opposite sex. The presence of the opposite sex in the classroom or a larger number of potentially interesting members of the opposite sex in class (Oosterbeek and van Ewijk, 2014) might cause more distractions and less focus on learning tasks, leading to relatively poor performance in exams (Riordan, 1985; Chadwell, 2010). In contrast, single-sex education restricts the access to individuals of the opposite gender for much of the day (Bigler et al, 2014), leading to relatively less obsession about appearance and sexual attractiveness (Riordan, 1985; Dyer and Tiggemenn, 1996), fewer engagement in heterosexual relationships, and more focus on school activities (Bigler et al, 2014).

3.1.2 Literature review and knowledge gap

Generally, two strands of research pertain to gender peer effect with the first evaluating binary cases of single-sex versus coeducation and the second assessing non-extreme cases of varying gender shares in mixed settings.

The large and growing body of research on the evaluation of single-sex education has produced inconsistent results, ranging from a statistically significant positive effect to a null or even negative impact on student achievement[30]. The large variation among the estimated effects was mainly attributed to methodological issues and selection bias (Jackson, 2012; Pahlke et al., 2014). Since most studies did not account for non-random selection of students into single-sex and coeducational settings, they confounded potential institutional differences

[30] See Pahlke et al. (2014) for a concise review of studies on single-sex schooling in different countries.

between either types of school or systematic differences among their entrants with the impact of class gender composition.

Nevertheless, methodological issues do not solely constitute the underlying reason behind the mixed and sometimes contradictory results on the effect of single-sex education. The findings are also inconsistent across the studies that resolved the selection issues by conducting a randomized control trial or using a natural experiment design. For example, utilizing a natural experiment in Switzerland, Eisenkopf et al. (2015) found a statistically significant positive impact of single-sex schooling on females' math achievement. Park et al. (2018) used the random assignment of students to schools in South Korea to evaluate the effect of single-sex schooling on students' performance in STEM subjects. They found a statistically significant and positive effect for males but no statistically significant effect for females. A recent review of more than 180 studies on single-sex schooling by Pahlke et al. (2014) also showed that the estimated effects, even produced by causal investigations at the same level of education, differed across various countries. The contrast in these findings could be an indication of the prominent role of differing contexts across countries. In their early cross-country analysis of single-sex education, Baker et al. (1995) addressed the "national context" as a reason for the different estimations across countries.

Despite the extensive literature in the former strand, only a few studies have assessed the impact of varying gender shares as a continuous variable in mixed classrooms. With an innovative approach to address the potential bias caused by omitted variables such as ability, Hoxby (2000) exploited the idiosyncratic variation of gender shares between two adjacent cohorts within the schools in Texas. She argued that this variation mainly stemmed from random fluctuations in birth gender ratios and was unlikely to be controlled by parents and school authorities. Therefore, she interpreted the positive estimated impact of an increasing female ratio in the classroom on both girls' and boys' test scores as causal. Nevertheless, the context of her research - Texas elementary public schools - restricted the possibility of generalizing her results to students at higher educational levels or from different settings. Two later studies by Whitmore (2005), who utilized the random assignment of students to classrooms in Tennessee, and by Kramarz, et al. (2008) who used the information of students' mobility across primary schools in England, also found a positive impact of having more

females in classrooms on students' achievements. More recently, using the same methodological approach as Hoxby's (2000), Lavy and Schlosser (2011) studied the gender peer impact at all three levels of schooling (elementary, middle, and high school) in Israeli schools. Their results also confirmed the positive effect of higher female proportions in classrooms on girls' and boys' academic achievements. At higher educational levels, in a randomized experiment at the University of Amsterdam, Netherland, Oosterbeek and van Ewijk (2014) manipulated the gender ratios in study working groups and found no substantial effect of increasing the proportion of females on student achievement.

Previous studies also suggest a nonlinear effect of the gender composition in classrooms on student achievement (Hoxby, 2000; Lavy and Schlosser, 2011). While the exact threshold is controversial[31], studies generally indicate that the impact rises when girls constitute the majority in class. This result is intuitively comprehensible since most studies on the underlying mechanisms attributed the effect to improvements in classroom atmosphere and a better learning environment and that changes in the overall classroom atmosphere is most likely nonlinear.

As reviewed above, most of the existing studies on the gender interaction effect focused on western countries where stereotypical views on gender roles are less pronounced. In this regard, the context of developing counties have remained nearly untouched, most probably due to the lack of adequate data on confounding factors from nationally representative samples. Given that the channels of the effect and the dominant mechanisms could vary markedly across nations with divergent cultural backgrounds[32], the existing results are hardly carried over into the unexamined countries. Additionally, although the gender composition in the learning environment has been shown to affect students' performance in different contexts (Hoxby, 2000; Lavy and Schlosser, 2011; Park et al., 2013; Eisenkopf et al., 2015),

[31] For example, Lavy and Schlosser (2011) maintained that the impact was largest in classrooms where the female proportion exceeded 58.7%. Hoxby (2000) found suggestive evidence for the largest peer effects among the cohorts with female shares above 66%. However, the results acquired by Oosterbeek and van Ewijk (2014) did not confirm the nonlinearity of the impact, probably as they mentioned, due to the fact that such high share of females did not exist in their sample from the economics curricula at the university.

[32] In their investigation of PISA 2003 achievement gap between girls and boys across the countries, Guiso, et al. (2008) suggested the relevance of country-level and cultural background variables such as cultural attitudes towards women, female economic activity, and women's political empowerment.

to the best of my knowledge, cross-country analyses of international test results have not yet focused on this class-level variable as a driving factor[33]. Therefore, the predominant focus of gender peer effect studies on limited number of countries with certain cultural background, along with the lack of international comparisons focusing on the gender peer effect pinpoint a clear gap in the literature, *the dearth of cross-country research on the gender peer effect.*

This chapter contributes to the existing literature on the *gender peer effect on student achievement* via a cross-country analysis of TIMSS 2015 international data. The TIMSS dataset provides rich information on student performance and background factors at individual, class and school levels. Using TIMSS results also enables to focus on students' performance in mathematics as a discipline most likely to be associated with students' future academic and career success (Claessens et al., 2009; Lubinski et al., 2014).

In order to examine the impact of the gender composition in classrooms in this study, I conduct two separate but related analyses. The first analysis focuses on *the impact of varying female proportions in coeducational classrooms on students' math achievements* for the 37 TIMSS2015 participating countries with considerable numbers of students in coeducation. Besides the model generated for the pooled sample of students in all 37 participating countries, country-specific models are also constructed. More specifically, the first analysis examines the following hypotheses:

- H1) *The higher the proportion of females in a mixed classroom, the better both male and female students perform in mathematics.*
- H2) *The association between the ratio of females in class and students' math scores differs across countries.*
- H3) *The impact of varying proportions of females in class on students' math scores follows a nonlinear pattern, increasing towards the right end of its range.*

[33] Existing international comparisons on determinants of educational outcomes have predominantly focused on the role of student-level factors such as socio-economic and migration background (Hanushek & Woessmann, 2010; Hanushek et al., 2014), or on the effect of school- and class-level variables such as class size (Woessmann, 2005; Woessmann & West, 2006), teacher characteristics and teaching styles (Schwerdt & Wuppermann, 2011; Lee & Huh, 2014; Caponera & Losito, 2016; Hanushek & Woessmann, 2017; Eriksson et al., 2018).

The second analysis focuses on *the effect of participation in single-sex versus mixed classrooms on students' math achievements*[34]. The sample in the second analysis is confined to 17 TIMSS2015 participating countries in which sizable number of students attend either type of education. Two hypotheses are tested in the second analysis:

- H4) *Students who attend single-sex classrooms outperform their counterparts in mathematics. The effect is especially positive for female students.*
- H5) *The impact of participation in single-sex classrooms on student achievement is different across various countries.*

While the findings about single-sex education impact is mixed, the hypothesis H4 is formulated based on the relatively dominant results that females academically benefit from single-sex classes while males perform equally well or even worse in all-male classes (eg. Adkinson, 2008; Lee & Bryk, 1986, Laster, 2004; Santos et al., 2013), and that an increased female ratio in classroom has academic benefit for both genders (Hoxby, 2000; Lavy & Schlosser, 2011).

3.2. Data

In this study, I use TIMSS 2015 data of eighth-grade students. The Trends in International Mathematics and Science Study (TIMSS) is an international assessment of student academic achievement in math and science conducted by the International Association for the Evaluation of Educational Achievement (IEA) in four-year cycles as from 1995. In order to measure students' performance in mathematics and describe student achievement at the population level for participating countries, TIMSS uses an elaborate matrix-sampling design. A sample of schools are drawn with probabilities proportional to their size and one or two intact classes of students are selected from each sampled school. The inclusion of

[34] Single-sex education might be considered as an extreme case of classroom gender composition. However, the cross-sectional TIMSS data used in this study is unable to take into account the potentially large systematic differences between single-sex and coeducational schools in most countries. As a result, including single-sex schools/classrooms in the analysis of female-ratio effect would confound the effect of several other factors with the impact of interest. Therefore, in the first analysis, I followed the approach previously used by Lavy and Schlosser (2011) and excluded all single-sex classes. Additionally, I conducted a separate analysis to compare the achievements of students participating in single-sex classes with those of the students attending mixed classrooms.

intact classes in the sample makes TIMSS - rather than other large-scale international datasets such as PISA[35] - well-suited for the purpose of this study, which considers the gender composition of classrooms as the main variable of interest. To examine the impact of varying gender ratios in classrooms (analysis1), a sample of 37 participating countries with an abundant number of students in coeducation has been used (sample 1)[36]. For the investigation of single-sex versus mixed classrooms (analysis2), 17 countries in which the share of students in single-sex education were between 10 to 90 percent of the country sample size, were included (sample 2).

To collect the data more efficiently, TIMSS utilizes complex procedures to distribute achievement items to sampled students. Each student is supposed to respond to only a subset of the full test. Using multiple imputation methodology, a set of five plausible values are then generated for each student according to Item Response Theory (IRT). Thus, the plausible values show the set of test scores that the student might have obtained if responding to the whole math item pool (which consists of around 200 items). As the dependent variable in the present study, I used all of the five plausible values provided by TIMSS as individual-level math scores[37] as recommended by Foy and Yin (2015).

To measure the gender composition in classroom, I relied on the sampling design of TIMSS regarding the selection of *intact classes* within each selected school. The students with the same class ID in each country's data were grouped together as one classroom. The class size and female ratio in each classroom were calculated accordingly[38].

---

[35] Program for International Student Assessment

[36] From the 40 participating countries in TIMSS 2015, I excluded Saudi Arabia, Iran, and Jordan from the first analysis due to the governmental obligation for offering single-sex education in these countries. In none of these three countries the share of students in coeducation exceeded 3% of the country sample size.

[37] As investigated by Carstens and Hastedt (2010), the improper use of plausible values such as averaging or using one of them by random, will likely produce biased and inaccurate results.

[38] Class size and number of students with certain sex in class have not been asked from individual students in TIMSS questionnaires. There is however, a question on class size in the teachers' questionnaire, for which a substantial number of teachers in most countries did not provide an answer. According to the TIMSS 2015 report on methods and procedures (Martin et al., 2016), more than 90% of eligible students in most countries attended a TIMSS test session, and in none of the participating countries did the share of absent students exceed 11%. Therefore, the calculated female ratio for individual schools could be regarded as fairly free of measurement error. The classes with lower than 4 members were excluded from the analysis (only lower than

In addition to evaluating student academic performance, TIMSS administers a set of context questionnaires to students, their math and science teachers and their school principals, collecting information about students' home background and school environment[39]. To ensure that the estimation does not confound with the major predictors of student performance, additional controls were derived from the related TIMSS questionnaires. Following the literature on *determinants of international educational achievement*[40], control variables from distinct levels of student, class, and school are used as inputs in the education production function.

Family and home background: Students' characteristics and family background, particularly socio-economic status (SES), proved to be the key influencing factor in the international education production function (Yang, 2003; Hanushek & Woessmann, 2010; Hanushek et al., 2019). Most of the international educational assessments used the number of books at home as a valid proxy for the student's SES (Woessman, 2008). In this study, I used a broader variable - home educational resources - which has been constructed by combining important factors of educational, social and economic background of students' families[41] (Hanushek & Woessmann, 2010). As a proxy for the immigration status of students, I used a dichotomized variable indicating whether the student *always or often* speaks the language of the test at home or *only sometimes or never* does so[42]. Moreover, despite the small variation in the age of eighth-graders, I controlled for student's age (available in two-decimal points in the data) as it could correlate with the main variables in this study (female ratio or single-sex dummy),

300 individuals in the whole sample) as the female ratio for these outliers were much more prone to measurement error in the case of student's absence in the test session.

[39] More details on the sampling design, imputation methods, context variables and the generation of plausible values in TIMSS database are available in Foy and Yin (2015) and Martin et al. (2016).

[40] For a broad review of contextual factors and determinants of international student achievement, see Hanushek and Woessmann (2011), and Martin et al. (2013).

[41] This composite variable has been constructed using several items such as the number of books at home, and whether the student has access to their own room, study desk, computer, laptop, Internet connection, etc. at home (Martin et al., 2016).

[42] There might exist national ethnic minorities with different dialects who do not speak the language of test at home. However, as the proportion of the observations with missing information about the place of own/parents' birth was high, I merely used the language dummy to account for migration background. After all, the focus of this study is not on the effect of migration or language proficiency.

for instance if older students or repeaters of a grade tended to be assigned to classes with a certain gender composition.

Teacher and class characteristics: At the second level, variables on *teacher's gender, level of education, and years of experience*, as well as *class size* were included among the controls. The non-response rate for teacher education level was rather high (exceeding 15 percent of the data in some countries such as the UAE, Bahrain and Egypt), and quite unbalanced across the countries. Given the high relevance of teacher educational level and potentially selective nature of the missing items, I considered a separate category for missing information on teacher education. While the coefficient of the non-response dummy is not interpretable, it prevents from potential selective attrition of the sample and the resulting bias.

School Characteristics: As the main variable in this study (class gender composition) is defined at classroom level (which in many cases is not distinguishable from school level in TIMSS data structure), it was not possible to include school dummies in the analyses as otherwise, the impact of other class-level variables including class gender composition would be largely underestimated. Given that most school-level variables proved to only trivially affect student achievement (Woessmann, 2005), I only control for the two major school-level factors used frequently in the educational production function in the literature, namely the overall socio-economic background of the student body in school[43], and the degree of *school discipline problems* (Caponera & Losito, 2016). Following the suggestion by Caponera and Losito (2016), the variable *School SES* was calculated by taking the average of the individual-level variable *home educational resources* of all students with the same school ID.

Table 3.1 shows the main descriptive statistics and overall sampling information for sample 1, including 7465 intact classes from 5503 schools in 37 countries, and for sample 2, including 2573 mixed classes and 2246 single-sex classes from 3227 schools in 17 countries. These samples are used for the first and second analyses respectively. For the sake of

[43] Cordero et al. (2017) and Eriksson et al. (2019) suggest that besides controlling for students' socioeconomic status (SES) an aggregate measure of class or school SES also matters.

comparison, the statistics for sample 2 are given separately for single-sex and coeducational subsamples[44].

Table 3.1. Descriptive statistics and sampling information for the samples of analyses 1 and 2.

	Sample 1 (Mixed classes)		Sample 2 (Mixed classes)		Sample 2 (Single-sex classes)	
	male	female	male	female	male	female
Student-level variables:						
Math performance	485.63	484.20	504.66	497.78	446.74	463.05
	(0.95)	(0.93)	(1.62)	(1.75)	(2.24)	(1.93)
Age	14.49	14.38	14.23	14.17	14.09	14.01
	(0.91)	(0.80)	(0.63)	(0.61)	(0.75)	(0.70)
Language dummy						
Sometimes/never speak the language at home	0.24	0.23	0.25	0.24	0.31	0.32
Always/often speak the language at home	0.76	0.77	0.75	0.76	0.69	0.68
Home educational resources						
Few resources	0.09	0.10	0.05	0.05	0.07	0.08
Some resources	0.41	0.45	0.41	0.47	0.43	0.50
Many resources	0.50	0.45	0.54	0.48	0.50	0.42
Class- and school-level variables						
Female ratio	0.46	0.52	_	_	_	_
	(0.11)	(0.13)				
Teacher sex						
Female	0.59	0.61	0.58	0.62	0.22	0.82
Male	0.40	0.38	0.41	0.38	0.77	0.18
Multiple non-same-sex teachers	0.01	0.01	0.01	0.01	0.00	0.00
Teacher education						
Below bachelor	0.10	0.11	0.04	0.05	0.05	0.05
Bachelor degree	0.55	0.55	0.62	0.63	0.62	0.67
Beyond bachelor	0.26	0.25	0.25	0.23	0.18	0.16
Non-response	0.09	0.09	0.09	0.09	0.15	0.12
Teacher experience						
Less than 5 years	0.21	0.21	0.23	0.23	0.17	0.20
Between 5 and 10 years	0.19	0.18	0.24	0.23	0.23	0.24
Between 11 and 18 years	0.20	0.20	0.20	0.21	0.24	0.27
Between 19 and 27 years	0.17	0.18	0.18	0.18	0.21	0.17
More than 27 years	0.23	0.23	0.15	0.15	0.15	0.12
Average SES in school	2.68	2.68	2.80	2.79	2.62	2.68
	(0.68)	(0.68)	(0.63)	(0.64)	(0.58)	(0.55)
Class Size	27.24	27.35	26.12	26.41	27.18	27.60
	(9.93)	(9.90)	(7.60)	(7.60)	(7.64)	(7.21)
School discipline problem						
Hardly any problem	0.39	0.40	0.47	0.47	0.41	0.56
Some problems	0.48	0.48	0.47	0.46	0.45	0.33
Serious problems	0.13	0.12	0.07	0.07	0.15	0.11

[44] To ensures that the weighted sample corresponds to the actual sample size (Foy, 2015), the TIMSS variable "HOUWGT" has been used as individual-level weight for calculating the statistics for each subsample.

Table 3.1 - continued. Descriptive statistics and sampling information for the samples of analyses 1 and 2.

	Sample 1 (Mixed classes)		Sample 2 (Mixed classes)		Sample 2 (Single-sex classes)	
	male	female	male	female	male	female
Number of students	90,619	90,314	29,461	30,219	27,714	27,006
Number of classes	7465		2573		2246	
Number of schools	5503		1736		1491	
Number of countries	37		17		17	

Note: Own calculations based on IEA Database for TIMSS 2015-G8. For categorical variables, the numbers indicate the percentage of the sample group in the respective category. For continuous variables, the values represent the mean and standard deviations (in parentheses). For the dependent variable (math performance), the means and estimated standard errors (in parentheses) were calculated using all five plausible values based on the Item Response Theory as recommended by the IEA guidelines (Foy and Yin, 2015).

The metric for national average mathematics performance in TIMSS has been set to the mean of 500 and the standard deviation of 100 for all participating countries in each wave (Martin et al., 2016). As shown in table 3.1, the mean math performance is especially low for the subsample in single-sex classrooms. This naïve comparison might lead one to conclude that single-sex education resulted in relatively lower performance in mathematics. However, one should note that since single-sex education is more prevalent in low-performing countries (eg. the middle-eastern countries) than in high-performing countries (such as Japan, New Zealand, and so on), in dividing sample 2 into single-sex and mixed classrooms, most of the students from low-performing countries were classified in the single-sex subgroup. Thus, the large difference in the mean performance between single-sex and coeducational subsamples probably confounds, among others, the impact of several country-specific factors with the effect of single-sex education.

Regarding the explanatory variables, it is notable that the distribution of individuals across the categories of home educational resources is rather similar across the subgroups. This in part reflects the reasoning behind the choice of countries for the second analysis in this study. To examine the impact of single-sex education, only the countries with sizable single-sex education have been included to ensure that single-sex schooling is part of the general education system in the selected countries and not merely confined to a highly selective group of families. If countries with highly selective single-sex education had been chosen for the

second analysis, the average SES would have been notably higher (or lower) for the subsample of single-sex classrooms.

It is important to note that part of the variations between the subgroups regarding the explanatory variables might stem from international differences and the role of country-level variables. For instance, the slightly lower mean of school SES in single-sex subgroups might also reflect the relatively higher share of low-income countries in single-sex education. While the statistics for other variables are almost comparable across the sample subgroups, as might be expected, the degree of school discipline problems tends to be particularly higher in all-male and lower in all-female subgroups.

3.3. Method

To examine the gender peer effect, a simple approach is to estimate an MLR model with the OLS method. However, according to Snijders and Bosker (2012), using single-level methods to analyze the relationship between variables of different levels (class gender composition at the class-level and student mathematics achievement at the individual-level) will likely lead to biased results and erroneous conclusions. Thus, accounting for the nesting structure of the data, i.e. individuals nested in classrooms nested in schools in each country, I also apply the hierarchical linear modeling (HLM) approach, which enables to model both within- and between-group variability.

3.3.1 Analysis 1: Varying female ratios in mixed classrooms

As a starting point to test the first hypothesis (H1), I estimate the equation 3.1 below with the OLS method and standard errors clustered at country level[45].

$$y_{ijkc} = \theta_0 + \Sigma_{\mathrm{n}=1}^{p}\gamma_n I_{nijkc} + \Sigma_{\mathrm{n}=1}^{q}\gamma_n C_{njkc} + \Sigma_{\mathrm{n}=1}^{w}\gamma_n S_{nkc} + CFE_c + \epsilon_{ijkc} \qquad (3.1)$$

In equation 3.1, y_{ijkc} stands for the math achievement of student *i* in class *j* of school *k* in country *c*, which is estimated as a linear function of *p=3* individual-level variables in vector

[45] It has been suggested by Bottomley et al. (2016) that, particularly when there are multiple levels of clustering, it is generally preferable to use confidence intervals that account for the highest level of clustering, except for the cases with few clusters and high intra-class correlation. Accordingly, the standard errors in this study are clustered at the country level for the general models (using the pooled sample of all countries), and at the school level for country-specific models.

I (*age, language-at-home dummy, home educational resources*), *q=5* class-level variables in vector *C* (*female ratio in classroom, teacher gender, teacher education, teacher experience, and class size*) and *w*=2 school-level variables in vector *S* (*School SES* and the degree of *school discipline problems*). θ_0 denotes the average intercept for all students in the sample for the analysis 1 (students from 37 countries participating in TIMSS2015). Additionally, the model captures the country-dependent variations in the average scores caused by different national educational policies and procedures through the country fixed effects (CFE_c). ϵ_{ijkc} is the individual-level random error term capturing unobserved variability between the individual students (due to individual-level unobserved variables such as the omitted ability, etc.).

Nevertheless, the zero conditional mean assumption in the MLR model is only valid if, among other factors, no relevant class-level variable remains unobserved. This assumption is likely to be violated by systematic differences between the classrooms, due to namely the ability peer effect or the level of competition in class. In fact, equation 3.1 assumes that the whole variation in the outcome variable comes either from observed differences (in terms of student-, class-, school-, and country-level variables), or from *individual* deviations from the averages. In fact, it is assumed that the observations even within each classroom are not systematically correlated. This is however, not a plausible assumption because the two-stage sampling design of TIMSS, in which the students in the second stage were not independently selected, has been ignored. One could reasonably argue that even in a single country, different classrooms might deviate from the average country-specific performance by a random class-level error term reflecting the unobserved variability across the classrooms. To account for this between-class-variation, a random intercept specification is used according to equation 3.2 below:

$$y_{ijkc} = \theta_{0000} + \Sigma_{n=1}^{p}\gamma_{n000}I_{nijkc} + \Sigma_{n=1}^{q}\gamma_{0n00}C_{njkc} + \Sigma_{n=1}^{w}\gamma_{00n0}S_{nkc} + CFE_c + r_{0jkc} + \epsilon_{ijkc} \qquad (3.2)$$

In the mixed-model equation above, the same outcome variable (y_{ijkc}) is estimated as a linear function of certain fixed parameters plus the random errors at different levels. The vectors of explanatory variables in the fixed part are defined likewise the MLR model except that the

subscripts of the coefficients follow the standard notation in HLM literature in which the non-zero digit represents the level at which the respective variable varies. The random part in equation 3.2 includes an extra term for the unexplained variability in the average performance of classrooms, i.e. variations that have not been explained by the class-level parameters in the fixed part formulated as r_{0jkc}. This might include the impact of class-level unobserved variables such as the teacher's motivation or the degree of competition in the classroom. In fact, with r_{0jkc} adding a random offset to the outcome variable for each participant from classroom *j* of school *k* in country *c*, equation 3.2 estimates a unique random intercept for each individual student in the sample as $\theta_{0000} + r_{0jkc} + \epsilon_{ijkc}$. The model is estimated with the maximum likelihood method.

In the random intercept specification (equation 3.2), the associations between the driving factors and student performance are assumed to be constant across all classrooms and all countries. Since it is highly unlikely that the impacts of the variables, including female ratio in classroom, are the same across different countries, I also construct country-specific models that allow for heterogeneous effect of the variables across countries to provide evidence for testing the second hypothesis (H2).

Moreover, to test the third hypothesis (H3), I follow Lavy and Schlosser's (2011) approach to examine how the gender peer effect changes across its range, calculating quartiles of female proportions in classrooms based on its distribution and replacing the female ratio variable in equation 3.2 with the generated quartiles of female ratio.

3.3.2 Analysis 2: Single-sex education vs. coeducation

In the second analysis, using the sample of 17 countries with sizable single-sex and coeducation, I follow a similar approach to analysis 1 to examine how participation in single-sex classrooms affects student math scores (H4). First, by using a linear regression model, I assume that the observations between and within the classrooms are independent. As this simplifying assumption is mostly violated in two-stage sampling designs such as the TIMSS', I secondly construct a hierarchical (mixed-effect) model similar to equation 3.2 with the female ratio variable replaced by a dummy variable (single-sex) equal to zero for the students in mixed classrooms and one for those in single-sex classes.

Finally, to test whether the impact of single-sex education is different across countries (H5), I run a separate random intercept model for each country. Country-specific estimations allow all factors, including single-sex dummy variable, to have a differing impact across countries.

For implementing the modelling process stated above, separate models are constructed for each gender group so that all models offer full flexibility in terms of the potentially different impacts for male and female students. Furthermore, it is important to note that according to Foy and Yin (2015), the multi-stage cluster sampling design and the incomplete item responses by each individual student in TIMSS necessitate the use of sampling weights, specific way of calculating variances (using Jackknife Repeated Replication(JRR) technique), and the aggregation of the results across the five plausible values. Therefore, given the purpose of each analysis in this chapter, I choose the proper sampling weights according to the suggestion by Foy (2017). More specifically, since the student total weight variable in TIMSS (TOTWGT) inflates sample sizes to estimate the population size, in order to avoid larger countries disproportionately affecting the estimates in the analyses of pooled samples, I use the sampling weight variable "SENWGT" (a transformation of "TOTWGT" that results in a weighted sample size of 500 in each country). In country-specific analyses, I used "HOUWGT" (household weight) to ensure that the weighted sample corresponds to the actual sample size in each country. Also, since weights enter into the log likelihood at both class and student levels in mixed-effect models, school-level weight variable "SCHWGT" is additionally applied for hierarchical models. Following Rabe-Hesketh and Skrondal (2006), sampling weights are then rescaled to sum to the cluster (class) size. Finally, each analyses is performed five times (once for each plausible value), and the final results are then aggregated across the five values as suggested by Foy and Yin (2015).

3.4. Results

3.4.1 Estimation of female-ratio effect (Analysis 1)

Testing H1

Table 3.2 shows the estimated coefficients for the uncontrolled and controlled MLR and HLM models using the pooled sample in the first analysis.

It is notable that the direction of the coefficients are consistent with expectations from the previous literature. For instance, the amount of educational resources at home and its aggregation across the school (general SES of the student body in school) have the largest positive impacts on student achievement. Native students or those who often speak the language of the test at home are more likely to perform better in the exam. Student age is negatively associated to performance as the repeaters of a grade are less likely to get higher scores in the exams. Students in more disciplined schools generally perform better. Exceptionally, in contrast to previous findings (Angrist & Lavy, 1999; Woessmann, 2005), the association with class size is positive in general. This might be due to the impact of outliers (very small classes in rural areas that suffer from other shortages in class facilities, etc.). The relationship might be more precise if a quadratic term of class size was included. Teachers' higher level of education is associated with students' better performance. None of the categories of teacher experience turned out to affect student achievement with statistical significance, but the sign and size of the impact are not inconsistent with expectations. Finally, students who study in schools with less discipline problems generally perform better in the test[46].

Table 3.2. Estimated coefficients produced by one- and multi-level (mixed) models (analysis1).

Independent variable	**SLR** (Naïve model)		**MLR** (equation 3.1)		**HLM/Mixed** (equation 3.2)	
	male	**female**	**male**	**female**	**male**	**female**
Student-level variables						
Age			-13.40***	-11.42***	-11.90***	-10.06***
			(0.63)	(0.68)	(0.68)	(0.73)
Language dummy			15.95***	14.58***	8.95***	9.11***
			(1.35)	(1.50)	(1.27)	(1.62)
Home educational resources						
Some resources			16.49***	19.09***	9.83***	11.37***
			(1.74)	(1.42)	(2.75)	(2.06)
Many resources			19.78***	25.00***	11.10***	15.49***
			(1.94)	(1.72)	(2.75)	(2.27)
Class/School-level variables						
Female ratio	0.69***	-0.16*	0.51***	0.29***	0.46***	0.18**
	(0.09)	(0.09)	(0.06)	(0.06)	(0.09)	(0.08)

[46] As the effects of other context factors are not the focus of this study and are widely investigated in the previous literature, sometimes with rigorous non-cross-sectional designs to unfold causal links, I am not going to interpret the coefficients in more detail.

Table 3.2 - continued. Estimated coefficients produced by one- and multi-level (mixed) models (analysis1).

Independent variable	SLR (Naïve model)		MLR (equation 3.1)		HLM/Mixed (equation 3.2)	
	male	**female**	**male**	**female**	**male**	**female**
Teacher sex						
Male			-2.52*	-1.65	-1.62	-0.24
			(1.32)	(1.29)	(2.14)	(2.15)
Multiple non-same-sex teachers			-1.35	4.10	-8.57	5.05
			(6.75)	(6.70)	(7.20)	(6.76)
Teacher Education						
Below bachelor			-6.36**	-6.23**	-3.17	-0.31
			(2.89)	(2.45)	(3.45)	(3.36)
Beyond bachelor			3.52**	1.69	2.61	1.92
			(1.59)	(1.66)	(2.52)	(2.09)
Non-response			-0.44	-0.95	-0.44	-0.95
			(3.75)	(3.40)	(5.55)	(4.80)
Teacher experience						
Between 5 and 10 years			-0.86	-0.46	-0.79	0.35
			(1.83)	(1.98)	(3.20)	(2.95)
Between 11 and 18 years			1.44	2.89	5.62*	5.05
			(1.77)	(1.89)	(3.32)	(3.15)
Between 19 and 27 years			2.88	3.77*	6.44*	7.42**
			(2.32)	(2.07)	(3.51)	(3.20)
More than 27 years			3.45	4.12**	4.73	6.43*
			(2.19)	(2.09)	(3.59)	(3.29)
Class Size			0.50***	0.47***	0.54***	0.58***
			(0.11)	(0.11)	(0.15)	(0.15)
Average SES in school			54.80***	51.85***	52.70***	51.08***
			(1.41)	(1.49)	(2.12)	(1.96)
School discipline problem						
Some problems			-11.04***	-9.07***	-12.39***	-12.91***
			(1.44)	(1.13)	(2.06)	(2.00)
Serious problems			-17.45***	-18.57***	-19.49***	-21.25***
			(2.17)	(2.53)	(4.29)	(4.12)
Country dummies			✓	✓	✓	✓
Students	90619	90314	81009	81059	81009	81059
Classes					6661	6743
Schools					5020	5090
Countries					37	37

Note: Own calculations based on IEA Database for TIMSS 2015-G8. The reference groups for categorical covariates, home educational resources, teacher sex, teacher education, teacher experience, and school discipline problems are "few resources", "female teacher", "bachelor degree", "less than 5 years" and "hardly any problem" respectively. Language dummy equals 0 for the students who never or only sometimes speak the language of test at home and 1 for those who often or always do so. For the one-level models (SLR and MLR), the number of observations equals the number of students. For the HLM model, the number of cases at each level is reported. As recommended by the IEA guidelines (Foy and Yin, 2015), all five plausible values provided by TIMSS were used for estimations. Numbers in parentheses show standard errors. Stars represent statistical significance levels. * $p < .10$, ** $p < .05$, and *** $p < .01$.

Nevertheless, the focus here is on the coefficient of female ratio. A naïve analysis indicates that while changes in the proportion of females in class do not result in a statistically non-zero average change in the girls' math scores, boys on average perform better as the share of females among their peers increases. The OLS estimation of a simple linear regression model produces the coefficients for female ratio as of +0.69 (statistically significant at 1% level) for boys and -0.16 (statistically significant only at 10% level) for girls. This implies that when the proportion of females in the classroom rises by 10 percent, boys on average score around 7 points higher than what they would have scored in a class with 10 percent less females. Such an increase in the proportion of females does not affect the math performance of girls at any statistically significant level.

When student, teacher, class and school characteristics are taken into account, a 10 percent increase in the proportion of females in class improves males' average performance by 5.1 points. The association is weaker but in the same direction among female students, indicating an average improvement of 2.9 points in females' scores in the case of a 10 percent higher female proportion in class. Both coefficients are statistically highly significant (at $\alpha = 0.01$).

While the cross-sectional nature of the data limits the interpretation of the OLS results in terms of causality, the random intercept model in table 3.2 accounts for the unobserved teacher and class-level variables, producing estimations less prone to omitted variable bias. The mixed-effect estimated coefficients, including those of female ratio, are mostly attenuated, probably reflecting the part of the impacts captured by unobserved class-level factors. Accordingly, assuming a linear relationship, each 10 percent increase in the proportion of females in class improves males' and females' math scores by around 4.6 and 1.8 points respectively, equivalent to the improvements of approximately 4% and 2% of a standard deviation in score distributions.

Testing H2

Relieving the restricting assumption that the explanatory variables have the same impact across all countries, country-specific random intercept models produce the female-ratio coefficients and corresponding confidence intervals as depicted in figure 3.1.

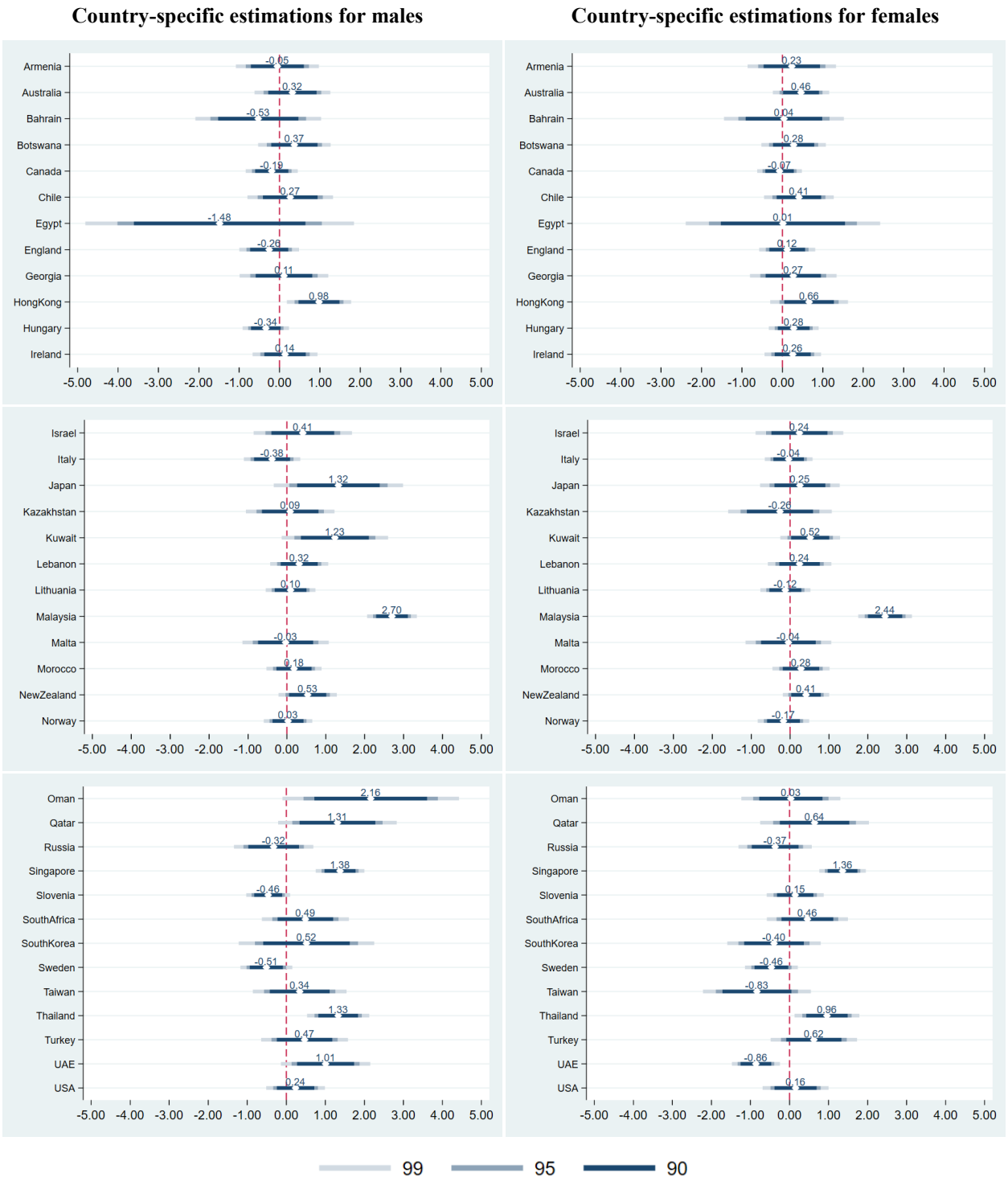

Figure 3.1. Country-specific coefficients of female ratio estimated by the random intercept (mixed) model

In this figure, for countries that the entire spectrum of the confidence interval does not include zero, the coefficient is statistically highly significant (at 1% level). For those that the lightest shading (99% confidence interval) crosses the zero line, the coefficient is still statistically significant (at 5% level). For other countries with 95% or 90% confidence intervals crossing the zero line, the coefficients are hardly or not statistically distinguishable from zero.

Accordingly, while in most countries the female ratio in classroom did not influence student math scores at any statistical significance level, in several countries the impact was considerable. For males in Hong Kong, Japan, Kuwait, Malaysia, Oman, Qatar, Singapore, Thailand, and the United Arab Emirates the impact is statistically significantly different from zero (at $\alpha = 0.05$ level or lower). The largest impact occurs in Malaysia and Oman where, assuming linearity, each 10 percent increase in the proportion of females in class is associated with around 27 and 22 points increase in boys' math scores respectively. Except for the statistically significant and positive coefficients for Malaysia, Singapore, and Thailand, females in most countries seem to be indifferent to the changes in the proportion of same-sex peers. The association is exceptionally reversed for girls from the United Arab Emirates (UAE) who scored about 9 points lower when placed in classes with 10 percentage more females. Summing up, the country-specific estimations show that the impact of the ratio of females in classroom on students' educational outcomes is heterogeneous across the countries.

Testing H3

Estimating equation 3.2 with female ratio in decimal points replaced by its quartiles suggests the nonlinear relationship between the proportion of females among peers and student achievement. Table 3.3 reports the main statistics of the generated quartiles (Q1 to Q4) of the female ratio and the estimations for the average change in boys' and girls' scores by switching to higher quartiles in the random intercept models. Accordingly, except for the top quartile in females' estimation, switching to higher quartiles of a female ratio generally produces larger estimations for the relationship between the female ratio in classrooms and student math scores. For male students, the strongest relationship seems to concentrate towards the tail end of the female ratio range (top quartile). For females however, the relationship seems to follow a parabola shape pointing downward as it initially gets stronger

up to the third quartile and weakens when the female ratio further increases. The increments in boys' estimations and the rise and fall in girls' estimations both suggest nonlinearity in the relationship.

Table 3.3. Main statistics of female-ratio-quartiles and the nonlinear impact of female-ratio.

	Males				Females			
	Q1	**Q2**	**Q3**	**Q4**	**Q1**	**Q2**	**Q3**	**Q4**
Range	2.6-42.3	42.4-50.0	50.1-56.0	56.1-97.4	2.6-42.3	42.4-50.0	50.1-56.0	56.1-97.4
Mean	34.7	46.8	53.4	62.3	36.6	47.0	53.4	66.2
Students	29400	30123	15486	15610	16276	26736	17887	29415
Estimation	_	8.90***	9.92***	12.77***	_	8.54***	10.32***	8.38**
		(2.46)	(2.76)	(2.67)		(2.48)	(2.72)	(2.54)
Students	81,009				81,059			
Classes	6661				6743			
Schools	5020				5090			
Countries	37				37			

Note: Own calculations based on IEA Database for TIMSS 2015-G8. The first three rows report the range, mean and the number of students in each quartile respectively. The estimation row reports the results from the multilevel regression (random intercept model) of students math performance on female-ratio-quartile and covariates at student, class, school, and country levels. The first quartile (Q1) is the base category. The last four rows report the number of cases at each level (students, classes, schools and countries) in the estimations. As recommended by Foy and Yin (2015), all five plausible values of TIMSS were used in the estimations. Numbers in parentheses show standard errors. Stars represent statistical significance levels. * $p < .10$, ** $p < .05$, and *** $p < .01$.

3.4.2 Estimation of single-sex education effect (Analysis 2)

Testing H4

Table 3.4 presents the coefficients of single-sex dummy estimated by the one-level linear regression model and the hierarchical linear or mixed model each with and without additional controls[47].

As shown by table 3.4, the correlation coefficients equal -54.01 for males and -31.45 for females, both statistically distinguishable from zero at $\alpha = 0.01$, implying that on average male and female students who participated in single-sex classrooms performed remarkably worse than did those who attended mixed-gender classrooms. This simple approach is,

[47] For reasons of brevity, I only present the coefficients of single-sex dummy estimated by different models. Table A.3.5 in the appendix provides the estimated coefficients of control variables in the models generated for all countries in the second analysis.

however, substantially prone to biases caused by several systematic differences between the two types of educational environment and the entrants.

Table 3.4. Estimated coefficients of single-sex dummy produced by one- and multi-level (mixed) models (analysis2).

Independent Variable	SLR (Naïve model)		MLR (Equation 3.1)		HLM/Mixed (Equation 3.2)	
	male	female	male	female	male	female
Single-sex dummy	-54.01***	-31.04***	-13.46***	-11.52***	-23.52***	-15.42***
	(2.76)	(2.46)	(2.53)	(2.14)	(3.42)	(2.88)
Students	57381	57493	51006	51153	51006	51153
Classes					3306	3322
Schools					2313	2321
Countries					17	17

Note: Own calculations based on IEA Database for TIMSS 2015-G8. For the one-level models (SLR and MLR), the number of observations equals the number of students stated in the last row. For the multilevel models (the last two columns), the number of cases at each level (students, classes, schools and countries) is reported. The estimated coefficients of control variables in analysis 2 are presented in table A.3.5 in the appendix. As recommended by Foy and Yin (2015), all five plausible values of TIMSS were used in the estimations. Numbers in parentheses show standard errors. Stars represent statistical significance levels. * $p < .10$, ** $p < .05$, and *** $p < .01$.

The OLS estimation of the controlled (MLR) model shows that the association between the single-sex dummy and test scores becomes weaker when pre-existing differences between the two types of institutions and their admitted students are at least partially taken into account. With the same demographic and socio-economic background and similar teacher, class, and school characteristics, students in all-male classrooms on average scored around 13 points lower than did their counterparts who attended in class with female classmates. Nor did female students benefit from participation in all-female classes as they also scored about 12 points lower than their counterparts in mixed classrooms.

The mixed-effect model addresses possible interdependence and correlations between individual observations from the same classroom. As reported in table 3.4, male and female students in coeducational settings outperformed their counterparts in math by 23.52 and 15.42 points respectively (both coefficients are statistically significant at the 1 percent level).

Testing H5

The country-specific estimates of the single-sex classroom effect are illustrated in figure 3.2.

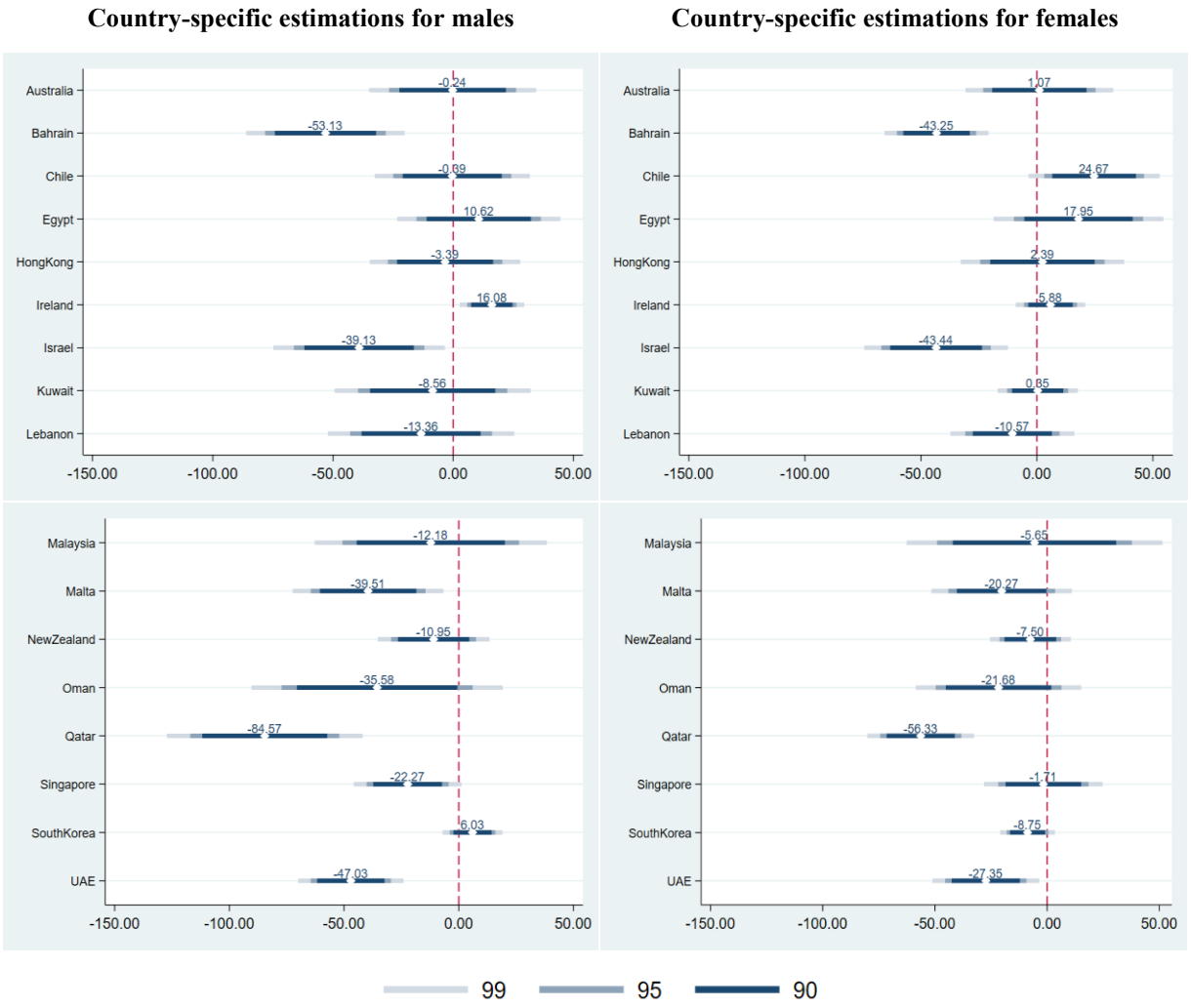

Figure 3.2. Country-specific coefficients of single-sex dummy estimated by the random intercept (mixed) model

As single-sex education is an extreme case of a female ratio in a classroom, it is expected that the sizes of the coefficients in the second analysis would be much higher than for those of the first analysis. Assuming a threshold of $\alpha = 0.05$ for statistical significance, in most of the countries under study, the math scores of participants in single-sex classes did not differ from those of students in a coeducational environment (Australia, Egypt, Hong Kong, Kuwait, Lebanon, Malaysia, New Zealand, Oman, and South Korea). Male students from Bahrain, Israel, Malta, Qatar, Singapore, and the UAE who attended all-male classrooms underperformed, by 23 to 85 points, in relation to their counterparts in mixed-gender classes

(statistically significant at 5 percent level). Like their male counterparts, females from Bahrain, Israel, Qatar, and the UAE did worse when participated in single-sex rather than mixed classes. In contrast, Irish boys and Chilean girls performed by around 16 and 25 points better (statistically significant) when placed in single-sex rather than mixed classrooms.

To sum up, the association between the single-sex dummy and student math score is different across countries, providing no evidence for the rejection of the fifth hypothesis.

3.5. Discussion

Exploiting the widespread coverage of the TIMSS data, this study set out to determine the effect of the gender composition in classrooms on students' math achievements across the participating countries. Although the cross-sectional nature of TIMSS data does not allow for value-added causal estimations, it still provides in-depth information on several relevant factors at different levels.

Given that single-sex and coeducational institutions in most countries systematically differ in relevant ways, two separate analyses were conducted to investigate the impact of the proportion of females in the classroom and of participation in single-sex versus mixed-gender classes.

In the first analysis, the results of the coeducational pooled sample of 37 countries participating in TIMSS2015 showed that students generally benefited from the presence of higher proportion of females in class. More specifically, when the share of females in the classroom increased by 10 percent, boys' and girls' scores improved by around 4 and 2 percent of a standard deviation on average. The impact was quite heterogeneous across the countries and across the range of female ratios, implying the nonlinearity in the relationship.

The second analysis using the sample of 17 TIMSS2015 participating countries in which the share of students in single-sex and coeducation both exceeded 10 percent of the country sample size revealed that male and female students generally do not benefit from participation in single-sex classrooms. On average, male and female students with solely same-sex classmates underperformed by 21 and 16 percent of a standard deviation in relation

to their counterparts in mixed-gender classes. The negative associations turned out to be larger in some countries.

Regarding the estimation strategy used in this study, among the hierarchical linear modeling approaches, I used the random intercept model to account for unobserved systematic differences between the classrooms. However, before running more complex multilevel models, I checked whether substantial variation actually existed at class level, i.e. whether a multilevel approach was required to explain the total variation, constructing the so-called null or empty model, which contains one fixed term (the constant term), and variance components at individual and class levels. The results for the first analysis showed that around 35% of the total variation in the outcome variable came from between-class variability, both for males (3,711 out of 10,494 units) and females (3,406 out of 9,735 units). The likelihood ratio test with inputs from the OLS and the random intercept models also confirmed that systematic differences in average math performance between the classrooms were not negligible, and a hierarchical linear modeling approach seemed necessary (the P-value of the test was statistically significant at 1% level). For the second analysis, the empty model showed that between-class variances constituted about 32% of the total variances both for male and female estimations. More specifically, the variance of residuals amounted to 3855 out of 12248 units of total variance for males' scores, and to 3495 out of 10425 units of total variance for females' scores.

Nevertheless, one might argue that due to certain class-level unobservable factors, some variables might have differentiated impacts across the classrooms even within countries. For instance, the impact of teacher experience is probably larger in a school with a rigorous and structured in-service training program for teachers. As for the variable of interest in this study, namely the ratio of females in class, the impact in a classroom taught by a teacher with gendered and biased beliefs about student competencies might differ from that in a class whose teacher does not hold such gendered views. To account for these concerns, one might need to add a class-level random part to the coefficients of all such class-level variables in equation 3.2, and estimate the resulting *random slope model*. The model however would become too complicated, and in this case, did not converge. In order to ensure that the use of a random intercept instead of a random slope model does not considerably harm the

estimations, I constructed parsimonious country-specific models with a random slope for the female ratio variable. Given that not all country-specific random slope models converged and that the estimated effects for the converging country models were nearly identical to those obtained from random intercept models, I refrained from constructing a random slope model and conducted the main analyses with the random intercept approach.

Given the selection of independent variables in this study, one might argue that student's attitude towards the subject could be a relevant variable in determining student performance (Caponera and Losito, 2016, Geesa et al., 2019). Since as stated in the theory section (3.2), students' self-concept could be affected by the share of same-sex students in class, some correlations might exist between the gender composition in class and student's attitude towards math. Furthermore, if as evidenced by several studies (eg. Caponera & Losito, 2016; Eriksson et al., 2019), teaching style is among the relevant factors affecting students' outcomes, one would need to control for this variable due to the potential correlation between teaching style and gender composition in class. It could be the case that in a classroom with a higher proportion of male students with more frequently disruptive behaviors, the teacher might be more inclined to use a teacher-centered style (giving lectures, asking students to listen and copy the solutions to problem sets from the board, etc.), and be conservative in using a student-oriented style with practices such as interactive teaching, constructing working groups and asking students to work on problem sets together. While TIMSS provides relevant information to measure students' math attitude and teacher's style for teaching math, I excluded them from the models in the analyses of this study. Besides the issue of severe sample attrition due to the large non-response rate for the related items, the inclusion might lead to over-controlling bias in the results because teaching style and math attitude could constitute the *mechanism* variables through which the impact of class gender composition operates. Therefore, in this study, I relied on the models without including them as controls[48].

[48] I conducted sensitivity analyses to investigate how the main results would change when the variables were controlled in the random intercept models in both analysis 1 and 2. The results showed that the inclusion of these two variables only trivially influenced my main results. As the t-statistics for the coefficients of interest were rather large, the small changes in the estimated coefficients (about max 0.02 in female-ratio effect and maximum 2.00 points in single-sex effect) did not affect the statistical significance or direction of the impacts.

The general trend in the results presented in this chapter is consistent with previous studies that both male and female students' academic outcomes improve when the female proportion in class increases (eg. Hoxby, 2000; Lavy & Schlosser, 2011). Since previous studies on the impact of single-sex education have found almost all possible directions and different sizes for the impact, the results of the second analysis in this study could not be regarded as inconsistent with the existing findings.

However, the estimated coefficients are not of the same size or in the same direction across the countries. As briefly mentioned in the introduction, these heterogeneities could partly stem from differences in cultural values and educational systems among the countries that have implications on the impacts. While the cross-sectional nature of the data does not allow to address the underlying reasons for individual coefficients and the possible mechanisms behind these effects, it is worth drawing attention to some patterns in country-specific estimations. For example, in the first analysis, the relationship between the students' outcome and the proportion of females in class has the largest positive values in Asian countries - Malaysia, Hong Kong, Singapore, and Thailand. Geographical proximity in this case might reflect cultural proximity or some resemblance between the education systems of these countries. Regarding the second analysis, it is interesting that the relationships between attending single-sex classrooms and students' math scores is strongly negative in several Muslim-majority and Arab countries (Bahrain, Qatar, and the UAE) where single-sex education is more prevalent. This is a somewhat puzzling result. Due to the overall preferences for single-sex education over coeducation in these societies, one would expect that single-sex education is more beneficial or at least less harmful for the students' academic performance[49]. While various mechanisms might intervene, a possible explanation refers to cultural distinctions (Salikutluk & Heyne, 2017) in these countries that naturally lead to a transformation of gender dynamics in the classroom. Given that public life is more segregated and stereotypical gender roles are more endorsed among Muslim societies, adolescents are more likely to perceive stereotypes as self-relevant and identify themselves with gendered

[49] According to Jackson (2012), students' preferences play a key role in determining the impact of single-sex education on student achievement. In his study on a sample of randomly allocated students to single-sex and mixed classes in Trinidad and Tobago, he found that the female students who expressed a strong preference for single-sex education performed significantly better in exams when participating in all-female classrooms.

beliefs about academic competencies (Maher, 2012). Thus, stereotype threat is probably more detrimental to student performance in such societies. Another explanation could refer to the benefits of classroom interactions among genders. If, as some proponents of coeducation claim, interactions among genders could stimulate better learning opportunities, it might be expected that the omission of these interactions in classrooms is more harmful in segregated countries than in societies where students have many other possibilities for interacting with the opposite gender.

Nonetheless, while some possible explanations for the observed associations were discussed above, I acknowledge the limitations of the study for causal interpretations and investigation of the mechanisms. To uncover the causal links and investigate the channels of the impact in each country, more detailed data on relevant cultural and educational factors and possible systematic differences between the schools and classrooms are required.

3.6. Chapter Overview and Conclusion

This chapter examined how students' educational outcomes could be affected by the gender composition of the learning environment. In two separate cross-country analyses, I investigated how the proportion of females in classrooms or the participation in a single-sex rather than mixed-gender setting influences students' math achievement. Findings showed that students generally gain academic benefit from a higher proportion of female classmates, but at the same time, from the presence of the opposite gender in the classroom environment. The impacts are heterogeneous across countries. The gender peer impact shows a nonlinear pattern across its range, with larger effects at the extreme values in single-sex classrooms.

The results of this study contribute to the dearth of empirical evidence on the gender peer effect in several unexamined countries for which detailed and sufficient data are mostly inaccessible despite the primary relevance of the issue. In addition, international comparisons, even in terms of associations, provide certain benefits. As noted by Hanushek and Woessmann (2010), if one assumes that any bias is constant across countries, then cross-country comparisons could offer valuable insights even if interpretation of each estimation is not feasible. With such a rich set of relevant variables provided by TIMSS, the comparison among the coefficients, between the countries and between genders within a country provides valuable insights for efficient educational policies. Therefore, while the limitations for causal inference from the present results inhibit a clear message for the reallocation of students and placement of low-achievers in individual countries, the findings reveal the potential importance of class gender composition in the education production function, a factor that has been less emphasized by previous international comparisons of student achievements.

Last but not least, despite their relatively lower impacts on student achievement (Woessmann, 2016), class- and school-level factors are more under the control of educational policymakers (than are family background and institutional features of school systems). Given that school and class factors are more closely related to student performance in developing countries than in the developed world (Hanushek and Woessmann, 2010), they could play an instrumental role in shaping efficient developmental policies and reforms in the realm of education.

Appendix

Table A.3.5. Estimated coefficients produced by one- and multi-level (mixed) models (analysis 2).

Independent Variable	SLR		MLR		HLM/Mixed	
	male	female	male	female	male	female
Age			-12.15***	-10.48***	-12.32***	-10.31***
			(0.96)	(0.85)	(0.85)	(0.83)
Language dummy			2.19	4.61***	5.46***	4.72***
			(1.47)	(1.50)	(1.18)	(1.51)
Home educational resources						
Some resources			29.04***	26.52***	16.32***	15.74***
			(2.52)	(2.76)	(2.58)	(2.20)
Many resources			33.64***	34.80***	16.71***	18.01***
			(2.74)	(3.01)	(2.67)	(2.40)
Single-sex dummy	-54.01***	-31.04***	-13.46***	-11.52***	-23.52***	-15.42***
	(2.76)	(2.46)	(2.53)	(2.14)	(3.42)	(2.88)
Teacher sex						
Male			-13.63***	-1.77	-10.29***	-6.09**
			(2.14)	(2.17)	(3.09)	(2.98)
Multiple non-same-sex teachers			-0.72	-7.96	-19.71*	-13.93
			(11.82)	(13.31)	(10.88)	(10.57)
Teacher Education						
Below bachelor			-4.10	-9.63**	-6.83	-6.93
			(4.45)	(3.92)	(5.59)	(4.55)
Beyond bachelor			8.31***	10.36***	1.57	3.88
			(2.77)	(2.68)	(3.31)	(3.23)
Non-response			-3.99	1.66	-1.75	-1.92
			(5.40)	(4.19)	(7.45)	(5.54)
Teacher experience						
Between 5 and 10 years			4.70**	-2.00	6.50	-1.50
			(2.31)	(2.50)	(4.25)	(3.85)
Between 11 and 18 years			5.05**	1.14	11.13**	4.88
			(2.42)	(2.44)	(4.42)	(3.85)
Between 19 and 27 years			2.92	5.56*	13.69***	7.58*
			(3.09)	(2.86)	(4.62)	(4.39)
More than 27 years			2.32	10.15***	7.91	14.41***
			(3.30)	(2.99)	(4.97)	(4.49)
Class Size			1.73***	1.22***	1.69***	1.46***
			(0.17)	(0.15)	(0.25)	(0.20)
Average SES in school			68.85***	64.08***	65.13***	65.22***
			(2.17)	(1.79)	(3.37)	(2.47)
School discipline problem						
Some problems			-8.12***	-9.26***	-10.77***	-13.27***
			(1.90)	(1.62)	(2.76)	(2.35)
Serious problems			-13.74***	-16.20***	-20.07***	-20.33***
			(4.09)	(3.82)	(4.88)	(5.19)

Table A.3.5 - continued. Estimated coefficients produced by one- and multi-level (mixed) models (analysis 2).

Independent Variable	SLR		MLR		HLM/Mixed	
	male	female	male	female	male	female
Country dummies			✓	✓	✓	✓
Students	57381	57493	51006	51153	51006	51153
Classes					3306	3322
Schools					2313	2321
Countries					17	17

Note: Own calculations based on IEA Database for TIMSS 2015-G8. The reference groups for categorical covariates, home educational resources, teacher sex, teacher education, teacher experience, and school discipline problems are "few resources", "female teacher", "bachelor degree", "less than 5 years" and "hardly any problem" respectively. Language dummy equals 0 for the students who never or only sometimes speak the language of test at home and 1 for those who often or always do so. For the one-level models (SLR and MLR), the number of observations equals the number of students stated in the last row. For the multilevel models (the last two columns), the number of cases at each level (students, classes, schools and countries) is reported. As recommended by the IEA guidelines (Foy and Yin, 2015), I used all five plausible values provided by TIMSS for all estimations. Numbers in parentheses show standard errors. Stars represent statistical significance levels. * $p < .10$, ** $p < .05$, and *** $p < .01$.

Chapter 4

Who teaches me? Teacher Gender and Student Achievement in Muslim-majority Countries

4.1. Introduction

Does it matter who teaches you? Do you learn the same content from a female teacher better than from a male teacher? Does your own gender play a role herein?

Numerous studies on determinants of student achievement have emphasized the key role of the teachers. While no other attribute of schools comes close to having this much influence on student performance, previous research has failed to identify any specific characteristics of teachers reliably related to student achievement (Hanushek, 2011). Among the potential teacher-related factors, teacher's gender and the possibility of student-teacher gender interaction effects have received considerable and increasing attention in the literature (Dee, 2007; Marsh et al., 2008; Cho, 2012; Winters et al., 2013; Paredes, 2014; Lim & Meer, 2015).

Many scholars pinpointed that student-teacher gender-match could provide academic benefit for the students due to the role model effect (Bettinger & Long, 2005; Holmlund & Sund, 2008; Paredes, 2014; Antecol et al., 2015; Egalite et al., 2015), or the teachers expressing

higher expectations to students of the same sex (Dee, 2007; Lim & Meer, 2015; Gershenson et al., 2016), leading to their higher motivation and improved performance. However, the empirical results are quite mixed. The inconsistencies are mostly attributed to limitations in research designs due to small sample sizes (eg. Bielock et al., 2010), insufficient number of male teachers in the samples (eg. Winters et al., 2013), and most importantly, non-random sorting of teachers and students between and within schools (eg. Chudgar & Sankar, 2008). If high-ability students are more likely to be assigned to teachers of a certain gender, the cross-sectional analysis could suffer from severe omitted variable bias. To address the selection issues, Dee (2005) introduced an innovative strategy, using within-student score variations across subjects to difference out key unobserved student traits such as ability level. He found that in case of assignment to same-sex teachers, the students were less likely to be perceived as disruptive and inattentive (Dee, 2005), and more likely to perform better in the subject (Dee, 2007).

While several studies have examined the causal impact of student-teacher gender interaction in the developed world, only few studies focused on this issue in developing countries[50], most probably due to the lack of adequate data on confounding factors from nationally representative samples. However, the issue of gender interaction effect is highly relevant in these countries because on the one hand, studies have shown that teacher and school factors are generally more influential in the developing countries than in the developed world (Hanushek and Woessmann, 2010). On the other hand, the results of the existing studies are not easily applicable in different contexts, particularly in Muslim-majority countries (MM countries) where distinctive cultural norms have important implications on social interactions between genders[51]. For example, while the gender of the teacher might not be an issue for the girls who have had several male teachers from childhood, it is probably the most salient

[50] See for example, Saha (1983), UNESCO (2006), Aslam and Kingdon's (2007), Chudgar and Sankar (2008), Rawal and Kingdon (2010), and Okoro et al. (2012), none of which addressed the potential non-random sorting of students into classes.

[51] This study does not intend to focus on the role of religion in determining the effect of teacher gender. However, given the prominent role of religion in determining one's cultural identity (Dahl et al., 2020), the term "Muslim-majority countries" is used to refer to the countries with mainly similar cultural norms with certain implications on gender issues.

attribute of the teacher for the girls who firstly encounter a male teacher in secondary education. In the latter case, which is common in a Muslim society, the teacher of the opposite gender might induce entirely disparate studying behaviors. The students might put extra effort to prove their educational competencies or shy away from being involved in class discussions to secure themselves from embarrassment. Similarly, having a female math teacher with profound math background could induce a stronger role model effect for a girl from a traditional family who does not typically interact with highly-educated women outside the classroom than for a girl from western and gender-equal cultural background[52]. Therefore, due to the basically different gender dynamics in educational environments, previous findings are not easily extrapolated to MM countries and a separate investigation is required. To the best of my knowledge, no empirical study so far has investigated the causal impact of teacher gender in MM countries.

The current study examines the impact of teacher gender on student achievement in MM countries. In particular, I seek the answers to these questions: Do girls in MM countries gain academic benefit when taught by a female rather than a male teacher? Do boys in MM countries perform better in classes instructed by a male rather than a female teacher? I also investigate whether and how the impact differs between single-sex and mixed-gender classes and across the countries. Given the high prevalence of single-sex education in MM countries, this research question is highly relevant to be investigated in this context.

Using TIMSS2015 international scores of eighth-graders from the MM countries, the analyses in this chapter estimate the impact of teacher gender on student achievement at secondary educational level when students are more socialized and the moderating role of culture is more prominent. I also focus on test scores in math and science, two STEM subjects with larger potential of being influenced by the gender interactions in classroom due to the typically higher reinforcement of gender stereotypes in these subjects (Lim & Meer, 2015). Using the estimation strategy introduced by Dee (2005), the results of this chapter is free

[52] I used the term "traditional family" rather than "Muslim family" because this study focuses on the role of culture (not religion). While in many Muslim families, women are highly-educated, the example used here refers to certain cultural norms (Unlike men, women do not need to pursue their study up to high degrees), which are somewhat prevalent among *some* Muslim families.

from potential bias caused by unobserved student effects. Additionally, the study accounts for the two-stage sampling design in TIMSS and the interdependence among the units of study in the sampled classrooms by applying a multilevel modelling approach. In an attempt to provide methodological insights for the investigation of the issue, I compare the results obtained from different specifications and various estimation methods. This chapter also contributes to the literature by using a new methodological approach, which combines the advantages of multilevel modeling and first-differencing[53], namely accounting for between-class variability and addressing unobserved student effects.

The next section explains some of the most important channels of the impact and the potential mechanisms at work.

4.2. Theory and Possible Mechanisms

Student-teacher gender interactions could affect academic performance through relevant changes in the behavior of both parties. Following Dee's (2005) suggestion, it would be helpful to differentiate between "active" and "passive" channels of the impact[54].

Active channels refer to gender-based imbalances in classroom interactions resulted from *teachers' mostly-unintended behavioral bias* towards the students of a certain sex. While most teachers are unlikely to systematically discriminate against students of a certain sex (Cho, 2012; Paredes, 2014), pupils usually experience differentiated classroom interactions based on gender (Cho, 2012). This is because the majority of classroom interactions occur at a pace that does not allow teachers to monitor or study their own behavior (She, 2000). As a result, teachers' beliefs and expectations might guide their behavior towards the students. A teacher who attributes higher ability to students of a specific gender might display higher expectations for their academic success (Pygmalion effect[55] as descried by Rosenthal &

[53] Although the term "first-difference" estimation is mostly used in panel data studies with multiple observations over time periods, following Dee (2005) and Cho (2012), I use the same term to refer to differencing the equations across the subjects (not time periods).

[54] It seems that the classification is made from the teachers' perspective. Paredes (2014) used a similar classification referring to "active" and "passive" channels as "those in which the teacher reacts to student's gender" and "those in which the student reacts to teacher's sex", respectively.

[55] In their influential study of Pygmalion effect, Rosenthal and Jacobson (1968) used fake information about students' ability and manipulated teachers' beliefs. They found that students who were falsely identified as of

Jacobson, 1968) by subtly allocating more time, assigning more difficult questions, giving more freedom to call out answers, and providing more positive feedback to them in class (Lim & Meer, 2015). All such mechanisms potentially lead to the higher performance for the presumably high-performing students.

But how does teacher gender play a role in this scenario? Studies have shown that teachers are more inclined to think positively about the academic potential of same-sex students (Ehrenberg et al., 1995; Lavy, 2008; Winters et al. 2013; Gershenson et al., 2016). In particular, it has been shown that female teachers are less likely to hold stereotypical gender beliefs against girls' math ability (Antecol et al., 2015), are more likely to perceive the behavior of boys as problematic (Dee, 2005; Zeeuw et al., 2014), and tend to react less patiently to boys' disruptive behaviors in class (Klein, 2004). Therefore, to the extent that teachers' attitudes towards students are influenced by teacher gender, teacher-student gender-match could have an impact on student achievement.

Passive effects are those triggered by the teacher's identity rather than the teacher's explicit behaviors (Dee, 2005) and operate through *changes in student's attitude and study-related behaviors*. Two related theories have been widely used to explain passive channels, "role model" theory (Almquist & Angrist, 1971; Basow & Howe, 1980), and "stereotype threat" theory (Steele & Aronson, 1995). The former states that the mere presence of a teacher with similar demographic background could enhance students' academic identification and motivation (Ammermueller & Dolton, 2006; Holmlund & Sund, 2008; Egalite & Kisida, 2018). A gender-congruent teacher could better serve as a role model for the students, and improve their performance accordingly (Bettinger & Long, 2005; Paredes, 2014; Egalite & Kisida, 2018).

The latter, the "stereotype threat" theory, suggests that negative stereotypes against one's group might act as a source of distress and apprehension about confirming them, and thereby inhibit the person's productivity (Steele & Aronson, 1995). Feeling worried to be viewed through the lens of negative stereotypes rather than personal merit, the person might prefer

high ability had higher school-year gains (Gershenson et al., 2016). Thus, communication of teacher beliefs and expectations to their students will turn them into self-fulfilling expectations.

not to take any action or feel too anxious when forced to take any. When stereotypical perceptions against female ability are endorsed in classroom environment, female pupils may modify their academic expectations, and unconsciously conform to perceived biases (Egalite & Kisida, 2018). For instance, when a teacher simply does not hear a female student's question or overlooks her comment in math class, she might infer that the teacher considers the comments irrelevant or incorrect (She, 2000). She might therefore refrain from getting involved in classroom discussion, shy away from asking and answering questions, and feel higher anxiety at the exams, all preventing her from performing at full potential[56]. In this regard, the teacher's gender matters because a teacher of the opposite sex is more likely to strengthen the pressure of negative stereotypes (Dee, 2005; Cho, 2012). In contrast, when a student shares gender with the teacher -the most informed and knowledgeable person in classroom- she would be less likely to feel pressure from gender stereotypes against her ability.

Due to data limitations, not many studies on the gender interactions between student and teacher have addressed the underlying mechanisms. Few studies compared the active and passive channels in terms of the effects on students' performance-related outcomes. Using between-subject student fixed-effect specification, Sansone (2017) found a statistically significant association between teacher gender and student's interest and self-confidence in the subject. However, once they controlled for the teacher's behavior, attitudes and expectations, the relationship became indistinguishable from zero. They concluded that teacher gender did not matter per se, and that the ostensible impact stemmed from the possibly unconscious differences between the way male and female teachers treat their students of different sexes (Sansone, 2017). In contrast, Paredes (2014) maintained the positive causal link between having a female teacher and females' outcomes due to the role model effects rather than the teacher's bias effects.

Overall, it is not yet clear-cut form the literature how much each type of the channels or mechanisms contribute to the impact. Nonetheless, the theories and related empirical

[56] A contrasting evidence found by Marsh et al. (2008) is that the level of motivation, anxiety, and persistence among the students assigned to male teachers did not differ substantially from those of students assigned to female teachers in Australia.

evidence certainly illustrate the paramount importance of cultural norms and sociological factors that collectively determine the dominant mechanisms and the overall effect of teacher's gender on academic outcomes. In particular, studies have emphasized on the moderating role of "national context" (Baker et al., 1995) and the relevance of country-level indices such as cultural attitudes towards women, female economic activity, women's political empowerment (Guiso et al., 2008), as well as individual-level beliefs about gender norms (Salikutluk & Heyne, 2017). Therefore, in countries with disparate cultural norms, different mechanisms are possibly at work. While my study in this chapter is not able and does not intend to uncover the underlying mechanisms of teacher gender effect among Muslim societies, by reporting the size and direction of the causal impact, it provides suggestive evidence that the gender of the teacher might operate through different channels in MM countries from those in western countries.

4.3. Literature Review and Knowledge Gap

Scholars and educators mostly investigated the impact of teacher gender to address concerns about the feminization of teacher profession, and how this trend potentially affected the size and direction of gender gaps in student achievement in these countries (eg. Cho, 2012; Lim & Meer, 2015; Sansone, 2017)[57].

While in theory most rationales for student-teacher gender interactions predict a positive impact of gender-matching on students' outcomes, empirical studies in this field have produced mixed results. Several reasons might explain the inconsistency of the results.

First, due to the existence and activation of different mechanisms, it would be expected that the effect differs across various levels of education. At primary level, when children's educational identity and self-perception about ability are formed (Antecol et al., 2015), role model effect might be dominant. Between the age of 7 and 12, as children develop an awareness of commonly-held gendered beliefs and gender stereotypes (Pahlke et al., 2014; Antecol et al., 2015), the related mechanisms come into play. Thus, it is suggested from the literature that student-teacher gender interaction effect is stronger in secondary education,

[57] It has been suggested that recruiting more male teachers could alleviate the gap in favor of males (Carrington et al., 2008; Marsh et al., 2008; Holmlund & Sund, 2008; Zeeuw et al., 2014).

where conformity to gender stereotypical roles become increasingly important (Marsh et al., 2008; Antecol et al., 2015).

Nevertheless, even the literature at, say secondary educational level, offers almost all possible results, ranging from null or negligible effect of teacher gender (Ammermueller & Dolton, 2006; Holmlund & Sund, 2008; Cho, 2012; Sansone, 2017), to positive impact of a certain teacher gender for all students (Klein, 2004; Lim & Meer, 2015), or positive effect of student-teacher gender match (Dee, 2007; Paredes, 2014).

A second source of the inconsistencies relates to the examination system under study. For the evaluation of the impact, studies used different outcome variables. Some research used subjective teacher assessments (eg. Dee, 2005; Gershenson et al., 2016), some analyzed objective internal exam results (eg. Klein, 2004; Holmlund & Sund, 2008), and some research focused on international standardized test scores (Ammermueller & Dolton, 2006; Cho, 2012). It could be the case that teachers' expectations and attitude (active channels of the impact) dominate when subjective teacher assessments are chosen as the dependent variable, while role model effect (passive channel) be stronger when objective test results are taken as the measure.

The inconsistencies could also be attributed to the diverse contexts of the studies. In their early study, using data from the U.S. Educational Longitudinal Study, Ehrenberg et al. (1995) analyzed the students' score gains from 8^{th} to 10^{th} grades and found no evidence for a statistically significant impact of same-sex teachers on students' test scores. Dee's (2005 & 2007) influential papers on teacher-student demographic-match provided contradicting evidence that assignment to a same-gender teacher had academic benefit for students. Ammermueller & Dolton (2006) followed Dee's identification strategy using four waves of TIMSS&PIRLS[58] data for the US and England, and found nearly null impact in the US and only a slight advantage in teaching math to boys by male teachers in the UK. More recently, Winters et al. (2013) analyzed a five-year-period panel dataset from students in Florida, USA. They found that both male and female high school students in their sample performed better when assigned to a female teacher. Furthermore, using Swedish secondary-school panel data,

[58] Progress in International Reading Literacy Study

Holmlund and Sund (2008) found a statistically significant relationship between assignment to a same-sex teacher and student achievement. However, their estimations were not robust to student fixed-effect specification (Holmlund & Sund, 2008). Lim and Meer (2015) utilized the random assignment of students in South Korea. According to their results, while male students did not appear to benefit from assignment to a same-gender teacher, female students performed about 8% of a standard deviation better when they were taught by a female rather than by a male teacher (Lim & Meer, 2015). In Germany, a study by Neugebauer et al. (2011) did not provide support for having a teacher of the same sex. In an attempt to rule out family background confounders, Zeeuw et al. (2014) took advantage of the unique Netherlands Twin Register data and compared the educational outcomes of same-gender twin pairs assigned to teachers of different genders, or of opposite-sex twins taught by one single teacher or two different but same-sex teacher. Like most of the previous studies, their research did not provide support for the assignment of students to same-sex teachers. Finally, in her international comparison, Cho (2012) used four waves of TIMSS data and applied the student fixed-effect strategy to examine the impact of student-teacher gender-match on pupils' test scores. He restricted the analyses to 15 OECD[59]-member countries in which lower secondary education participation was almost universal. According to Cho's (2012) country-specific estimations, student-teacher gender-match had no impact on student test scores in eight countries (Czech Republic, Finland, Hungary, Netherlands, Norway, New Zealand, Slovak Republic, and USA), had a positive impact on boys' achievements in four countries (Canada, Japan, Portugal, and Spain), and improved girls' test scores in the remaining three countries (France, Greece, and Sweden).

Overall, most of the existing studies on gender interaction effect focused on western and non-Muslim countries where social interactions are less gendered and stereotypical views on gender roles are less pronounced. The few existing investigations in MM country contexts did not account for the non-random assignment of the students to teachers and mostly produced inconsistent results. Examples are Saha's (1983) study on 21 less developed countries using IEA data, Okoro et al.'s (2012) study in Nigeria, and UNESCO's (2000) in

[59] Organization for Economic Co-operation and Development

Pakistan. This review therefore, points to a clear gap in the literature, the dearth of research on teacher gender effect in MM countries with distinctive cultural norms and potentially disparate mechanisms and results, a gap which is addressed by the study in this chapter.

4.4. Data

The Trends in International Mathematics and Science Study (TIMSS) is an international test conducted by the International Association for the Evaluation of Educational Achievement (IEA) on a regular basis since 1995. The test assesses math and science skills of the sampled fourth- and eighth-grade students from the participating countries. With a two-stage cluster sampling design, TIMSS draw a random sample of schools in each country and selects one or two *intact*[60] classes from each participating school to take part in the exam. In addition to measuring student performance in certain subjects, TIMSS collects information about students' background and educational environment through a set of questionnaires targeted to students, their teachers, and school principals[61].

In this study, I use TIMSS2015 data of eighth-grade students from eight MM countries (Bahrain, Egypt, Lebanon, Malaysia, Oman, Qatar, Turkey, and United Arab Emirates). The use of TIMSS international test scores enables to exclude any impact of subjective teacher assessment. Furthermore, as the TIMSS data contains each student's performance in two subjects, it is possible to make within-student comparisons and eliminate potential biases generated by unobserved student traits such as ability level.

Among the forty countries participating in TIMSS2015, I considered fourteen countries with a share of Muslims above 50 percent of their populations. Among these countries, Iran, Jordan, Kuwait, and Saudi Arabia were excluded due to the insufficient number of students assigned to teachers of the opposite sex[62]. Additionally, all students with multiple math or

[60] The selection of intact classes in TIMSS sampling design makes this test rather than other international standard tests such as PISA suitable for the purpose of this study as it gives the possibility to calculate the gender composition of classroom for each participant.

[61] More details on the sampling design, imputation method, context variables and generating the plausible values in TIMSS database are available in Martin et al. (2016).

[62] In these four MM countries, single-sex schooling is mandatory or highly prevalent. Due to the high tendency of assigning teachers of the same sex to single-sex schools, less than 5 percent of the country sample size in these countries were assigned to teachers of the opposite sex.

science teachers were dropped from the sample[63]. This attrition led to a loss of lower than 25% of the country samples of Bahrain, Lebanon, and United Arab Emirates. The samples of Kazakhstan and Morocco were totally removed as more than 99% of the students had multiple teachers in either subjects. Thus, the final sample includes 114,442 observations (student-teacher pairs), containing the collected information of 60,359 students (30,330 boys and 30,029 girls) in 2320 classes, linked to 4045 teachers (1798 male and 2247 female teachers) and 1689 schools from eight MM countries (Bahrain, Egypt, Lebanon, Malaysia, Oman, Qatar, Turkey, and United Arab Emirates)[64]. Table 4.1 provides overall sampling information for the countries in the final TIMSS sample used in this study.

Table 4.1. Overall sampling information by country.

Country	Number of students	% female	Number of teachers	% female	Number of schools	Share of students in single-sex schools
Bahrain	4628	46.7	305	52.8	99	81%
Egypt	7776	51.6	421	51.1	210	72%
Lebanon	3831	53.4	340	54.6	137	17%
Malaysia	9637	51.6	575	78.6	206	10%
Oman	8883	49.3	677	50.3	301	83%
Qatar	5054	49.3	395	46.8	121	76%
Turkey	6079	48.4	437	48.1	218	0%
United Arab Emirates	14471	48.5	895	55.9	397	89%
Total	60,359	49.8	4045	55.6	1689	57%

Note: Own calculations using IEA Database for TIMSS 2015-G8. Since the TIMSS data structure does not allow to distinguish between class and school levels (in most cases only one class per school was included in the TIMSS sample), this study does not differentiate between single-sex classrooms and single-sex schools. Thus, the single-sex classes/schools were identified as those to which both girls and boys in the sample were assigned.

The proportion of students in single-sex classrooms are remarkably higher in most Arab countries, and relatively lower in the two countries with the lowest share of Muslim population (Lebanon and Malaysia) and in Turkey, where single-sex education was officially

[63] Since these cases were mainly from certain schools with systematic differences from typical schools (eg. higher socio-economic status), they were not comparable to the observations with single teacher in each subject. Thus, even those cases with same-sex multiple teachers were excluded from the sample.

[64] As in this chapter, I also use differenced models to estimate the causal impact, it is important to note that the stated sampling information relates to the full sample used for non-differenced models. For the differenced models, the number of students is confined to those whose performances have been observed for both math and science subjects. As a result, for the differenced models, the sample consisted of 27,183 boys and 26,900 girls.

banned to promote secular values. Furthermore, while insufficient number of male teachers has been an obstacle for a reliable evaluation of the effect of teacher gender in several previous research (eg. Winters et al, 2013), I exploit the abundance of male teachers in MM countries to investigate the teacher gender effect. According to table 4.1, in most MM countries male teachers constitute a considerable proportion of the teaching profession (somewhat around 23 to 53 percent of the teachers in the country subsamples).

It is also notable that TIMSS uses an elaborate method to measure students' performance. The pool of questions for each subject is divided to five subsets of questions and one is randomly assigned to each participant. With multiple imputation methodology[65] and using the student's responses to the assigned items, five plausible values are generated as measures for student's performance. In this study, I use all five plausible values as the dependent variables as recommended in TIMSS user-guide by Foy and Yin (2015) and Foy (2017).

In order to ensure that the estimation does not confound the effect of teacher gender with that of the major predictors of student performance, besides the *teacher gender* as the main independent variable, certain explanatory variables are extracted from TIMSS questionnaires and added to the models. Regarding the determinants of student performance used in previous studies (eg. Kramarz et al., 2008; Hanushek & Woessmann, 2017; Hanushek et al., 2019), the following variables are added as controls.

- *student's age* (measured in years by two decimal points to control for possible grade repetitions),
- *student's migration background*[66] (a dummy variable indicating whether the student always/often speaks the language of the test at home or only sometimes/rarely does so),

[65]See Martin et al. (2016) for more details.

[66]Unfortunately, the proportion of the observations with missing information about the place of own/parents' birth was high. Thus, given that one of the main channel through which migration background might affect student performance is regarded as language difficulties, I merely used the language dummy to account for students' migration background. After all, the focus of this study is not on the effect of migration or language proficiency.

- *student's home educational resources* (a TIMSS-constructed categorical variable, *the level of educational resources* such as books, Internet, study room, etc. available at home),
- *teacher's experience* (a categorical variable with five categories using percentiles of teacher's years of experience, from lower than 5 years of experience to above 23 years for math teachers, and 24 years for science teachers),
- *teacher's education* (a categorical variable with four categories: below bachelor, bachelor degree, beyond bachelor, and non-response[67]),
- *teacher's major in post-secondary study* (a categorical variable with four categories: no formal education after upper secondary level or majored in an unrelated field[68], majored in teaching but not the subject itself, majored in the subject but not in teaching the subject, majored in the subject and teaching the subject),
- *class size* (calculated by grouping the students with the same school and class ID in each country to prevent potential bias resulted from higher probability of allocating teachers of a certain gender to larger classrooms),
- *single-sex dummy* (to control for systematic differences between the single-sex and coeducational schools[69] in terms of relevant factors such as curriculum, emphasis on academic success, teachers' motivation, etc.),
- *School SES* (whether more than 50% of the student body in the school are from disadvantaged versus affluent families or the shares are balanced between social classes, (Cordero et al., 2017; Eriksson et al., 2019),
- *school discipline problems* (a categorical variable constructed by TIMSS to show the degree of frequency in discipline problems such as theft, bullying, etc. in school).

[67]The non-response rate for this item was rather high (exceeding 15 percent of the data in some countries such as the UAE, Bahrain and Egypt). In order to prevent the loss of a large share of the sample size due to missing information and also avoid a potential risk of bias due to selective attrition, I considered a separate category for missing information on teacher education.

[68] As the number of teachers with no formal education or some unrelated study were few (lower than 10 percent of the teachers), and given that I have a separate variable for educational level, I combined the two categories as having no formal post-secondary education or majoring in an unrelated field to teaching the subject.

[69]As in most cases only one class per school has participated in the exam, in this chapter no distinction has been made between the gender composition of classrooms and that of the schools.

In contrast to Dee (2007) and like Cho (2012), I could not use school or classroom fixed-effects because in TIMSS sample only one or two intact classes from each school participated in the test, and doing so would cause a perfect collinearity with the class-level teacher gender variable in most cases.

Table 4.2 shows the main descriptive statistics for the pooled sample of eight MM countries examined in this study by student-teacher gender-pair.

Table 4.2. Summary of descriptive statistics for the pooled sample by student-teacher gender pair.

Variable	Boys		Girls	
	Male teacher	Female teacher	Male teacher	Female teacher
Student-level variables				
Math score	416.21	455.31	441.90	445.07
	(2.06)	(2.89)	(3.08)	(1.45)
Science score	424.46	456.39	447.30	466.52
	(2.31)	(3.50)	(3.50)	(1.67)
Age	14.03	14.09	14.06	13.98
	(0.75)	(0.66)	(0.65)	(0.69)
Language dummy				
Sometimes/never speak the language at home	0.30	0.39	0.35	0.32
Always/often speak the language at home	0.70	0.61	0.65	0.68
Home educational resources				
Few resources	0.21	0.19	0.24	0.20
Some resources	0.71	0.73	0.68	0.72
Many resources	0.08	0.08	0.08	0.08
Class- and school-level variables				
Teacher experience				
Lower than 6 years	0.17	0.28	0.22	0.26
Between 6 and 10 years	0.22	0.28	0.20	0.26
Between 11 and 15 years	0.21	0.16	0.20	0.20
Between 16 and 23 years	0.24	0.17	0.20	0.18
More than 23 years	0.17	0.11	0.17	0.09
Teacher education				
Below Bachelor	0.05	0.07	0.08	0.05
Bachelor degree	0.70	0.64	0.68	0.72
Beyond bachelor (master or doctorate)	0.13	0.21	0.15	0.14
Non-response	0.12	0.07	0.09	0.09
Teacher Major (post-secondary study)				
No-formal/unrelated education beyond upper-secondary	0.06	0.11	0.11	0.08
Majored in teaching the subject, but not the subject	0.16	0.15	0.17	0.15
Majored in the subject, but not in teaching the subjects	0.45	0.40	0.35	0.43
Majored in teaching the subject and the subject	0.34	0.34	0.37	0.34
Single-sex dummy				
Mixed class	0.31	0.69	0.67	0.38
Single-sex class	0.69	0.31	0.33	0.62

Table 4.2 - continued. Summary of descriptive statistics for the pooled sample by student-teacher gender pair.

Variable	Boys		Girls	
	Male teacher	Female teacher	Male teacher	Female teacher
Class size	29	30	30	27
	(9)	(9)	(9)	(8)
School SES				
More disadvantaged	0.36	0.41	0.45	0.39
More affluent	0.37	0.31	0.27	0.30
Almost balanced shares	0.28	0.28	0.29	0.30
School discipline problem				
Hardly any problem	0.38	0.42	0.37	0.50
Some problems	0.43	0.41	0.39	0.36
Serious problems	0.19	0.16	0.23	0.14
Number of students	19,462	10,868	6,926	23,103
Number of classes	1291		1029	
Number of schools	935		754	

Note: Own calculations based on IEA Database for TIMSS 2015-G8. For continuous variables, the values represent the mean and standard deviations (in parentheses). For categorical variables, the numbers show the proportions of individual observations (students) in the respective subgroup. For the performance variables (math and science scores), the means and estimated standard errors (in parentheses) were calculated using all five plausible values based on the Item Response Theory as recommended by the IEA guidelines (Foy and Yin, 2015). The statistics for class-size variable have been rounded to the nearest integer.

The mean performance for each group has been estimated using five plausible values for each student performance in each subject. Therefore, for these *estimated* variables, the numbers in parentheses show standard errors, not the standard deviations. Given that the metric for the national average mathematics and science performance in TIMSS has been set to the mean of 500 and the standard deviation of 100 for all participating countries in each wave (Martin et al., 2016), the below-500 means for all subgroups in both subjects reflect the mix of lower-than-average-performing countries in the selected sample.

As shown by table 4.2, in both math and science subjects, the groups of students with female teachers considerably outperformed their counterparts with male teachers. This however could not be taken as an evidence for the academic benefit of assignment to female teachers because students might be *non-randomly* assigned to teachers of different sexes. For example, the ostensible improvement in the average scores in case of assignment to female teachers might stem from other reasons, say in this case the higher access to educational resources at home, or lower degrees of discipline problems in school for the outperforming groups. The benefit might indeed exist or even be larger than it initially seems since, for

example among boys, the outperforming group were in worse condition with respect to their migration background or class size[70]. Furthermore, other confounding factors such as single-sex environment or country-level variables (such as the examination system and overall education policy) might intervene. For instance, it could be the case that in countries with more rigorous education systems and higher-performing students, the teaching profession was more feminine. In this case, the effect of such relevant country-level factors are captured as the impact of having a female teacher.

To more solidly assess the need for controlling the extracted variables in my analyses, I used the t-test for continuous variables and the chi-square test for categorical variables to examine the distribution of students across the range/levels of explanatory variables. The large test-statistics (t-statistics and Pearson chi-square statistics) and thereby small p-values for the tests revealed a statistically significant association between teacher gender and each of the confounding factors at different levels. In other words, the possibility of *gender-based sorting* of teachers was not rejected by the statistical tests. Therefore, it would be crucial to control for the extracted variables at different levels.

It is also notable in table 4.2 that the share of students in single-sex classes are exceptionally high among the subgroups with matching-gender pairs, probably due to the general tendency of single-sex schools for matching the gender of the assigned teachers with students'.

4.5. Method

To examine whether the teacher gender affects student academic outcomes using TIMSS data, I construct two models with different hierarchical approaches, multiple linear regression model (MLR) and hierarchical linear model (HLM). Each model is estimated using two different strategies to address causal concerns[71]. The selected approach is then used to investigate whether the impact differs between single-sex and mixed-gender classrooms and across countries.

[70]Angrist and Lavy (1999) maintained that at least for higher graders, reducing the number of students in classroom promotes better performance.

[71] As Wooldridge (2002) suggests, I distinguish between the "models", which *define the relationships between variables*, and their "estimable equations" which *specify how the parameters in the model are estimated*.

As a general remark, all coefficients are allowed to vary between genders as separate models are estimated for girls and boys. Additionally, to account for the elaborate sampling design in TIMSS and prevent skewed results, I use relevant sampling weights as recommended by Foy (2017). Accordingly, to avoid larger countries disproportionately affecting the estimates in the analyses of the pooled sample, I use "SENWGT" (a transformation of "TOTWGT" that creates a weighted sample size of 500 in each country). For country-specific analyses, I use "HOUWGT" to ensure that the weighted sample corresponds to the actual sample size in each country[72]. Since in mixed-effect models, weights enter into the log likelihood at both class and student levels, in the hierarchical approach the TIMSS school-level weight variable "SCHWGT" is also applied. Following Rabe-Hesketh and Skrondal (2006) about the need for rescaling in hierarchical models, sampling weights are then rescaled to sum to the cluster (class) size. Additionally, as suggested by Foy and Yin (2015), all estimations are conducted five times (once for each plausible value), and the final results are aggregated across the five values.

4.5.1 The effect of teacher gender on student achievement

Multiple Linear Regression Model (MLR)

Using the selected variables indicated in data section (4.4), an MLR model explains student achievement as a linear function of explanatory variables plus a random error term for individual deviations from the averages as follows:

$$y^{s}_{ijkc} = \alpha + \Sigma^{p}_{\mathrm{n}=1}\beta_n I_{nijkc} + \Sigma^{m}_{\mathrm{n}=1}\gamma_n T^{s}_{njkc} + \Sigma^{q}_{\mathrm{n}=1}\delta_n C_{njkc} + \Sigma^{z}_{\mathrm{n}=1}\eta_n S_{nkc} + \lambda Subj^{s} + CFE_c + \epsilon_{ijkc} \qquad (4.1)$$

In equation 4.1, y^{s}_{ijkc} stands for the test score of student *i* in class *j* of school *k* from country *c* in the subject *s* (mathematics or science) exam. Explanatory variables include *p=3* individual-level variables in vector *I* (*age, language dummy, the level of home educational*

[72]Another widely-used individual-level sampling variable introduced by TIMSS is "TOTWGT", which according to Foy (2017) inflates sample sizes to estimate the population size. As this study is seeking the impact of a specific variable rather than inferring national population metrics, I used "HOUWGT" instead. The same weighting variable has been used by previous secondary analyses of TIMSS data such as Caponera and Losito's (2016).

resources), *m=4* subject-*s* teacher characteristics[73] in vector T^s (*teacher gender, teacher experience, teacher education, and teacher major), q=2* class-level variables in vector *C (single-sex dummy, and class size*), and *z=2* school-level variables in vector *S* (*overall socio-economic status of the student body in school and the level of discipline problems in school*). The α denotes the average intercept for all observations of subject *s* performance in the sample. The error term ϵ_{ijkc} captures the unobserved variability between the individual students due to omitted variables such as student's ability, motivation, etc. Since the subject dummy ($Subj^s$) captures the subject fixed-effects, the intercept and the error term are assumed to be constant across the subjects. Finally, the effects of the national educational systems and procedures and of other relevant country-level variables are captured by the country fixed-effects (CFE_c).

As a starting point, I estimate equation 4.1 by the OLS method for the pooled sample of all observations (pooling across the countries and subjects). However, the major problem with the OLS approach is that the estimated effect likely suffers from omitted variable bias due to the violation of the zero-conditional-mean assumption. If for example, students with lower ability level were more likely to be assigned to female teachers, the model above would overestimate the potential benefit of having a male rather than a female teacher.

Nevertheless, containing a pair of two observations per student, the TIMSS data structure provides an opportunity to avoid the potential threats to causality that come from student fixed-effect. Thus, as Dee (2005) innovatively proposed, I exploit the paired nature of TIMSS data to difference out all student-level variables that have possibly remained constant across the subjects, including unobserved ability and motivation. To find the differenced estimable equation, the subject-specific MLR models are written with the intercept and error term containing subject subscripts (Ma and Sc for math and science respectively). Assuming that the unobserved student factors (ϵ_{ijkc}^{Ma} and ϵ_{ijkc}^{Sc}) are decomposable into those that are constant

[73] These variables are also defined at class-level. However, as the teachers are different for the subjects, a separate vector is defined for the subject-specific teacher characteristics (T^s), for which the upper index *s* denotes the subject.

across the subjects (μ_{ijkc}) and those that vary between the subjects (u^{S}_{ijkc}), the subject-specific MLR models read as:

$$y^{Ma}_{ijkc} = \alpha^{Ma} + \Sigma^{p}_{\mathrm{n}=1}\beta_n I_{nijkc} + \Sigma^{m-1}_{\mathrm{n}=1}\gamma_n T'^{Ma}_{njkc} + \theta TchrF^{Ma} + \Sigma^{q}_{\mathrm{n}=1}\delta_n C_{njkc} + \Sigma^{z}_{\mathrm{n}=1}\eta_n S_{nkc} + CFE_c + \mu_{ijkc} + u^{Ma}_{ijkc}$$

$$y^{Sc}_{ijkc} = \alpha^{Sc} + \Sigma^{p}_{\mathrm{n}=1}\beta_n I_{nijkc} + \Sigma^{m-1}_{\mathrm{n}=1}\gamma_n T'^{Sc}_{njkc} + \theta TchrF^{Sc} + \Sigma^{q}_{\mathrm{n}=1}\delta_n C_{njkc} + \Sigma^{z}_{\mathrm{n}=1}\eta_n S_{nkc} + CFE_c + \mu_{ijkc} + u^{Sc}_{ijkc}$$

in which the vector T' includes teacher traits except for the teacher gender dummy ($TchrF$) which equals one for female and zero for male teachers. The differenced estimable equation thus reads:

$$y^{Ma}_{ijkc} - y^{Sc}_{ijkc} = (\alpha^{Ma} - \alpha^{Sc}) + \Sigma^{m-1}_{\mathrm{n}=1}\gamma_n\left(T'^{Ma}_{njkc} - T'^{Sc}_{njkc}\right) + \theta\,(TchrF^{Ma} - TchrF^{Sc}) + (u^{Ma}_{ijkc} - u^{Sc}_{ijkc}), \qquad (4.2)$$

which no longer contains subject-invariant factors at student, class, school, or country levels. A positive and statistically significant estimation for the coefficient θ would indicate that assignment to a female rather than a male teacher improves student scores.

According to equation 4.2, for the first-differenced[74](FD) estimator θ to be consistent, the identifying assumption of strict exogeneity (corr(Δu_{ijkc} , ΔT_{jkc}) = 0) is translated into the condition that the unobservable student fixed-effects that might correlate with teacher gender are not subject-specific, and thus do not exist in the differenced equation. In other words, I assume that the possibly nonrandom sorting of students was not based on those unobserved students traits that might vary between the subjects. This is a weaker and thus more plausible assumption than what we need to suppose for the consistency of the OLS estimation. While the non-differenced MLR model requires that the student's ability or motivation be

[74]The term "first-differenced" estimator is probably more appropriate to be used in the analyses of panel data where the sequential nature of the information over time allows using the term "first" in the differenced equation. While in TIMSS data, each individual's performance is observed for two subjects rather than in ordinal time periods, I use the common name for the estimator (first-differenced or FD-estimator). The same terminology has been used in prominent previous studies such as Cho's (2012).

uncorrelated with teacher gender, the FD estimator merely requires the students to be equally able or motivated to do well in math and science.

Nonetheless, equation 4.2 assumes that the teacher gender effect is the same across the subjects. This could be violated if for example, the role-model effect or stereotype threats are stronger in either subject. Relieving this assumption leads to equation 4.3 in which the coefficients of teacher gender for each subject could be estimated separately:

$$y_{ijkc}^{Ma} - y_{ijkc}^{Sc} = (\alpha^{Ma} - \alpha^{Sc}) + \Sigma_{n=1}^{m-1}\gamma_n\left(T'^{Ma}_{njkc} - T'^{Sc}_{njkc}\right) + \theta^{Ma}TchrF^{Ma} + \theta^{Sc}(-TchrF^{Sc}) + \left(u_{ijkc}^{Ma} - u_{ijkc}^{Sc}\right) \quad (4.3)$$

The parameters θ^{Ma} and θ^{Sc} show the average change in the student's score when assigned to a female rather than a male teacher in the subject. While in equation 4.3 I assume for simplicity that the impacts of other teacher traits are the same between the subjects, one could easily estimate entirely new vectors of coefficients for each subject[75].

Hierarchical Linear Model (HLM)

The MLR model does not account for the nesting structure of the data, i.e. students nested in classes, nested in schools, and thereby ignores the potential and likely dependence between the observations from the same classroom. Indeed, the MLR model assumes that the whole variation in the outcome variable comes either from observed differences (in terms of student-, class-, and school-level variables as well as country fixed-effects) or from *individual* deviations from the averages. However, one could reasonably argue that different classrooms might deviate from the average country-specific performance by *a random class-level error term*. To account for the systematic differences between the sampled classrooms in each country, a hierarchical model (HLM) introduces a random effect at classroom level as in equation 4.4:

[75]With only two observations per units of study -student- the estimation of the first-differenced and fixed-effect approaches are exactly identical. However, I focus on the differenced equation (FD estimator rather than FE estimator) since the same approach is going to be used in the hierarchical modeling approach as well and technical problems did not let me to use fixed-effect estimation for the HLM model (the FE estimation of the hierarchical model *with plausible values* did not converge).

$$y^s_{ijkc} = \gamma_{0000} + \Sigma^{p}_{\mathrm{n}=1}\gamma_{n000}I_{nijkc} + \Sigma^{m}_{\mathrm{n}=1}\gamma^{s}_{0n00}T^{s}_{njkc} + \lambda_{0500}Subj^{s} + \Sigma^{q}_{\mathrm{n}=1}\gamma'_{0n00}C_{njkc} + \Sigma^{z}_{\mathrm{n}=1}\gamma_{00n0}S_{nkc} + CFE_c + \nu_{0jkc} + \epsilon_{ijkc} \quad (4.4)$$

In the above "mixed" model, the same outcome variable (y^s_{ijkc}) is estimated as a linear function of certain fixed parameters plus *random errors at student and class levels*. In the fixed part, the vectors of explanatory variables are defined likewise the MLR model, but the subscripts of the coefficients follow the standard notation in HLM literature in which the non-zero digit represents the level at which the respective independent factor varies. In the random part, the HLM model accounts not only for the *within-class (between-student)* variations, but also for the *between-class variations* that have not been explained by the class-level parameters in the fixed part[76]. In fact, the term ν_{0jkc} adds a random offset to the intercept for each classroom ("random intercept"), reflecting the potential impacts of *unobserved class-level confounders* such as the degree of emphasis on academic success or competition in classroom, overall teachers' motivation, or curriculum. These factors are probably identical for all individuals in the same classroom, but randomly vary across the classrooms. Using the maximum likelihood estimation (MLE) method, the mixed model above is estimated for the pooled sample of the eight MM countries.

Here again, for a consistent estimation, both error terms (ϵ^s_{ijkc} and ν^s_{0jkc}) must be uncorrelated with the explanatory variables. This is however, not a tenable assumption. After all, the unobserved-ability-problem with the MLR approach is carried over to the mixed model due to the potential correlation between teacher gender and unobserved class-level factors such as the *omitted peer ability variable*[77]. To deal with this issue, with a similar approach, I assume that some of the relevant but unobservable class characteristics such as the overall peer ability, the level of competition in classroom, the level of available facilities in class, or part of the teacher motivation that relates to the overall compensation system in

[76]As the variable of interest in this study (teacher gender) lies at the class level, adding random errors at the higher levels of school and country would not be necessary.

[77]Several empirical studies maintained that apart from the ability level of the student, the overall ability of the peers in classroom plays a key role in determining individual performance. See for example, Epple and Romano (2011), and Sacerdote (2011).

the school[78] are constant between the subjects. Thereby, the same differencing approach is used to cancel these subject-invariant factors out in the estimable equation. With both error terms in the mixed model above substituted with their respective subject-invariant and subject-specific components ($\epsilon^{s}_{ijkc} = \mu_{ijkc} + u^{s}_{ijkc}$, and $\nu^{s}_{0jkc} = \eta_{0jkc} + r^{s}_{0jkc}$), the subject-invariant error terms at both student and class levels (μ_{ijkc} and η_{0jkc}) are removed in the differenced equation 4.5:

$$y^{Ma}_{ijkc} - y^{Sc}_{ijkc} = (\gamma^{Ma}_{0000} - \gamma^{Sc}_{0000}) + \Sigma^{m-1}_{n=1}\gamma_{0n00}\left(T'^{Ma}_{njkc} - T'^{Sc}_{njkc}\right) + \theta(TchrF^{Ma} - TchrF^{Sc}) + \left(r^{Ma}_{0jkc} - r^{Sc}_{0jkc}\right) + \left(u^{Ma}_{ijkc} - u^{Sc}_{ijkc}\right) \quad (4.5)$$

or

$$y^{Ma}_{ijkc} - y^{Sc}_{ijkc} = (\gamma^{Ma}_{0000} - \gamma^{Sc}_{0000}) + \Sigma^{m-1}_{n=1}\gamma_{0n00}\left(T'^{Ma}_{njkc} - T'^{Sc}_{njkc}\right) + \theta^{Ma}TchrF^{Ma} + \theta^{Sc}(-TchrF^{Sc}) + \left(r^{Ma}_{0jkc} - r^{Sc}_{0jkc}\right) + (u^{Ma}_{ijkc} - u^{Sc}_{ijkc}), \quad (4.6)$$

if heterogeneous effects across the subjects are allowed.

The advantage of the FD approach to estimate the mixed (HLM) model is the cancellation of the subject-invariant factors and thereby removing the potential bias that could have been produced otherwise. In the estimation of equations 4.5 and 4.6, the strong assumptions of *no correlation between teacher gender and unobserved student- and class-level variables* such as individual ability and ability peer effect are not needed. We only need to assume that the *changes* in these factors between the subjects, not the factors themselves, are uncorrelated with teacher traits (random with a zero-conditional-mean). More specifically, at individual level, we must only assume that a given student is equally able to perform well in math and science. Likewise, at class level, we no longer need to assume that male and female teachers are equally likely to be assigned to classrooms with overall higher ability, an assumption which could easily be violated if for example highly-selective private schools are more likely to hire male/female teachers. Instead, we could merely aggregate the individual-level ability assumption over the class-level, i.e. we suppose that the overall ability level of peers in one

[78]In contrast, subject-specific class-level variables include among others the teacher's instructing methods, and the ability for managing student interactions in class and promote a good learning environment.

classroom does not change between the subjects. The coefficients θ in equation 4.5, and θ^{Ma} and θ^{Sc} in equation 4.6 would then be consistent estimators for the impact of teacher gender on student performance.

4.5.2 Heterogeneity by single-sex and coeducational classrooms

Although the gender composition in classroom has been controlled in the models, the coefficients are restricted to have the same impact in single-sex and coeducational classrooms. However, based on the underlying theories and mechanisms of the impact, it seems reasonable to expect a differing impact between the single-sex and coeducational environments. Since the single-sex dummy is differenced out in the FD approach, it is not possible to examine the potential heterogeneity by interacting single-sex dummy with the teacher gender variable. Therefore, I estimate the selected model separately for either type of the classrooms so that all coefficients may vary across these types.

4.5.3 Heterogeneity of the impact across countries

Pooling the observations across different countries with cultural proximity, and thus, acquiring higher estimation power might seem reasonable. However, different education systems in various countries might induce considerable heterogeneities in the impacts of context factors, including teacher gender. In the result section (4.6), I also provide the country-specific estimations of the teacher gender impact. The equations at country level are rather similar (except for the removal of country dummies), but different sampling weights are used in the estimations for the pooled sample and for individual country-samples.

4.6. Results

4.6.1 Teacher gender and student achievement (pooled sample)

Table 4.3 reports the coefficients of female-teacher dummy estimated by different equations stated earlier in section 4.5.1 for the pooled sample. For the sake of comparison, the naïve estimator is also provided in the first two rows (SLR model). Accordingly, in MM countries students' performances are positively associated with having a female teacher. On average, male and female students who had female teachers outperformed their counterparts with male teachers by around 27 and 6 points respectively (both coefficients are different from zero at

α =0.01 and α =0.05 statistical significance levels). As shown in table 4.3, when student's background and educational environment's characteristics are taken into account by the OLS estimation of the MLR mdoel, the boys' coefficient is remarkably attenuated and becomes unlikely to differ from zero. Imposing the ceteris paribus condition on the OLS estimation reverses the sign for girls' coefficient of female-teacher dummy, implying that with same student traits and educational environment characteristics, girls do better by almost 5 points when assigned to male rather than female teachers. The impact is not statistically significant at α = 0.05 or α = 0.01 though.

Table 4.3. Estimated teacher-gender effect by student gender using different modeling approaches.

Model	Estimation Method	Student Gender	Female Teacher	S.E.	Observation	Student	Class	School	Country
SLR	OLS	Boys	27.19***	(3.10)	57513	30330			
		Girls	5.88**	(2.91)	56929	30029			
MLR	OLS	Boys	5.45	(3.56)	45982	25040			
		Girls	-5.02*	(2.84)	45749	24986			
	FD	Boys	-14.19***	(2.68)	24780	24780			
		Girls	-16.66***	(3.16)	24365	24365			
HLM	MLE	Boys	-10.86***	(2.79)	45982	25040	1409	1080	8
		Girls	-13.86***	(3.34)	45749	24986	1389	1070	8
	FD	Boys	-12.78***	(3.39)	24780	24780	1354	1088	8
		Girls	-13.36***	(3.71)	24365	24365	1320	1071	8

Note: Own calculations based on IEA Database for TIMSS 2015-G8. The fourth and fifth columns report the main estimation results, coefficients of female-teacher dummy variable (which equals 1 for female and 0 for male teachers) estimated by each estimation strategy and the standard errors (in parentheses). While the pooled sample included 30,330 boys and 30,029 girls, the number of observations for the non-differenced estimation strategies (SLR-OLS, MLR-OLS and HLM-MLE), exceeds the sample size because for the majority of the students two observations exist, one for math and one for science. For the differenced estimation strategies (MLR-FD and HLM-FD), the sample is confined to the students whose performances have been observed for both math and science subjects (27,183 boys and 26,900 girls). However, the number of observations for the differenced models is smaller than the paired sample size due to the missing information for different control variables for around 2500 boys and almost 2500 girls across the eight countries. For the multilevel models (HLM), the number of cases at each level (student, class, school and country) is reported. As recommended by the IEA guidelines (Foy and Yin, 2015), I used all five plausible values provided by TIMSS for all estimations. Stars represent statistical significance levels. * p < .10 , ** p < .05, and *** p < .01.

Nevertheless, the possibility of nonrandom sorting of students to male and female teachers based on unobserved factors such as student ability threatens the causality in the OLS estimation. Differencing out subject-invariant factors within students' observations leads to a large transformation in the estimated effects. The coefficients both become negative and statistically significant (at 0.01 level), indicating that boys and girls have outperformed by around 14 and 17 points in TIMSS exams in case of assignment to a male rather than a female

teacher. Using equation 4.3, I checked whether the FD-estimator of the MLR model varied between the subjects. Accordingly, θ^{Ma} was estimated as of -9.57 for boys and -24.85 for girls (both statistically significant at α = 0.01), and θ^{Sc} as of -17.86 (statistically significant at α = 0.01) for male and -4.04 (not statistically significant) for female students.

Accounting for the interdependence among the sampling units in the same classroom (or the so-called *between-class variability*) by equation 4.4, the random intercept model for the pooled sample indicates that on average boys and girls with female teachers have underperformed their counterparts with male teachers by about 11 and 14 points respectively. Interestingly, while the OLS estimations of the MLR model with additional controls are highly distant from the FD-estimators of the impact, the mixed model produces somewhat similar results to the FD model. If we take Dee's (2005, 2007) fixed-effect approach as reliably estimating the teacher gender causal impact, this implies that even when a cross-sectional hierarchical dataset does not allow for individual fixed-effect estimation (i.e. when only one observation per individual exists), the bias could largely be reduced by taking into account the multi-stage sampling structure and using multilevel rather than one-level models. This implication is further checked by the country-specific models. Finally, the FD-estimation of the random intercept model indicates that boys and girls have performed nearly equally worse (by around 13 points) when assigned to female rather than male teachers. Using equation 4.6 for estimating θ^{Ma} and θ^{Sc} separately shows that both boys and girls gain significant benefit from having a male math teacher (θ^{Ma}equals -11.43 for boys and -21.02 for girls, both statistically significant at α = 0.01), while only boys benefit from being assigned to a male science teacher (θ^{Sc}equals -13.97 for boys and -3.31 for girls, only boys' coefficient is statistically significant at α = 0.01).

For brevity reasons, the coefficients of control variables estimated by the models above are provided in the appendix, table A.4.6 for non-differenced models, and table A.4.7 for the differenced models. A brief overview on the coefficients in table A.4.6 indicates that the sign/direction of the associations are mostly in line with the expectations. For instance, having access to more educational resources at home, speaking the same language as the test's at home, having a teacher with higher educational level, or studying in a more orderly and safe school are all associated with higher individual test scores. Having a look at the

estimated coefficients in table A.4.7, it is notable that when subject-invariant factors at all levels are ruled out, none of the estimated coefficients for the differenced teacher-traits are as large as that of the teacher gender. While the existing low variation in the differenced traits might have led to large standard errors and thereby non-statistically-significant effects, it is important that the teacher gender effect is substantially large and still statistically distinguishable from zero.

4.6.2 The moderating role of class gender composition

Table 4.4 shows the coefficients of female-teacher dummy acquired from the FD estimation of the HLM model (FD-mixed estimator) for the stratified sample of students in single-sex and coeducational classrooms. The table shows that the positive impact of having a male teacher is quite unbalanced across the two types of classrooms. For both genders, the size of the impact is considerably larger in single-sex classrooms. While in mixed-gender classrooms, boys and girls scored respectively around 11 and 10 points higher in case of having a male rather than a female teacher, in single-sex environment having a male teacher produced an academic benefit of around 19 and 22 points for male and female students respectively (with all coefficients statistically highly significant).

Table 4.4. The FD-Mixed estimator for teacher gender effect in single-sex and coeducational classrooms by student gender.

Classroom Gender Composition	Student Gender	Female Teacher	S.E.	Student	Class	School	Country
Mixed-gender classrooms	Boys	-10.90***	(3.89)	10544	801	649	8
	Girls	-10.17***	(4.06)	10649	802	650	8
Single-sex classrooms	Boys	-18.71***	(5.13)	14236	553	441	7
	Girls	-22.27***	(6.28)	13716	518	423	7

Note: Own calculations based on IEA Database for TIMSS 2015-G8. The third and fourth columns report the coefficients of female-teacher dummy variable (which equals 1 for female and 0 for male teachers) and the standard errors (in parentheses) estimated by the FD-Mixed model separately for mixed-gender and single-sex classrooms. The number of observations for the FD-Mixed model equals the number of students reported in the fifth column. The number of cases at each level (student, class, school and country) is reported. As recommended by the IEA guidelines (Foy and Yin, 2015), I used all five plausible values provided by TIMSS for all estimations. Stars represent statistical significance levels. * $p < .10$, ** $p < .05$, and *** $p < .01$.

It is important to note that since the FD-mixed estimation strategy uses *within-student comparisons*, all subject-invariant factors are ruled out from the causal links. Therefore, the estimated effects in table 4.4 are not vulnerable to the common counterargument that the single-sex and coeducational classrooms might systematically differ in several unobserved variables that might have caused the different performances. As these systematic differences such as the selectivity in admitting students and hiring teachers, or school facilities are unlikely to vary *between the subjects*, they could not constitute an alternative explanation for the variations in student achievement.

4.6.3 Heterogeneous impacts across countries

Figure 4.1 shows the country-specific impacts of the teacher gender (the coefficient of female-teacher dummy) estimated by each of the four estimation strategies explained in the method part.

Importantly, the remarkable discrepancies between the country-specific estimations in the first row with those of other methods reflect the potential threat for large biases if one merely relies on cross-sectional controlled analysis. The MLR-OLS estimations likely suffer from biases due to the omitted ability and other relevant factors at different levels.

When subject-invariant factors are differenced out in the MLR model (the second row), the signs of most countries' coefficients are reversed, and the precision of the estimations increases (the confidence intervals get tighter). Moreover, the variability of the impact across the countries decreases. According to the FD-estimators of the MLR model, except for males from Oman and the UAE, students from most countries are not affected by the gender of their teachers at any statistically significant level.

The comparison among the country-specific coefficients obtained from different estimation strategies also suggests that when multilevel data are to be examined, one might improve the consistency and precision of the estimated effects by accounting for the sampling design and interdependencies among the units of observation.

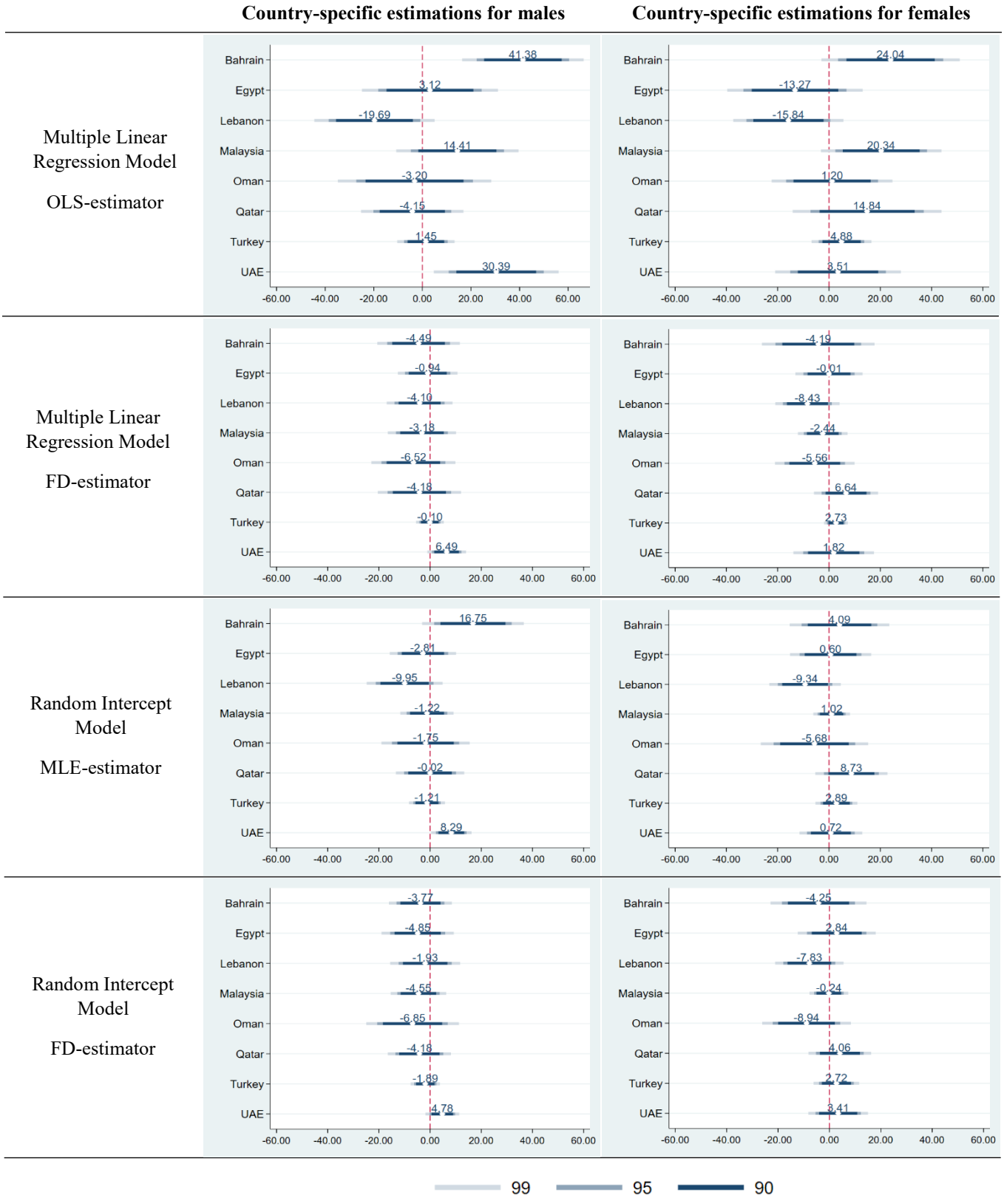

Figure 4.1. Country-specific coefficients of female-teacher dummy, produced by different estimation strategies

To illustrate, while the OLS coefficient of teacher gender for male students from the UAE equals 30.39, the estimated effects by the differenced equation (row 2) and mixed-effect model (row 3) turn out as 6.49 and 8.29 respectively. The latter two estimations are much closer in size and do not vary in such a large range as does the former OLS estimation. Some of the other countries (eg. Egypt and Malaysia) show similar patterns, indicating that it worth to at least take into account between-cluster variabilities when analyzing multi-level data.

The bottom row shows the country-specific coefficients of teacher gender estimated by an FD-estimator of the random intercept model, which not only differences out all subject-invariant factors, but also accounts for the sampling structure of TIMSS data. In contrast to the statistically significant and relatively sizable FD-Mixed estimator for the pooled sample of all countries (more than 10 percent of a standard deviation), it appears that most of the country-specific estimates are not statistically significant. In fact, the only subgroup whose performances were improved (by around 5 points, statistically significant at 5% level) in case of having a female rather than a male teacher were boys from the United Arab Emirates.

Similarly, in her investigation of the impact in 15 OECD countries using differenced equations across the subjects, Cho (2012) also found most of the country estimations indistinguishable from zero. The most likely explanation for the statistically non-significant country-specific impacts is the technical limitations imposed by differencing the equations. In his explanation of possible limitations of estimating differenced equations, Wooldridge (2012) points to the potentially large reduction in the variations of explanatory variables, particularly categorical variables, which leads to a small denominator in the calculation of coefficients' variance, and thereby large standard errors and low precision of the estimated coefficients. In the pooled sample analysis, I benefited from efficiency gains by providing relatively higher variations in explanatory variables. Therefore, despite their statistical insignificance, the size of the FD-estimators of country-specific impacts deserves the attention from educational policymakers.

4.7. Discussion

This section briefly discusses some of the major concerns which might be raised about the methodology and the interpretation of the results stated in this chapter.

The study focused on the impact of teacher gender in MM countries. Despite the high relevance of gender issues in such cultural contexts, these countries have remained nearly untouched in terms of such research, mainly due to the lack of nationally representative quality data. Some previous studies using international dataset intentionally excluded non-OECD countries, including Muslim countries, from their analyses on the ground that secondary-level education is non-universal in these countries[79] and the impact might differ for the selective sample of students who, at any age, *chose* to pursue their studies at secondary level[80]. While this self-selection indeed imposes an important limitation on the interpretation of the results, in this chapter, I emphasize the high relevance of the topic and the non-applicability of previous results to these contexts. Thus, instead of ruling out such rather gendered-culture countries from the analyses, this chapter uses the available rich data collected by TIMSS for the evaluation, and suggests a cautious interpretation of the results, i.e. bearing in mind that the estimation pertains to the sample of secondary-education participants and therefore is generalizable to the corresponding subpopulation. Regarding the interpretation of the main result for example, I found that in MM countries those boys and girls who enrolled in secondary education performed better (by around 13 points on average) in case of being assigned to a male rather than a female teacher.

While the main findings are more likely than previous results to be applicable in MM countries, I note that the context of my study is not quite homogenous. Among eight participating countries, six are from the Arab world (Bahrain, Egypt, Lebanon, Oman, Qatar, and the United Arab Emirates). Conducting a sensitivity analysis, I examined whether the academic benefit of having a male teacher is more an Arab phenomenon[81]. To do so, I re-estimated the coefficients for the reduced sample of Arab countries (Malaysia and Turkey were excluded from the initial sample). According to the results provided by table 4.5, while

[79] According to the World Bank Education Statistics report (2020), the gross secondary-level enrollment ratio in the Arab World was 70.74% in 2014. This number reflects the total enrollment in secondary education regardless of age, as a percentage of the population of official secondary education age. GER can exceed 100% due to the inclusion of over-aged and under-aged students because of early or late school entrance and grade repetition.

[80] See for example Cho (2012).

[81] Due to the low variations in the differenced factors in country-specific models, one cannot address this concern simply by comparing the country-specific estimations.

the directions and statistical significance of the estimated impacts remained unchanged, the size of the impact considerably changed. According to table 4.5, Arab boys and girls perform on average by 16 and 20 points better when assigned to a male teacher (compared with the impacts of about 13 points for boys and girls in the full sample). Therefore, it seems that in the Arab world, girls' and boys' performances are more influenced by the gender of their teachers.

Table 4.5. The FD-Mixed estimator for teacher gender effect using the reduced sample of Arab countries.

	Boys	**Girls**
Female-teacher dummy	-15.56*** (4.52)	-20.20*** (4.80)
Student	18056	17541
Class	893	863
School	701	690
Country	6	6

Note: Own calculations based on TIMSS 2015-G8 for Bahrain, Egypt, Lebanon, Oman, Qatar, and the United Arab Emirates. FD-Mixed estimations for female-teacher dummy coefficients and the standard errors (in parentheses). The number of observations for the FD-Mixed model equals the number of students. The number of cases at each level (student, class, school and country) is reported in the last rows. As recommended by the IEA guidelines (Foy and Yin, 2015), all five plausible values provided by TIMSS were applied for all estimations. Stars represent statistical significance levels. * $p < .10$, ** $p < .05$, and *** $p < .01$.

In closing the discussion, I note the differing size and direction of my estimates compared to the existing estimations. Although previous findings about the effect of teacher gender on student achievement are mixed, to the best of my knowledge, none of the studies found a statistically significant *positive* effect for having a *male teacher* for both genders. Besides the reasons discussed by Cho (2012) to explain her divergent results from those of Dee (2007)[82], I acknowledge the limitations of my data to address all causal concerns. Although the estimation strategy used in this research have ruled out the subject-invariant factors as potential threats to causal links, concerns about gender-based differences in teacher quality still exist. More conspicuously, when cultural norms in a society reinforce stereotypical

[82] Cho (2012) gives three reasons for the dissimilar results of her study to those of Dee (2007), namely differences in the data sources and time periods, and the distinctive underlying assumptions for identifying the effect.

gender roles, it is likely that men and women in teaching profession are not comparably qualified. Differing qualifications would lead to different effectiveness in teaching, and ultimately influence student achievement (Hanushek, 1992; Rockoff, 2004; Rivkin et al., 2005; Aaronson et al., 2007). Furthermore, male and female teachers might systematically differ in terms of unobserved yet relevant variables. For instance, male and female teachers might differ in terms of motivation levels due to differentiated socialization of genders. They might face different opportunities for professional development and teacher training programs. More importantly, I do not know about i.a. teachers' possibly gendered attitude towards educational competencies, whether and how often they communicate such views in classroom environment, etc. All these factors potentially confound with teacher gender impact.

Lastly, the contrasting results is most likely attributable to the culturally disparate context of this study. In MM countries, social interactions between genders are entirely different from those in non-Muslim countries, implying distinctive mechanisms for the teacher gender impact. Unfortunately, with such cross-sectional international data like TIMSS, it is not possible to address the underlying mechanisms. It could be the case that having a male teacher improves the educational outcomes of both girls and boys, but for differing reasons. The dominant mechanism for boys might be the role model effect or male teachers displaying higher academic expectations to them, while girls might put extra effort in studying the subject when taught by a male teacher because they get motivated to disprove the negative stereotypes against females' STEM ability, which are highly reinforced in their society. To sum up, distinctive cultural background could cause the activation of certain potential mechanisms in MM countries, which have not been even mentioned in previous studies focusing on western culture, and therefore, it appears that a comparison between the results in such different contexts is basically irrelevant.

4.8. Chapter Overview and Conclusion

The study in this chapter set out to unfold the causal impact of teacher gender on student achievement in Muslim-majority (MM countries), where, despite the relevance of gender issues due to certain cultural norms, empirical research on gender interactions in classroom is scarce. As the TIMSS 2015 data provides students' test scores in two subjects, I used student fixed-effect strategy to difference out all subject-invariant factors such as the unobserved ability. Estimating a differenced random intercept model, I exploited two features of TIMSS data at the same time, combining the advantages of hierarchical modeling (accounting for between-class variation) with the benefits of student fixed-effect strategy. Findings showed that girls and boys in MM countries generally perform better by around 10% of a standard deviation when assigned to male rather than female teachers. The impact was nearly twofold in single-sex classrooms and varied across individual countries.

Comparing the estimated coefficients obtained from different models in this study provides important insights for the choice of methods in the examination of teacher gender impact. First and foremost, the OLS could produce estimations far off those of the other methods which account in some ways for unobserved variabilities at student or class levels. This relatively large difference most probably stems from the failure of the OLS method to address the nonrandom sorting of students to male and female teachers. Moreover, the estimations obtained from the random intercept model, which accounts for class-level variance component are essentially comparably with estimations of the FD model. If according to Dee (2007) and his following scholars such as Cho (2012), we take for granted that FD-estimator is more precise, it seems that the mixed modeling approach might be able to considerably reduce the bias, even if student fixed-effect are not accounted for. The valuable insight from this finding is that when data limitation does not allow for fixed-effect estimations, say when only one performance per student is available in cross-sectional data, a multilevel approach which adds a random offset at class-level might considerably reduce the selection bias, and is thus advisable.

Moreover, although with so many factors playing a role at country level the interpretation of the individual country results must be made with caution, an important general finding of this

study is that in the context of MM countries, the gender of the teacher matters, particularly in single-sex learning environments. Given that no other attribute of teachers are as much strongly linked to student performance, the moderate-sized impact of teacher gender deserves attention from policymakers and education authorities in MM countries. Although the ratio of male to female teachers in MM countries are relatively more balanced than in other countries, it appears from my results that hiring more male teachers or assigning the existing male teachers to single-sex classrooms boost overall students' performance. More importantly, it seems sensible to design and implement parallel policies for improving the teaching quality of female teachers, namely providing more in-service training opportunities for female teachers in MM countries. Moreover, when closing the achievement gap between male and female students is targeted by policymakers, assigning a teacher of a specific sex to gender-homogenous educational environments could be considered as a profound policy lever to promote educational equality in MM countries.

Last but not least, the present findings provide valuable insights for policymaking in the western countries with large inflows from MM countries, or more specifically, the Middle-East. In designing policies to achieve an integrated society, the policymakers in the destination countries should consider the potentially different impact of gender interactions in classroom on the students who migrated to the country from a distinct cultural background.

Appendix

Table A.4.6. Estimated coefficients of control variables in non-differenced equations (4.1 and 4.4) for the pooled sample.

Independent Variable	Non-differenced models			
	MLR		HLM	
	Boys	Girls	Boys	Girls
Female-teacher dummy	5.45	-5.02*	-10.86***	-13.86***
	(3.56)	(2.84)	(2.79)	(3.34)
Age	-22.93***	-19.53***	-20.94***	-17.71***
	(1.39)	(1.24)	(1.32)	(1.42)
Language Dummy	12.60***	14.87***	10.93***	11.53***
	(2.44)	(2.58)	(1.97)	(1.69)
Home educational resources				
Some resources	36.08***	33.03***	19.10***	16.08***
	(2.28)	(1.99)	(2.37)	(1.95)
Many resources	74.68***	73.98***	37.60***	39.66***
	(4.00)	(3.60)	(3.77)	(3.22)
Subject-dummy	1.77	17.71***	4.81***	18.95***
	(1.53)	(1.21)	(1.58)	(1.53)
Teacher experience				
Between 6 and 10 years	4.30	2.93	4.08	3.92
	(3.68)	(3.20)	(3.46)	(2.86)
Between 11 and 15 years	7.74	9.55**	-1.32	-0.29
	(4.76)	(3.86)	(3.55)	(3.44)
Between 16 and 23 years	-0.34	9.46**	4.32	1.66
	(3.71)	(4.00)	(3.47)	(3.48)
More than 23 years	3.73	12.38***	0.84	1.38
	(4.49)	(4.52)	(3.93)	(4.76)
Teacher education				
Below Bachelor	-4.08	-14.25***	-13.66**	-11.67*
	(8.10)	(5.26)	(6.54)	(7.08)
Beyond bachelor (master or doctorate)	23.03***	13.30***	1.50	1.61
	(6.26)	(3.70)	(3.45)	(3.41)
Non-response	-9.06**	-2.44	-7.15**	-1.99
	(4.47)	(4.28)	(3.28)	(3.87)
Teacher Major				
Majored in teaching but not the subject	-4.08	6.61	-2.34	-7.78*
	(7.18)	(5.61)	(4.55)	(4.24)
Majored in the subject but not teaching	0.30	4.81	-0.73	-5.22
	(6.49)	(4.32)	(3.98)	(4.04)
Majored both in the subject and teaching	-2.68	10.49**	-2.60	-6.50
	(6.37)	(5.20)	(4.09)	(4.04)

Table A.4.6 - continued. Estimated coefficients of control variables in non-differenced equations (4.1 and 4.4) for the pooled sample.

Independent Variable	Non-differenced models			
	MLR		HLM	
	Boys	Girls	Boys	Girls
Single-sex dummy	-38.25***	-18.00***	-55.68***	-20.25***
	(5.99)	(5.23)	(6.77)	(5.93)
Class size	0.95***	0.62**	0.77**	0.48
	(0.28)	(0.25)	(0.34)	(0.30)
School SES				
Almost balanced shares	16.37***	15.84***	15.01**	20.65***
	(4.31)	(3.39)	(6.18)	(4.94)
More affluent	29.00***	21.58***	31.71***	27.14***
	(5.90)	(3.59)	(7.41)	(6.51)
School discipline problem				
Some problems	-11.02***	-10.85***	-13.77**	-15.33***
	(3.85)	(3.57)	(5.39)	(4.25)
Serious problems	-17.06***	-14.17***	-18.78**	-22.03***
	(5.88)	(4.33)	(7.67)	(7.46)
Country dummies	✓	✓	✓	✓
Observations	45982	45749	45982	45749
Student			25040	24986
Class			1409	1389
School			1080	1070
Country			8	8

Note: Own calculations based on IEA Database for TIMSS 2015-G8. The reference groups for categorical covariates, home educational resources, teacher experience, teacher education, teacher major, school SES, and school discipline problems are "few resources", "less than 6 years", "bachelor degree", "no formal post-secondary education or studying in an unrelated field", "more disadvantaged", and "hardly any problem" respectively. Language dummy equals 0 for the students who never or only sometimes speak the language of test at home and 1 for those who often or always do so. Subject dummy equals 1 for science subject and 0 for math. Female-teacher dummy variable equals 1 for female and 0 for male teachers. While the pooled sample included 30,330 boys and 30,029 girls, the number of observations in the exceeds the sample size because for the majority of the students two observations exist, one for math and one for science. For the HLM model, the number of cases at each level (student, class, school and country) is reported in the last rows. As recommended by the IEA guidelines (Foy and Yin, 2015), I used all five plausible values provided by TIMSS for all estimations. Numbers in parentheses show standard errors. Stars represent statistical significance levels. * $p < .10$, ** $p < .05$, and *** $p < .01$.

.

Table A.4.7. Estimated coefficients of control variables in differenced equations (4.2 and 4.5) for the pooled sample.

Differenced Independent Variable	Differenced models (FD-estimators)			
	MLR		HLM	
	Boys	Girls	Boys	Girls
Female-teacher dummy_ Diff	-14.19***	-16.66***	-12.78***	-13.36***
	(2.68)	(3.16)	(3.39)	(3.71)
Subject dummy_ Diff	0.48	0.39	0.33	1.13
	(0.59)	(0.71)	(0.56)	(0.70)
Teacher experience_ Diff				
Less than 6 years_ Diff	2.75	0.00	1.04	-0.09
	(3.09)	(34.55)	(4.68)	(5.72)
Between 6 and 10 years_ Diff	5.84*	4.09	6.47	7.14
	(3.55)	(34.27)	(4.30)	(5.54)
Between 11 and 15 years _ Diff	-0.17	-3.75	1.08	0.40
	(3.57)	(34.34)	(4.14)	(5.50)
Between 16 and 23 years_ Diff	6.26*	1.09	7.06**	4.88
	(3.65)	(34.59)	(3.54)	(5.80)
More than 23 years_ Diff	0.00	-6.97		
	(2.33)	(34.48)		
Teacher education_ Diff				
Below Bachelor_ Diff	-0.49	0.00	-5.80	-5.47
	(24.30)	(18.92)	(8.98)	(9.44)
Bachelor degree_ Diff	5.51	5.96	6.76	6.53
	(23.39)	(18.13)	(4.89)	(4.89)
Beyond bachelor (master or doctorate) _ Diff	7.19	7.89	7.87	8.75
	(22.98)	(18.17)	(6.58)	(5.80)
Non-response_ Diff	0.00	1.88		
	(23.11)	(17.68)		
Teacher Major_ Diff				
No formal post-secondary or unrelated field_ Diff	4.26	0.00	6.35	12.85***
	(17.48)	(81.90)	(4.83)	(4.95)
Majored in teaching but not the subject_ Diff	0.00	-11.04	2.37	2.41
	(17.41)	(81.64)	(3.07)	(3.36)
Majored in the subject but not teaching_ Diff	1.76	-9.67	-0.69	1.32
	(17.08)	(81.50)	(2.41)	(2.56)
Majored both in the subject and teaching_ Diff	0.16	-10.53		
	(17.03)	(81.58)		
Observations	24780	24365	24780	24365
Student			24780	24365
Class			1354	1320
School			1088	1071
Country			8	8

Table A.4.7 - continued. Estimated coefficients of control variables in differenced equations (4.2 and 4.5) for the pooled sample.

Note: Own calculations based on IEA Database for TIMSS 2015-G8. The estimated coefficients and standard errors (in parentheses) in this table relate to the differenced variables in equations 4.2 and 4.5. Therefore, for each individual student, the differenced variable "variable-name_Diff" was calculated as the value for math minus the value for science, which equals zero for non-varying factors between the subjects. While the pooled sample included 30,330 boys and 30,029 girls, the number of students for differenced models is confined to those whose performances have been observed for both math and science subjects (27,183 boys and 26,900 girls). The number of observations in the MLR (one-level) and HLM (multilevel) differenced models are smaller than the sample size due to the missing information for different control variables for around 2500 boys and almost 2500 girls across the eight countries. As recommended by the IEA guidelines (Foy and Yin, 2015), I used all five plausible values provided by TIMSS for all estimations. Stars represent statistical significance levels. * $p < .10$, ** $p < .05$, and *** $p < .01$.

Bibliography

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and Student Achievement in the Chicago Public High Schools. *Journal of Labor Economics, 25(1)*, 95-135.

Acker, S. (1995). Chapter 3: Gender and Teachers' Work. *Review of Research in Education, 21(1)*, 99–162. DOI:10.3102/0091732X021001099.

Adkinson, J. E. (2008). Does cooperative learning affect girls' and boys' learning and attitudes toward mathematic transformation skills in single-sex and mixed-sex classrooms? Dissertation Abstracts International: Section A. *Humanities and Social Sciences*, *68(11)*, 4639.

Almquist, E. M. & Angrist, S. S. (1971). Role Model Influences on College Women's Career Aspirations. *Merrill-Palmer quarterly, 17*, 236-279.

Ammermueller, A., & Dolton, P. (2006). Pupil-Teacher Gender Interaction Effects on Scholastic Outcomes in England and the USA. SSRN Electronic Journal. *Advance online publication*. DOI:10.2139/ssrn.927689.

Angrist, J. & Lavy, V. (1999). Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement, *The Quarterly Journal of Economics, 114(2),* 533-575. Oxford University Press.

Antecol, H., Eren, O., & Ozbeklik, S. (2015). The Effect of Teacher Gender on Student Achievement in Primary School. *Journal of Labor Economics, 33(1),* 63–89. DOI:10.1086/677391.

Aslam, M. & Kingdon, G. (2007). What Can Teachers Do to Raise Pupil Achievement? *Economics of Education Review, 30(3)*, 559-574. DOI:10.1016/j.econedurev.2011.01.001.

Baker, D. P., Riordan, C., & Schaub, M. (1995). The effects of sex-grouped schooling on achievement: the role of national context. *Comparative Education Review, 39*, 468-482.

Banu, D. P. (1986). Secondary school students' attitudes towards science. *Research in*

Science & Technological Education, 4, 195–202.

Barone, C. (2011). Some Things Never Change: Gender Segregation in Higher Education across Eight. *American Sociological Association, Sociology of Education, 84(157)*. DOI:10.1177/0038040711402099.

Barro, R. & J. Lee. (1994). Sources of Economic Growth. *Carnegie-Rochester Conference Series on Public Policy, 40*, 1-46.

Barro, R. (1991). Economic Growth in a Cross- Section of Countries, *Quarterly Journal of Economics, 106,* 407-443.

Basow S. A. & Howe, K. G. (1980). Role-model Influence: Effect of Sex and Sex-role Attitude in College Students. *Psychology of Women Quarterly, 4,* 558-572.

Beilock, S. L., Gunderson, E. A., Ramirez, G., & Levine, S. C. (2010). Female teachers' math anxiety affects girls' math achievement. *Proceedings of the National Academy of Sciences, 107(5),* 1060–1063.

Bettinger, E. P., & Long, B. T. (2005). Do Faculty Serve as Role Models? The Impact of Instructor Gender on Female Students. *American Economic Review, 95(2),* 152–157. DOI:10.1257/000282805774670149.

Bigler, R. S., & Liben, L.S. (2006). A Developmental Intergroup Theory of Social Stereotypes and Prejudice. In R. V. Kail (Ed.), *Advances in Child Development and Behavior*, 34, 39-89.

Bigler, R. S., Hayes, A. R., & Liben, L.S. (2014). The Role of Gender in Educational Contexts and Outcomes. *Advances in Child Development and Behavior*, 47.

Billger, S. M. (2009). On Reconstructing School Segregation: The Efficacy and Equity of Single-Sex Schooling. *Economics of Education Review, 38(3),* 393-402.

Booth, A., Cardona-Sosa, L., & Nolen, P. (2018). Do Single-Sex Classes Affect Academic Achievement? An Experiment in a Coeducational University. *Journal of Public Economics, 168,* 109-126.

Bos, K. & Kuiper W. (1999). Modelling TIMSS Data in European Comparative Perspective: Exploring Influencing Factors on Achievement in Mathematics in Grade 8. *Educational Research and Evaluation, 5(2),* 157-179.

Bottomley, C., Kirby, M., Lindsay, S. W., & Alexander, N. (2016). Can the Buck Always Be Passed to the Highest Level of Clustering? *BMC Medical Research Methodology, 16 (29),* PMCID: PMC4784323. doi: 10.1186/s12874-016-0127-1.

Bracey, G. W. (2006). Separate but superior? A review of issues and data bearing on single-sex education. Tempe, AZ: *Educational Policy Research Unit.*

Brathwaite, D. A. (2010). A comparative analysis of single-sex schools in terms of achievement in reading and math and student attendance. Dissertation Abstracts International: Section A. *Humanities and Social Sciences, 71(4)*, 1148.

Campbell, K. T., & Evans, C. (1997). Gender issues in the classroom: A comparison of mathematics anxiety. *Education, 117*, 332–338.

Caponera, E., & Losito, B. (2016). Context factors and student achievement in the IEA studies: evidence from TIMSS. *Large-Scale Assessments in Education, 4(1).* DOI:10.1186/s40536-016-0030-6.

Carrington, B., Tymms, P., & Merrell, C. (2008). Role models, school improvement and the gender gap- Do men bring out the best in boys and women the best in girls? *British Educational Research Journal, 34(3),* 315–327. DOI:10.1080/01411920701532202.

Carstens, R. & Hastedt, D. (2010). The Effect of Not Using Plausible Values When They Should Be: An illustration using TIMSS 2007 grade-8 mathematics data. *IEA Data Processing and Research Center.*

Chadwell, D. (2010a). A Gendered Choice: Designing and Implementing Single-sex Programs and Schools. *Thousand Oaks*, CA: Corwin.

Cho, I. (2012). The effect of teacher–student gender matching: Evidence from OECD countries. *Economics of Education Review, 31(3),* 54–67. DOI:10.1016/j.econedurev.2012.02.002.

Chudgar, A., & Sankar, V. (2008). The relationship between teacher gender and student achievement: evidence from five Indian states. *Compare: A Journal of Comparative and International Education, 38(5),* 627–642. DOI:10.1080/03057920802351465.

Claessens, A., Duncan, G. J., & Engel, M. (2009). Kindergarten Skills and Fifth-grade Achievement: Evidence from the ECLS-K. *Economics of Education Review28 (4),* 415-427.

Coleman, J. (1961). The Adolescent Society. *New York: Free Press*.

Cordero, J. M., Cristobal V., & Santin, D. (2017). Causal Inference on Education Policies: a Survey of Empirical Studies Using PISA, TIMSS, and PIRLS. *Munich Personal RePEc Archive Paper, No. 76295*.

Corso, J. F. (1959). Age and Sex Differences in Pure-Tone Thresholds. *The Journal of the Acoustical Society of America, 31*, 498.

Dahl, G. B., Felfe, C., Frijters, P., & Rainer, H. (2020). Caught Between Cultures: Unintended Consequences of Improving Opportunity for Immigrant Girls. *National Bureau of Economic Research Working Paper No. 26674*.

Dee, T. S. (2005). A Teacher Like Me: Does Race, Ethnicity, or Gender Matter? *American Economic Review, 95(2),* 158–165. DOI:10.1257/000282805774670446.

Dee, T. S. (2007). Teachers and Gender Gaps in Student Achievement. *Journal of Human Resources, 42(3)*. 528-554.

Doris, A., O'Neill, D., & Sweetman, O. (2013). Gender, single-sex schooling and math achievement. *Economics of Education Review, 35*, 104–119.

Duflo, E. (2012). Women's Empowerment and Economic Development. *Journal of Economic Literature, 50(4),* 1051-1079.

Dyer, G., & Tiggemann, M. (1996). The Effect of School Environment on Body Concerns in Adolescent Women. *Sex Roles, 34,* 127–138.

Edwards, S. R. (2002). Gender-based and mixed-sex classrooms: The relationship of mathematics anxiety, achievement, and classroom performance in female high school math students. Dissertation Abstracts International: Section A. *Humanities and Social Sciences, 62(8),* 2639.

Egalite, A. J., & Kisida, B. (2018). The Effects of Teacher Match on Students' Academic Perceptions and Attitudes. *Educational Evaluation and Policy Analysis, 40(1)*, 59–81. DOI:10.3102/0162373717714056.

Egalite, A. J., Kisida, B., & Winters, M. A. (2015). Representation in the classroom: The effect of own-race teachers on student achievement. *Economics of Education Review, 45*, 44–52. DOI:10.1016/j.econedurev.2015.01.007.

Egbochuku, E. O., & Aihie, N. O. (2009). Peer group counselling and school influence on adolescents' self-concept. *Journal of Instructional Psychology, 36*, 3–12.

Ehrenberg, R. Goldhaber, D., & Brewer, D. (1995). Do Teacher's Race, Gender and Ethnicity Matter? Evidence from the National Educational Longitudinal Study of 1988. *Industrial and Labor Relations Review. 48(3).* 547-561.

Eisenkopf, G., Hessami, Z., Fischbacher, U. (2015). Academic performance and single-sex schooling: Evidence from a natural experiment in Switzerland. *Journal of Economic Behavior & Organization, 115*, 123-143.

Epple, D. & Romano, R. E. (2011). Peer Effects in Education: A Survey of the Theory and Evidence. *Handbook of Social Economics, (1B), Ch. 20*. DOI:10.1016/S0169-7218(11)01025-2.

Epstein, C. F. (1997). The Myths and Justifications of Sex Segregation in Higher Education: VMI and the Citadel. 4Duke Journal of *Gender Law & Policy,* 101-118.

Eriksson, K., Helenius, O., & Ryve, A. (2019). Using TIMSS items to evaluate the effectiveness of different instructional practices. *Instructional Science, 47(1),* 1–18. https://doi.org/10.1007/s11251-018-9473-1

Esfandiari, M., & Jahromi, S. (1989). A comparison of Iranian high school students in single-sex and mixed-sex bilingual schools: Intelligence and vocational aspiration. *International Journal of Intercultural Relations, 13,* 447–464.

Foy, P. (2017). TIMSS 2015 User Guide for the International Database. *TIMSS & PIRLS International Study Center,* Lynch School of Education, Boston college and International Association for the Evaluation of Educational Achievement (IEA). ISBN: 978-1-889938-38-7.

Foy, P., & Yin, L. (2015). Data Analyses with IEA's TIMSS and PIRLS International Databases. *Contemporary Educational Research Quarterly, 23 (4),* 41-61. DOI:10.6151/CERQ.2015.2304.02.

Geesa, R. L., Izci, B., Song, H., & Chen, S. (2019). Exploring Factors of Home Resources and Attitudes Towards Mathematics in Mathematics Achievement in South Korea, Turkey, and the United States. *EURASIA Journal of Mathematics, Science and Technology Education, 15(9).*

Gershenson, S., Holt, S. B., & Papageorge, N. W. (2016). Who believes in me? The effect of student–teacher demographic match on teacher expectations. *Economics of Education Review, 52*, 209–224. DOI:10.1016/j.econedurev.2016.03.002.

Gordon, D. M., Iwamoto, D. K., Ward, N., Potts, R., & Boyd, E. (2009). Mentoring urban Black middle school male students: Implications for academic achievement. *Journal of Negro Education, 78*, 277–289.

Guiso, L., Monte, F. Zingales, L., & Sapienza, P. (2008). Culture, Gender, and Math. *Science, Education Forum, 320*, 1164-1165.

Hakura, D., M. Hussain, M. Newiak, V. Thakoor, and F. Yang, (2016) Inequality, Gender Gaps and Economic Growth. *International Monitory Fund Working Paper No. 16/111*.

Halpern, D., F., Eliot, L. Bigler, R. S., Fabes, R. A., Hanish, L., D., Hyde, J., Martin, C. L. (2011). The Pseudoscience of Single-sex Schooling. *Science, 333*, 1706-1707.

Hannan, D. F., Smyth, J. Mccullagh, R. O'Leary, & Mcmahon, D. (1996). Coeducation and Gender Equality. *Oak Tree Press/ ESRI*. Dublin.

Hanushek, E. A. & Rivkin, S. G. (2010). Constrained job matching: Does teacher job search harm disadvantaged urban schools? *National Bureau of Economic Research .NBER Working Paper 15816*.

Hanushek, E. A. & Woessmann, L. (2010). The Cost of Low Educational Achievement in the European Union. *European Commission Education and Culture*. European Expert Network on Economics of Education (EENEE) Analytical Report No. 7.

Hanushek, E. A. & Woessmann, L. (2011). The Economics of International Differences in Educational Achievement, *Handbook of the Economics of Education, 3*. San Diego, CA: Elsevier.

Hanushek, E. A. & Woessmann, L. (2017). School Resources and Student Achievement: A Review of Cross-Country Economic Research. *Springer International Publishing,*

Methodology of Educational Measurement and Assessment, Ch. 8. DOI 10.1007/978-3-319-43473-5_8.

Hanushek, E. A. (1992). The Trade-off between Child Quantity and Quality. *Journal of Political Economy, 100(1),* 84-117.

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review, 30(3),* 466–479. DOI:10.1016/j.econedurev.2010.12.006.

Hanushek, E. A. (2013). Economic Growth in Developing Countries: The role of human capital. *Economics of Education Review 37*, 204-212.

Hanushek, E. A. (2020). The Economics of Education, A Comprehensive Overview (Second edition), *Academic Press*, ISBN: 978-0-12-815391-8. Retrieved from *https://doi.org/10.1016/C2017-0-02304-2*.

Hanushek, E. A. Peterson, P. E., Shakeel, M. D., Talpey, L. M., Woessmann, L. (2019). The Unwavering SES Achievement Gap: Trends in U.S. Student Performance. *NBER Working Paper No. 25648.*

Hanushek, E. A., & Kimko, D. D. (2000). Schooling, labor force quality, and the growth of nations. *The American Economic Review, 90(5),* 1184-1208.

Hanushek, E. A., & Woessmann, L. (2008). The role of cognitive skills in economic development. *Journal of Economic Literature, 46(3),* 607-668.

Hanushek, E. A., & Woessmann, L. (2010). The Economics of International Differences in Educational Achievement. *IZA Discussion Paper No. 4925*.

Hanushek, E. A., & Woessmann, L. (2012). Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *Journal of Economic Growth, 17(4),* 267-321.

Hanushek, E. A., & Woessmann, L. (2015). The knowledge capital of nations: Education and the economics of growth. *Cambridge,* MA: MIT Press.

Hanushek, E. A., Peterson, P. E. & Woessmann, L. (2014). U.S. Students from Educated Families Lag in International Tests. *Education Next 14(4),* 8–18.

Hayes, R. A., Pahlke, E. & Bigler, R. (2011). The efficacy of single-sex education: Testing for selection and peer quality effects. *Sex Roles: A Journal of Research,* 65, 693-703.

Hoffman, B. H., Badgett, B. A., & Parker, R. P. (2008). The effect of single-sex instruction in a large, urban, at-risk high school. *Journal of Educational Research, 102,* 15–36.

Holmlund, H., & Sund, K. (2008). Is the gender gap in school performance affected by the sex of the teacher? *Labour Economics, 15(1),* 37–53. DOI:10.1016/j.labeco.2006.12.002.

Hoxby, C. (2000). Peer Effects in the Classroom: Learning from gender and race variation. *NBER Working Paper No. 7867*.

Huguet, P. & Régner, I. (2007). Stereotype threat among school girls in quasi-ordinary classroom circumstances, *Journal of Educational Psychology,* 99, 545-560.

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist, 60,* 581–592.

Institute for Research and Planning in Higher Education, Ministry of Science, Research, and Technology, (2017), Retrieved from https://irphe.ac.ir/content/1921/-لطال-عات-کت-اب-آمار-در-سال-۱۳۹۵-۹۶ In Persian.

Iranian Association for Scientific Development. (2011). Report on the Conference for Gender Separation at Universities. Retrieved from http://www.iasd.ir/find.php?item=1.61.18.fa. In Persian.

Jackson, C. & Smith, D. (2000). Poles Apart? An Exploration of Single-Sex and Mixed-Sex Educational Environments in Australia and England. *Educational Studies, 26,* 409-422.

Jackson, C. K. (2012). Single-sex Schools, Student Achievement, and Course Selection: Evidence from rule-based student assignments in Trinidad and Tobago. *Journal of Public Economics, 96*, 173–187.

Kessels, U. & Hannover, B. (2008). When Being a Girl Matters Less: Accessibility of Gender-related Self-knowledge in Single-sex and Co-educational Classes and Its Impact on Students' Physics-related Self-concept of Ability. *British Journal of Educational Psychology,* 78, 273-289.

Klasen, S. (2002). Low Schooling for Girls, slower Growth for All? *World Bank Economic Review 16,* 345-373 (2002).

Klasen, S. and F. Lamanna. (2009). The impact of gender inequality in education and employment on economic growth: New evidence for a panel of countries. *Feminist Economics, 15*, 91-132.

Klein, J. (2004). Who is most responsible for gender differences in scholastic achievements: pupils or teachers? *Educational Research, 46(2),* 183–193. DOI:10.1080/0013188042000222458.

Kramarz, F., Machin, S. Quazad, A. (2008). What Makes a Test Score: The respective contributions of pupils, schools, and peers in achievement in English primary education. *INSEAD Working Paper No. 2008/58/EPS.*

Laster, C. (2004). Why we must try same-sex instruction. *Education Digest, 70*, 59–62.

Lavy, V. (2008). Do Gender Stereotypes Reduce Girls' or Boys' Human Capital Outcomes? Evidence from a natural experiment. *Journal of Public Economics, 92,* 2083-2105.

Lavy, V., & Schlosser, A. (2011). Mechanisms and Impacts of Gender Peer Effects at School. *American Economic Journal: Applied Economics*, *3*(2), 1–33.

Lee, D. & Huh, Y. (2014). What TIMSS Tells Us about Instructional Practice in K-12 Mathematics Education, *Contemporary Educational Technology, 5(4),* 286-301.

Lee, V. E., & Bryk, A. S. (1986). Effects of single-sex secondary schools on student achievement and attitudes. *Journal of Educational Psychology, 78*, 381–395.

Lee, V. E., & Lockheed, M. E. (1990). The effects of single-sex schooling on achievement and attitudes in Nigeria. *Comparative Education Review, 34*, 209–231.

Lenroot, K. R., Gogtay, N., Greenstein, D. K., Molloy Wells, E., Wallace, G. L., Clasen, L. S., Blumenthal, J. D., Lerch, J., Zijdenbos, A. P., Evans, A. C., Thompson, P. M., & Giedd, J. N. (2007). Sexual dimorphism of brain developmental trajectories during childhood and adolescence*. NeuroImage. 36.* 1065-1073.

Liberman, M., (2008), Retrieved from "Language Log": *http://languagelog.ldc.upenn.edu/nll/?p=171*.

Lim, J., & Meer, J. (2015). The Impact of Teacher-Student Gender Matches: Random Assignment Evidence from South Korea. Cambridge, MA. DOI: 10.3386/w21407.

Lubinski, D., Benbow, C. P., & Kell, H. J. (2014). Life Path and Accomplishments of Mathematically Precocious Males and Females Four Decades Later. *Psychological Science, 25,* 2217-2232.

Mael, F., Alonso, A., Gibson, D., Rogers, K., & Smith, M. (2005). Single-sex versus

coeducational schooling: A systematic review. *American Institutes for Research.* Washington, DC.

Maher, A. H. (2012). Construct validity of self-concept in TIMSS's student background questionnaire: a test of separation and conflation of cognitive and affective dimensions of self-concept among Saudi eighth graders. *European Journal of Psychology of Education.* DOI: 10.1007/s10212-012-0162-1.

Maher, A. H. (2019).Relations among Engagement, Self-Efficacy, and Anxiety in Mathematics among Omani Students, *Electronic Journal of Research in Education Psychology*. DOI: 10.25115/ejrep.v17i48.2182.

Maher, M. AH., Al-Malki, H. (2014). Frame of Reference and Achievement across Gender among Omani Middle-school Students, *The International Journal of Educational and Psychological Assessment, 16(1),* 82-101.

Mallam, W. A. (1993). Impact of school-type and sex of the teacher on female students' attitudes toward mathematics in Nigerian secondary schools. *Educational Studies in Mathematics, 24,* 223–229.

Marsh, H. W., Martin, A. J., & Cheng, J. H. S. (2008). A multilevel perspective on gender in classroom motivation and climate: Potential benefits of male teachers for boys? *Journal of Educational Psychology, 100(1),* 78–95. DOI:10.1037/0022-0663.100.1.78.

Martin, M. O., Foy, P., Mullis, I. V. S., & O'Dwyer, L. M. (2013). Effective schools in reading, mathematics, and science at the fourth grade. *TIMSS and PIRLS 2011: Relationships among reading, mathematics, and science achievement at the fourth gradeimplications for early learning,* 109-178. Chestnut Hill: TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., Mullis, I. V. S., & Hooper, M. (2016). Methods and Procedures in TIMSS 2015. *International Study Center, Lynch School of Education,* Boston College.

Mehran, G. (2003). Gender and Education in Iran. *Paper commissioned for the EFA Global Monitoring Report 2003/4, The Leap to Equality.* Retrieved from *http://datatopics.worldbank.org/hnp/files/edstats/IRNgmrpro03.pdf*.

Morse, S. (1998). Separated by sex: A critical look at single-sex education for girls. *American*

Association of University Women Educational Foundation. Washington, DC.

Neugebauer, M., Helbig, M., & Landmann, A. (2011). Unmasking of the Same-sex Teacher Advantage, *European Sociological review, 27(5)*, 669-689.

Novotney, A. (2011). Coed vs. Single-sex Ed: Does Separating Boys and Girls Improve Their Education? *American Psychological Association, 42(2),* 58.

Okoro, C. C., Ekanem, I. E., & Udoh, N. A. (2012). Teacher Gender and The Academic Performance of Children in Primary Schools in Uyo Metropolis, Akwa Ibom State, Nigeria. *Journal of Educational and Social research, 2(1*). 267-273. DOI:10.5901/jesr.2012.02.01.267.

Oosterbeek, H. & van Ewijk, R. (2014). Gender Peer Effect in University; Evidence from a Randomized Experiment. *Economics of Education Review, 38,* 51-63.

Pahlke, E., & Hyde, J. S. (2016). The Debate Over Single-sex Schooling. *Child Development Perspectives, 10*(2), 81-86.

Pahlke, E., Hyde, J. S., & Allison, C. M. (2014). The Effects of Single-Sex Compared with Coeducational Schooling on Students' Performance and Attitudes: A Meta-Analysis. *Psychological Bulletin. 140*(4).

Paredes, V. (2014). A teacher like me or a student like me? Role model versus teacher bias effect. *Economics of Education Review, 39*, 38–49. DOI:10.1016/j.econedurev.2013.12.001.

Park, H., Behrman, J. R. & Choi, J. (2018). Do single-sex schools enhance students' STEM (science, technology, engineering, and mathematics) outcomes? *Economics of Education Review, 62*(C), 35-47.

Park, H., Behrman, J. R., & Choi, J. (2013). Causal Effects of Single-Sex Schools on College Entrance Exams and College Attendance: Random Assignment in Seoul High Schools. *Demography, 50(2).* 447-469. doi: 10.1007/s13524-012-0157-1.

Rabe-Hesketh, S., and A. Skrondal. 2006. Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society*, Series A 169, 805-827.

Rawal, S., & Kingdon, G. (2010). Akin to my teacher: Does Caste, Religious, or Gender Distance between Student and Teacher Matters? Some evidence from India. *Leading*

Education and Social Research, IOE London, DoQSS Working Paper, No. 10-18.

Raznahan, A., Lee, Y., Stidd, R. Long, R., Greenstein, D., Clasen, L., Addington, A., Gogtay, N., Rapport, J. L., & Giedd, J. N. (2010). Longitudinally mapping the influence of sex and androgen signaling on the dynamics of human cortical maturation in adolescence. *Proceedings of the National Academy of Science (PNAS), 107(39),* 16988-16993.

Riegle-Crumb, C., & Humphries, M. (2012), Exploring Bias in Math Teachers' Perception of Students' Ability by Gender and Race/ethnicity. *Gender and Society*, 26, 290-322. DOI:10.1177/0891243211434614.

Riordan, C. (1985). Public and Catholic Schooling: The Effects of Gender Context Policy. *American Journal of Education, 93*(4), 518-540.

Riordan, C. (1994). Single-gender schools: Outcomes for African and Hispanic Americans. *Research in Sociology of Education and Socialization, 10*, 177–205.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73(2),* 417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review 94(2),* 247–252.

Rosenthal, R. & Jacobson, L. (1968). Pygmalion in the Classroom. *Urban Review, 3*, 16-20.

Roth, D. J. (2009). The effectiveness of single-gender eight-grade English, history, mathematics, and science classes. *ProQuest Dissertations and Theses database. UMI No.3429036.*

Rustad, N. & Woods, J. (2004). Statement on the legality of single-sex education. *American Association of University Women (AAUW).* Washington DC.

Sacerdote, B. (2011). Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far? *Handbook of the Economics of Education, 3,* 249–277. DOI: 10.1016/S0169-7218(11)03004-8.

Sadker, M., Sadker, D., & Zittleman, K. R. (2009). Still failing at fairness: How gender bias cheats girls and boys in school and what we can do about it. New York, NY: Simon & Schuster.

Salikutluk, Z., & Heyne, S. (2017). Do Gender Roles and Norms Affect Performance in Maths? The Impact of Adolescents' and their Peers' Gender Conceptions on Maths Grades. *European Sociological Review*, *33*(3), 368–381. DOI:10.1093/esr/jcx049.

Salomone, R. C. (2006). Single-sex programs: Resolving the research conundrum. *Teachers College Record, 108,* 778–802.

Sansone, D. (2017). Why does teacher gender matter? *Economics of Education Review, 61*, 9–18. DOI:10.1016/j.econedurev.2017.09.004.

Santos, C. E., Galligan, K. M., Pahlke, E. & Fabes, R. A. (2013). Gender stereotyping, boys' achievement and adjustment during junior high school. *American Journal of Orthopsychiatry, 83,* 252-264.

Sax, L. (2005). Why Gender Matters. New York, Doubleday.

Sax, L. J., Arms, E., Woodruff, M., Riggers, T., & Eagan, K. (2009). Women graduates of single-sex and coeducational high schools: Differences in their characteristics and the transition to college. *Sudikoff Family Institute for Education and New Media.* Los Angeles, CA.

Schneeweis, N., & Zweimueller, M. (2012). Girls, girls, girls: Gender Composition and female school choice. *Economics of Education Review, 31*, 482-500.

Schwerdt, G., & Wuppermann.C. A. (2011). Is Traditional Teaching Really All That Bad? A Within-Student Between-Subject Approach. *Economics of Education Review 30(2),* 365–79.

She, H.-C. (2000). The interplay of a biology teacher's beliefs, teaching practices and gender-based student-teacher classroom interaction. *Educational Research, 42(1),* 100–111. DOI:10.1080/001318800363953.

Shirazi, F. (2014). Educating Iranian Women. *International Journal of Education and Social Science, 1(2).*

Smithers, A., Robinson, P. (2006). The Paradox of Single-sex and Coeducational Schooling. Buckingham. UK Carmichael Press.

Smyth, E. (2010). Single-sex education: What does research tell us? *Revue française de pédagogie. 171.*

Snijders, T. A. B. & Bosker, R. J. (2012). Multilevel Analysis, 2nd ed. Sage Publications.

Spielhofer, T., Benton, T., & Schagen, S. (2004). A study of the effects of school size and single- sex education in English schools. *Research Papers in Education, 19,* 133–159.

Statistical Center of Iran. (2019). available from https://www.amar.org.ir/-ارتباط-با-ما-آمارهای-کشوری In Persian.

Steele, C. M. & Aronson, J. (1995). Stereotype Threat and the Intellectual Test Performance of African Americans. *Journal of Personality and Social Psychology, 69*, 797-811.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52,* 613–629.

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. *Advances in experimental social psychology, 34,* 379-440.

Stephens, M. (2009). Effects of single-sex classrooms on student achievement in math and reading. Dissertation Abstracts International: Section A. *Humanities and Social Sciences, 70(6)*, 1934.

Stotsky, S., Denny, G., & Tschepikow, N. (2010). Single-sex classes in two Arkansas elementary schools: 2008–2009. Fayetteville: University of Arkansas.

Sullivan, A., Joshi, H., & Leonard, D. (2010). Single-sex schooling and academic attainment at school and throughout lifecourse. *American Education Research Journal,* 47(1), 6-36.

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. *The social psychology of intergroup relations,* 33–47.

The Supreme Council for Cultural Revolution. (2011). Islamisation of Universities and Higher Educational Centers. Retrieved from *http://sccr.ir/pages/?current=provlist*.

Turner, J. C., Ellemers, N., Spears, R., Doosje, B. (1999). Some current issues in research on social identity and self-categorization theories. *Social identity, Oxford: Blackwell*, 6–34.

Turnovsky, S. J. (2015). Economic Growth and Inequality: The Role of Public Investment. *Journal of Economic Dynamics & Control, 61,* 204-221.

U.S. Department of Education. (2004). Nondiscrimination on the Basis of Sex in Education

Programs or Activities Receiving Federal Financial Assistance. Federal Register, 69(46), 11276011285.

UNESCO. (2006). Advocacy Brief: The Impact of Women Teachers on Girls' Education. Bangkok: UNESCO.

van Ewijk, R., & Sleegers, P. J. C. (2010). The effect of peer socioeconomic status on student achievement: A meta-analysis. Educational research review, 5, 134-150.

Vrooman, M. K. (2010). An examination of the effects of single-gender classes on reading and mathematics achievement test scores of middle school students. Dissertation Abstracts International: Section A. *Humanities and Social Sciences, 70(8),* 2880.

Vuorinen-Lampila, P. (2016). Gender segregation in the employment of higher education graduates. *Journal of Education and Work, 29 (3*), 284-308. DOI:10.1080/13639080.2014.934788

Whitmore, D. (2005). Resource and Peer Impacts on Girls' Academic Achievement: Evidence from a randomized experiment. *American Economic Review. Papers and Proceedings 95 (2),* 199-203.

Wilson, K. L., & Boldizar, J. P. (1990). Gender Segregation in Higher Education: Effects of Aspirations, Mathematics Achievement, and Income. *American Sociological Association, Sociology of Education 63 (1),* 62-74. DOI: 10.2307/2112897.

Winters, M. A., Haight, R. C., Swaim, T. T., & Pickering, K. A. (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,* 69–75. DOI:10.1016/j.econedurev.2013.01.007.

Woessmann, L. & West, M. R. (2006). Class-Size Effects in School Systems around the World: Evidence from Between-Grade Variation in TIMSS. *European Economic Review 50(3*), 695–736.

Woessmann, L. (2005). Educational Production in Europe. *Economic Policy 20(43),* 446–504.

Woessmann, L. (2016). The Importance of School Systems: Evidence from International Differences in Student Achievement. *Journal of Economic Perspectives, 30(3),* 3–32.

Wooldridge, J. M. (2002). Econometric analysis of cross section and panel data. Cambridge, Mass: MIT Press.

Wooldridge, J. M. (2012). Introductory Econometrics: a modern approach. Mason, Ohio: South-Western Cengage Learning.

World Bank Education Statistics. (March 2020 update). Retrieved from *https://datacatalog.worldbank.org/dataset/education-statistics*.

World Bank. (2014). Voice and Agency. Empowering women and girls for shared prosperity. Washington DC: The World Bank.

World Economic Forum. (2019). Global Gender Index Report 2020, ISBN-13: 978-2-940631-03-2.

World Health Organization. (2021). Gender and Health. Retrieved from *https://www.who.int/health-topics/gender#tab=tab_1*.

Yang, Y. (2003). Dimensions of Socio-economic Status and their Relationship to Mathematics and Science Achievement at Individual and Collective Levels. *Scandinavian Journal of Educational Research, 47(1).* DOI:10.1080/0031383032000033317.

Zeeuw, E. L. de, van Beijsterveldt, C. E.M., Glasner, T. J., Bartels, M., Geus, E. J.C. de, & Boomsma, D. I. (2014). Do children perform and behave better at school when taught by same-gender teachers? *Learning and Individual Differences, 36,* 152–156. DOI:10.1016/j.lindif.2014.10.017.