LUND UNIVERSITY
School of Economics and Management

Master's Programme in Economics

The effect of same-sex teacher assignment on student outcomes: Evidence from Australia and New Zealand

by

Hugo Morgado Azevedo

NEKN01
Master's thesis (15 credits)
August 2022
Supervisor: Ana Rodriguez-Gonzalez

Abstract

This essay investigates the effects of a same-sex teacher assignment on female and male 8th grade students in Australia and New Zealand using data from the Trends in Mathematics and Science Study (TIMSS). I employ the first difference (FD) method introduced by Dee (2007) to analyze the impacts on students' math and science academic achievements, as well as their attitudes toward these subjects. TIMSS has the appealing aspect of providing observations of students attending both subjects and their teachers in the corresponding disciplines, allowing to account for unobservable subject-invariant student characteristics.

When restricting the assessment to the knowing cognitive domain part, I find that girls' test scores increase significantly by around 0.040 SD when assigned to a female teacher and that this effect is driven by advantaged girls. This effect is smaller when using the standard test scores, but these outcomes seem more likely to cause biased estimates due to spillovers. The gender match effects I find on students' subject perceptions are larger and mostly positive. The comparison with previous estimates shows that the gender match effects on academic performance are not primarily caused by a change in students' attitudes toward the disciplines.

Overall, my findings show that boys do not experience any strong and significant positive effects of gender match on academic achievement and that the effects on students' subject perceptions are also observed predominantly on girls. Therefore, my results provide little support for policies that are based on this argument to encourage the recruitment of male teachers to attenuate the growing feminization of the teaching profession.

Keywords: Education, Gender match, Fixed effects, First difference, TIMSS

Table of Content

1 Introduction

The preponderant role of the teacher on students' academic success has been a large focus of attention in economics of education. For instance, Hanushek (1971) questions whether teachers count in the achievement of students and which of their characteristics are of relevance. Aiming to improve the efficiency of the educational system, he was among the firsts providing experimental evidence of a teacher's effect on student achievement. However, further research evaluating the significance of different teacher determinants is needed, as the exact characteristics that matter remain unclear.[1]

These past decades, decision-makers have implemented policies to encourage the recruitment of male teachers due to the feminization of the teaching profession. This was intended to limit the potential negative consequences on boys caused by a lack of male role models. For example, the state of Queensland in Australia developed a plan to increase the rate of male teachers in 2002.[2]

Such initiatives have led researchers to focus on the effects of gender interactions between students and teachers. For instance, Dee (2007) finds a positive effects of student-teacher gender match in his paper, supporting the importance of a same-sex teacher on student performance. Even so, it remains elusive to what extent the assignment to a same-gender teacher would influence students' achievements as the literature provides mixed results.

This paper explores the effects of teacher gender match on female and male 8th grade students in Australia and New Zealand. I use the first difference (FD) method introduced by Dee (2007) to investigate the impact on students' math and science test scores but also their attitudes towards the subjects. The Trends in Mathematics and Science Study (TIMSS) provides observations of students following both subjects and their teachers in the respective disciplines. This enables me to remove student subject-invariant characteristics that could be a large source of bias.

I first find small and insignificant effects of same-sex teacher assignment on students' academic outcomes even when allowing for subject specific effects. However, the first difference method relies on within-student variation in gender match across subjects. This means that I difference out math and science test scores to identify the gender match effects. One identifying

[1] Studies exploring the appropriate matching between individuals are not limited to the student-teacher interactions. In fact, some studies have, for example, investigated the effects of physician-patient racial concordance on patient satisfaction and quality of care (Strumpf 2011; Laveist and Amani 2002).
[2] See: Education Queensland, 2002.

assumption of the model is that those test scores are only affected by the teacher of the assessed discipline. This assumption might not hold as the student could transfer the skills acquired with one teacher to the other subject. To limit the effects of these spillovers, I restrict the assessment to the knowing cognitive part.[3] I consider subject-specific facts and concepts to be more resilient to such academic overlapping. With the restricted test scores, the results show that girls' science test scores improves by 0.042 standard deviation on average when assigned to a female teacher. This estimate is significant at the 10-percent level and is driven by advantaged girls.

I also evaluate the influence of gender interactions on students' non-cognitive outcomes (enjoyment, valuation, and confidence in the subject).[4] I find stronger positive effects of student-teacher gender match, which I observe predominantly on girls. For instance, the results show that girls' subject confidence index increases significantly by just over 0.130 standard deviations on average in both math and science when assigned to a female math teacher. The comparison with the gender match effects on academic performance suggests that an increase in non-cognitive outcomes does not necessarily lead to an improvement of students' test scores.

This paper contributes to the literature by providing new evidence using the latest available data from the TIMSS study. This research goes into greater depth than previous ones using TIMSS data as I focus on the different cognitive parts of the assessment to limit spillover effects and improve the robustness of the results. Moreover, my study presents findings on both students' cognitive outcomes and non-cognitive outcomes, while also exploring potential heterogeneous effects.

In this paper, I first provide an overview of previous research on this topic followed by a presentation of the student fixed effects model. The TIMSS data is then introduced before reporting the results observed on students' academic outcomes with OLS regressions and the first difference model. In the robustness check section, I examine the consistency of using only one of the five plausible test score values, compare the student characteristics means between the two subgroups with and without teacher gender differences, and present the results found when restricting the achievement outcomes to the knowing cognitive part of the assessment. After, I explore potential heterogeneous effects of gender match on students' academic

[3] The assessment is divided into three cognitive part: knowing, applying and reasoning that will be better explained in the data section. The main intuition is that the applying and reasoning methods learned could be more easily transferable across disciplines, disturbing the identification of the gender match effects.

[4] These variables will be measured through multiple surveys items.

performances. I also present estimates assessing the impact on students' perceptions of the subjects. Finally, the observed results are discussed, conclusions are drawn, and limitations of the method are presented.

2 Previous Research

Numerous studies have investigated the effects of the gender match between teachers and students in the past decades. This interest in the subject has intensified with the changing gender differences in educational outcomes.

In fact, it was widely considered in most countries that girls generally underperform boys in both math and science while outperforming them in language and reading skills (Xin Ma 2008). However, as girls started to catch up with boys in math and science during the late 20th century (see: Baker and Jones 1993), attention turned toward the significant role that teachers' genders could hold.[5] The feminization of the teaching profession, which was thought to benefit girls, was especially suspected to be the reason for this decrease in the gender gap.

Sommer (2000) supports this theory, which interprets such a trend as a failure of male education. In her book, she even goes as far as calling this a "war against boys". This directly contradicts the work of Myra and David Sadker (1995), who argue that gender bias results in girls receiving a worse education than boys.

The literature provides various explanations for the relevance of gender interactions between students and their teachers. One suggested mechanism is the phenomenon of stereotype threat, where students may experience a decline in performance if they consider themselves at risk of being judged based on stereotypes. Spencer, Steele and Quinn (1999)[6] demonstrates that the apprehension of negative stereotypes for girls in math affects their performance. In their study, women performed better when keeping the stereotype threat low by presenting the examination as having no gender differences and significantly worse when presented with gender differences. It is possible to assume that a same-gender teacher would contribute to lower this stereotype threat.

[5] For the most recent trend see: Meinck and Brese, 2019.

[6] See also: Steele, 1997.

We might also observe a Pygmalion effect, where teachers would express different academic expectations for boys and girls, depending on their own gender. This would become a self-fulfilling prophecy with students reacting to these expectations (Rosenthal and Jacobson 1968). However, the role model effect is certainly the mechanism that has received the most attention to explain a possible positive effect of a same-gender teacher (Almquist and Angrist 1971; Basow and Howe 1980). Assuming that students perceive teachers as role models and that they identify more closely with a teacher of the same gender, having the teacher's gender match the student's gender could help improve students' academic achievement and their global perceptions of the discipline. González-Pérez, Mateos de Cabo and Sáinz (2020) provide evidence of a role model effect in the specific case of women in STEM (science, technology, engineering, and mathematics) fields. Using the participation of women with STEM careers who volunteered to present their field to teenage girls in schools, they find a significant positive impact on girls' appreciation and consideration for these disciplines when they experience such role model sessions.

A variety of alternative methodologies have been used in previous research investigating the effects of same-sex teacher on students.

The simplest and probably the most intuitive method is to conduct studies using data where the assignment of students to teachers is random. In this case, a simple linear regression provides reliable results. The effort is then focused on verifying that the assignment is truly random. For instance, Lim and Meer (2020) employ the 2010 Seoul Longitudinal Study of Education (SELS2010), which includes data on middle school students and teachers who are randomly assigned to a class. Other studies exploiting random assignment in different settings are done by Carrell, Page and West (2009), Antecol, Eren and Ozbeklik (2014), and Gong, Lu and Song (2018). However, this type of data could be rather rare, and this technique is therefore restricted to a limited number of cases.

These studies show positive and sizable effects of same-sex teacher on female students, with the exception of Antecol, Eren and Ozbeklik (2014) who find negative effects of gender match on female students in disadvantaged neighborhoods. The positive results are observed on both students' cognitive outcomes (e.g., standardized test scores) and non-cognitive outcomes (e.g., attitudes towards a discipline, choice of courses and major).[7]

[7] Interestingly, Bettinger and Long (2005) use an instrumental variable to estimate the impact of same-gender faculty members on student's choice of major in college and find positive effects. However, this method does not appear to be commonly used for this question.

Alternatively, one of the most popular approaches to address the endogeneity problem is presented by Dee (2007)[8], where he uses a fixed effects method to eliminate the fixed unobservable student characteristics. This enables to contain part of the bias caused by the non-random assignment of students to teachers. The strategy is similar to the one used for data on monozygotic twin pairs (Ashenfelter and Krueger 1994; Ashenfelter and Rouse 1998; and Rousse 1999), but this time instead of observing one pair of twins, the same student is observed in two academic disciplines. To conduct this research, he uses the 1988 National Education Longitudinal Study (NELS:88), which provides middle school students' outcomes in two academic subjects (math or science and English or history) where their teachers are also surveyed. This approach has also been applied, sometimes with some variations, in many other studies (Ammermueller and Dolton 2006; Hoffman and Oreopoulos 2007; Holmlund and Sund 2008; Neugebauer, Helbig and Landmann 2011; Cho 2012; Paredes 2014; Alfa and Hermann 2017), becoming the most standard method in this literature.[9]

These studies' findings lead to mixed results, with some reporting positive and sizable effects of gender match (Dee 2007; Ammermueller and Dolton 2006; Paredes 2014) while others detect little or no effects (Hoffman and Oreopoulos 2007; Holmlund and Sund 2008; Neugebauer, Helbig and Landmann 2011).[10] Positive effects are also observed more frequently on female students. Much of this research focuses on how students are currently impacted by their present teacher, whereas the studies mentioned in the previous paragraphs exploiting random assignment also take interest in the long-term effects.

In general, all these findings are essentially contributing to the enrichment of the literature as they occur in different contexts (countries, time periods, school level) and employ diverse methodologies. Table A1 provides an overview of these various studies.[11]

Recent studies have used TIMSS data to test the potential universal teacher gender match effects by examining several countries. For example, Cho (2012) assesses the impact on 4th and 8th graders in fifteen OECD countries with a similar first difference model to Dee (2007). As either math or science is paired with English or history in Dee's model, the identifying assumption is

[8] See also : Dee, 2005.

[9] This method has also been used to study the effects of teacher characteristics other than gender (E.g. Bietenbeck, 2014).

[10] Ehrenberg, Goldhaber and Brewer (1995) actually use the same data as Dee (2007), but employ a simple linear regression with controls. They report no impact of student-teacher gender match, whereas with the method controlling for unobservable student characteristics, positive effects are found.

[11] Table A1 briefly summarizes the data used by indicating the country, year of collection and students school level. It also presents the methods used in the empirical analysis and the key results found. This table is presented in the Appendix.

that unobservable student characteristics (e.g., skills and preferences) are the same across those academic subject pairs. Cho argues that unobservable student traits are more accurately removed considering math and science can be paired with TIMSS data and the abilities required in those disciplines are more alike. It actually seems reasonable to consider math and science as a better control for each other than would be math with any other social science subject (see: Gardner and Hatch, 1989), and so I will also prefer this pair of disciplines for my analysis. Using this pairing, Cho finds that in eight countries out of fifteen a same-sex teacher has no significant effects on students' achievement, while there is a positive impact for boys in four countries and for girls in three countries.

Alfa and Hermann (2017) also use TIMSS 2003, 2007 and 2011 waves to investigate gender match effects for 8th graders in twenty European countries, performing some modifications to improve the reliability of their results. For example, they focus only on 8th graders because the teaching profession is generally highly dominated by females in 4th grade and does not allow for the necessary variation in the fixed effects model. They also restrict their data by eliminating observations of students in advanced level groups that they suspect of being more prone to selection bias. They find positive effects of gender match on student's academic achievement in half of the countries.

This paper offers new evidence from Australia and New Zealand using TIMSS 2011, 2015 and 2019 waves. I replicate the method employed by Dee to eliminate unobservable student characteristics while attempting, like other researchers before, to refine the model by adding certain precisions.

I compare the results found by using only students' performances on the knowing cognitive domain part of the assessment with that of the standard test scores. This means that I use the items assessing the pure knowledge of the basic concepts and facts that are specific to each discipline and ignore those requiring application or reasoning. This additional precision contributes to limiting suspected spillovers between subjects, thus increasing the reliability of my findings.

In this paper, I provide estimates evaluating the gender match effects on students' standardized test scores, but also on their global perceptions of the discipline based on multiple survey responses. I also investigate heterogenous effects using an index measuring students' home resources as a proxy for socioeconomic background.

This study uses only observations of lower secondary students in math and science classes. Moreover, I conduct my analysis on a limited number of countries that might have specific gender interactions.[12] It is therefore important to remain cautious when generalizing these findings.

3 Student fixed effects method

As previously discussed, I use the first difference (FD) method proposed by Dee (2007) to investigate the effects of student-teacher gender match. In the first specification, I assume that the academic achievement of the student i for math subject m is a function of observable student characteristics (S_i), whether the student's gender is the same as the teacher's (GM_{im}), observable teacher and classroom characteristics (T_{im}), unobservable student fixed effects (η_i) and a mean-zero error term (ε_{im}):

$$Y_{im} = \alpha_m + \beta GM_{im} + \gamma S_i + \delta T_{im} + \eta_i + \varepsilon_{im} \qquad (1)$$

Assuming a similar specification for student observed in science subject s, I also derive the following equation:

$$Y_{is} = \alpha_s + \beta GM_{is} + \gamma S_i + \delta T_{is} + \eta_i + \varepsilon_{is} \qquad (2)$$

Those simple linear models are estimated separately for girls and boys, the gender match variable (GM_{im}) can therefore be interpreted as a simple teacher gender dummy (equal to 1 if female and 0 if male when observing girls, while the opposite is true for boys). A positive β coefficient would indicate that gender match improves student test performances on average.

However, exploring the effects of teacher gender on student test scores with a simple OLS regression is likely to yield biased results due to endogeneity. In fact, the gender match variable may actually be correlated with the unobserved student fixed effects (ηi) that affect students' academic outcomes. For example, it is possible to imagine that the most troublesome students, who typically have lower academic achievement, are more likely to be placed by the school direction with a male (or female) teacher. This non-random assignment of teachers to students

[12] The gender match effects could potentially depend on the level of gender inequalities in the students' environment. In comparison to the world average, Australia and New Zealand can be considered as low gender inequality countries. For example, in the seventh round of the World Values Survey (2017-2020), the inhabitants of Australia and New Zealand appear to strongly support gender equality in education. To the question " A university education is more important for a boy than for a girl?" only 2.3% of the surveyed participants agree or strongly agree in Australia and 2.7% in New Zealand. These are the two lowest rates among the 57 countries surveyed. This trend can also be observed on the numerous other survey questions.

will not lead to the correct identification of the causal relationship between test scores and teacher gender, as the estimates will be downward biased for male teachers in this case.

However, by using the math and science test scores I have two observations that allow me to address unobservable student characteristics that are fixed across subjects. Since it is the same student taking both subjects, I can subtract these two observations from each other to remove student fixed effects. This means that I implicitly control for all fixed individual-specific factors (η_i), potentially removing a large source of omitted variable bias. It is important to emphasize that this within-student estimation controls only for subject-invariant effects, which means only those student characteristics (e.g., ability, preferences) that similarly impact achievement in both subjects. The proximity between disciplines is therefore essential to properly remove the student fixed effects, which is the reason for selecting TIMSS data with math and science as subject pairs. The first difference model (FD) is derived by differencing out equations (1) and (2) as follows:

$$Y_{im} - Y_{is} = \alpha_m + \beta GM_{im} + \gamma S_i + \delta T_{im} + \eta_i + \varepsilon_{im} - (\alpha_s + \beta GM_{is} + \gamma S_i + \delta T_{is} + \eta_i + \varepsilon_{is})$$
$$= \alpha_m - \alpha_s + \beta(GM_{im} - GM_{is}) + \delta(T_{im} - T_{is}) + \varepsilon_{im} - \varepsilon_{is} \qquad (3)$$

The estimations generated with the first difference model are identical to the fixed effects model as there are only two observations for each student. A positive coefficient for the gender match variable (GM_{im} - GM_{is}) would suggest that assigning a teacher of the same gender leads to positive effects on students' educational outcomes.

It should be noted that with this model the effect is only identified for students who actually change treatment status (students with different teacher genders in the two subjects) as the parameter β is identified due to within-student variation in gender match across subjects. This could result in difficulties in comparing the results with OLS estimates, since the latter uses both between and within-individual variation. It is not evident that these students with teachers of different genders are representative as they may have characteristics that are on average substantially different from the group of students with no teacher gender difference. It is thus necessary to verify that these subgroups are indeed similar before generalizing the estimation results.

In the fixed effects model, measurement errors in the variable of interest could be an issue as it can be shown that the downward bias is amplified due to the within dimension (see: Griliches and Hausman 1986). However, if I assume that gender is binary (either male or female) and I ignore the fact that some individuals may consider themselves to be of another gender identity, then it seems reasonable to presume that the gender of teachers and students were correctly reported.

The previous model assumes that the effect of teacher gender does not depend on the subject although this may be the case. For instance, it could very well be conceivable that the role-model effect is stronger in science than in math for girls if for some reason a female teacher can break down some of the negative stereotypes more strongly in that subject. To investigate this type of heterogeneity the following transformation of equations (1) and (2) allow for subject specific effects of gender match:[13]

$$Y_{im} = \alpha_m + \beta_m GM_{im} + \gamma S_i + \delta T_{im} + \eta_i + \varepsilon_{im} \quad (4)$$

$$Y_{is} = \alpha_s + \beta_s GM_{is} + \gamma S_i + \delta T_{is} + \eta_i + \varepsilon_{is} \quad (5)$$

And first differencing these equations again to remove student fixed effects:

$$Y_{im} - Y_{is} = \alpha_m - \alpha_s + \beta_m GM_{im} + \beta_s(- GM_{is}) + \delta(T_{im} - T_{is}) + \varepsilon_{im} - \varepsilon_{is} \quad (6)$$

The effect of teacher gender differs across subjects if the math coefficient β_m differs significantly from the science coefficient β_s. In all first difference regression models, the standard errors are clustered at the school level to allow for heteroscedasticity[14] and intra-school correlation country x subject x year dummies are included to control for time and country trends.

The main identifying assumption for this research is that unobservable student characteristics are identical across subjects and have the same influence on math and science test scores. There are several reasons why this assumption might not hold. It is conceivable that the assignment to a teacher of the same gender is correlated with student unobservable subject-specific characteristics (contained in the error term, ε_{im} and ε_{is}). If those subject-specific characteristics also impact student educational outcomes, the internal validity of the model could be threatened. For this reason, I decided to restrict the observations by dropping those where the

[13] The model with subject specific effects was also introduced in Dee's (2007) paper.

[14] In TIMSS data most of the schools are represented by only one of their class, thus this is similar to cluster at the class level.

class size appeared abnormally low, suggesting that the students could potentially be following a specific advanced track where selection based on subject-specific skills is more common.
It can also be the case that some unobserved teacher and classroom characteristics (e.g., teacher quality) are correlated to the student outcome and the variable of interest, which would bias my results.
Another potential concern is that the overlap between math and science could actually enable students to transfer knowledge gained from one subject with its associate teacher to solve problems in the other subject. The identification of student-gender match effects would be further challenged by these spillovers. As a solution to decrease these undesired effects, I propose to replace the standard test scores by those achieved when restricting to the knowing cognitive area of the assessment. The acquisition of methods for applying theory and constructing complex thinking seems much more transferable between subjects than simple facts and basic concepts. The knowing domain items are purely subject specific facts and thus can be used with greater confidence as student educational outcomes. To the best of my knowledge, this strategy has not been used before to identify student gender match effects with TIMSS data.

In this research, I will investigate the effects of same teacher gender on student outcomes by restricting my analysis to Australia and New Zealand. The geographical proximity of these two countries allows me to provide results on a region with a similar educational culture when regrouping the observations. This data expansion is ideal for the desired study as the student fixed effects method requires a considerable number of observations due to the identification being based on within-student variation. This also enables the exploration of heterogeneous effects while relying only on the most recent TIMSS study waves.

I also present results employing this same model but with student self-perceptions of the subject as the dependent variable. I use the indices provided by the TIMSS survey assessing students' liking, confidence, and valuation of the disciplines. These indices are constructed through multiple item responses. The same identifying assumption applies as once again only the unobservable student characteristics that are fixed across subjects are accounted for. However, this time these characteristics should similarly impact student subject perceptions instead of test scores. The correlation between teacher and classroom characteristics with the assignment of a same-gender teacher may again be an issue. As well as the potential spillovers problem, where for instance students would also start liking math more if positively influenced by a teacher role model in science.

4 TIMSS data

The Trends in International Mathematics and Science Study is an international comparative study that assesses fourth and eighth grade students' academic achievement in mathematics and science. This program conducted by the International Association for the Evaluation of Educational Achievement (IEA) was launched in 1995 and has been conducted every four years since then. This study, in which more than 60 countries have collaborated over the years, provides the opportunity to improve teaching and learning by interpreting the differences between educational systems.

TIMSS employs a stratified two-stage cluster sample design. In the first stage, a sample of at least 150 nationally representative schools are drawn with probabilities proportional to their number of students taught. In the second stage, one or more complete classes are randomly selected from within these sampled schools. Students from these classes will then be incorporated into the study's data set. The sampling process is jointly developed by the country's National Research Coordinator (NRC) and TIMSS sampling experts who will monitor, in particular, the targeted population and permitted exclusions, the sampling size and precision, the participation rates as well as the attribution of sampling weights.[15]

The interesting feature of this study is the possibility of having public access to data on students' academic performances and perceptions of mathematics and science, for which we also have background information on the teacher of each discipline. We can therefore consider math and science as subjects pairs in a fixed effects model using within-student variation.

One implication of this approach is that sufficient variation in teacher gender across disciplines is necessary to estimate the effects of student-teacher gender match. The data on 4th graders (around 10 years old) is therefore inadequate for our purposes for two reasons. First, it is quite common that at this level students are taught math and science by the same teacher, whereas this is no longer the case for 8th graders (around 14 years old). Second, in most countries the teaching profession is heavily dominated by women in 4th grade while it is more equally

[15] For further details on TIMSS sampling process see: Methods and Procedures in TIMSS 2015.

distributed between men and women in lower secondary school level.[16] As a result, my study will focus only on 8th grade students as we do not observe enough variation of teacher gender in elementary school math and science classes. Dee (2007) demonstrates in his paper that gender gaps in educational outcomes actually widen in 8th grade. Furthermore, existing research suggests that students start to realize and apply gender stereotypes around this age (e.g. Ruble and Martin 2004). Therefore, this specific choice of school year seems to be highly relevant for my evaluation.

Considering that in this study I desire to conduct a more in-depth analysis than the previous papers using TIMSS data, it seems necessary to restrict the estimations to a limited number of countries. After careful consideration, Australia and New Zealand appeared to be the ideal candidates for several reasons. Firstly, these two countries have participated regularly in the study waves, which is not always the case with some countries participating only occasionally. The number of student observations for these nations tends to be quite large with also a reasonable number of missing values compared to other participants. Furthermore, due to the proximity of these two countries, I will be able to analyze the effects of same-sex teacher on students' outcomes in that specific region by combining the observations. This will further increase my capacity for precision, which is highly necessary for the various investigations I wish to perform.

Some reasons are specific to the Australian and New Zealand education systems, such as the compulsory education until the age of 16, which ensures that the students being assessed are representative of the whole population.[17] Students also have different teachers for math and science and the ratio of female teachers is about 55% in these countries. Moreover, unlike some countries where teachers teach exclusively to students of the same gender, here strong gender segregation in the classroom is not observed, which allows me to apply student fixed effects.

In this study, I therefore use data for Australia and New Zealand from the 2011, 2015 and 2019 waves of TIMSS. This ensures that I have results based on recent developments and can benefit from the latest set of variables added to the study.

[16] Most countries have more than 80% of female teachers in 4th grade, whereas it is closer to 60% in 8th grade. For further details about female teacher ratio by country in 2015 also see: TIMSS 2015 Fourth Grade Almanacs and TIMSS 2015 Eighth Grade Almanacs.

[17] Details on compulsory school can be found on the following government education websites:
https://www.studyaustralia.gov.au/english/study/education-system
https://www.education.govt.nz/our-work/our-role-and-our-people/education-in-nz/#primary

The fact that students take standardized tests independent of their teachers that could have biased the results is a clear advantage of this study. Furthermore, it is important to highlight that student assessment can be separated into different content areas. In mathematics the four sections tested are number, algebra, geometry, and data and chance while in science the students should be familiar with biology, chemistry, physics, and earth sciences. The examination is also divided into three different cognitive domains. The first one, knowing, requires the student to know the basic facts and concepts of the subject. The other two focus on the student's applying and reasoning skills.[18]

Given the similarities between the two disciplines, students might use abilities learned in one discipline with its associated teacher to solve questions from the other subject. However, it seems intuitive that these spillovers would be primarily caused by items requiring the application of knowledge to solve a problem or the development of complex reasoning since these skills seem to be more easily transferable than simple discipline-specific facts. For instance, a student could recondition the application or reasoning skills acquired with his or her math teacher for the science class, but it seems unlikely that his or her knowledge of basic math facts and concepts would be of any help in that science course. Therefore, in this study, I will compare the results achieved using the scores of the whole assessment with those restricting the scores to the knowing cognitive domain.

However, in order to limit the burden on students while ensuring comparable information on students' knowledge and capacities across different domains, TIMMS did not administer all items to every student. Instead, they responded to only a fraction of the assessment's items. These scores were then transformed into plausible values during the scaling process. These plausible values can be considered as multiple random draws from a distribution of score values derived from the students' responses on the limited test items and its background characteristics information.

Taking only one plausible value would result in unbiased estimates, as opposed to using the mean of plausible values as the test score, but the standard errors would certainly be underestimated. Aparicio, Cordero and Ortiz (2021) still recommend the adoption of a single plausible value to conduct exploratory research as the difference in results when using one or

[18] More precisely, in mathematics the test items concern the number part at 30%, algebra at 30%, geometry at 20%, and data and chance at 20%, while for science it is biology at 35%, chemistry at 20%, physics at 25% and earth science at 20%. For the cognitive skills needed to solve the items, in mathematics 35% knowing 40% applying and 25% reasoning, while in science 35% knowing 35% applying and 30% reasoning. Further details on the assessment can be found in: TIMSS 2015 Assessment Frameworks.

multiple plausible values in a large sample is minor. However, it is suggested to use all plausible values to increase the accuracy of the estimates.

The plausible values procedure was introduced by Mislevy (1991), where he actually applies the multiple imputation methodology developed by Rubin (1987). This approach consists in collecting the estimates individually for each plausible value and then computing the average of those estimates.[19]

In my case, I have five plausible values for the students' scores in each domain as well as for the whole assessment, which is adequate according to Bibby (2020).[20] In this paper, I will present the main tables using only one plausible value while showing the robustness of the results by using five plausible values in the Appendix.[21]

I thus reach a total of 46483 student observations without imposing any restrictions. I eliminate, however, some observations with missing values, 785 with unknown student gender, around 1800 with students having not exactly one (different) teacher in each subject, and 8000 with unknown teacher gender. I also remove around 4800 observations where all students and teachers in the classes are of the same gender. The purpose here is to typically exclude segregated schools (e.g. church schools) that do not permit in any case a first difference estimation with teacher-student gender match varying across subjects. Finally, the observations with a class size smaller than 6 have also been dropped as there could be some missing observations of students in these classes which are bigger in reality, or those students could be part of a specific program. This results in a total of 28045 student observations coming from 978 different schools.

TIMSS provides a large amount of information about students' home environments and school lives, as well as teacher characteristics and classroom conditions, which are all gathered through questionnaires. The fixed effects model used in this research includes the following teacher and classroom characteristics: Teacher age measured by 6 categorical dummies, years of experience, formal education level measured by 4 categorical dummies, teacher self-perception of disruption and disinterest in classroom measured by 3 categorical dummies, teacher self-perception of safety, satisfaction, and emphasis on academic success in school measured with

[19] Details of the plausible values procedure is also available in: Methods and Procedures in TIMSS 2015.

[20] Bibby, who questions in his paper the potential effect of increasing plausible values beyond five, demonstrates that the sample size has a larger impact on the estimations than the number of plausible values. Therefore, considering our large number of observations, five plausible values seem quite sufficient.

[21] For more details on how to use plausible values in Stata see: Macdonald, 2008.

indices constructed from multiple questionnaire responses[22], class size, female students ratio in classroom, and student born in the country ratio in classroom. This model also includes country x subject x year dummies.

I also use some student characteristics in the OLS regressions I perform, which are not relevant in the fixed effects estimation using within-student variation. I have included variables describing the age of the student, whether they were born in the country or not, a home resources index and other indices for subject liking, confidence, and value.[23] These variables are also calculated through the questionnaire questions.

Summary statistics of these variables are reported in Table A2 that can be found in the Appendix. In this table, we can observe that math seems to be more valued than science, both for students and students' parents. It is noteworthy that teacher characteristics are fairly similar between subjects except for the age, years of experience and formal education. Math teachers tend to be on average a bit older, with more experience and with lower formal education. Table A3 also indicates that gender differences in teacher characteristics are small compared to those observed between subjects.

The appropriateness of identifying gender as solely male or female can be debated, as a broad range of gender identities could be possible. The simple separation into conventional gender identities possibly does not represent self-perceptions of gender differences well. My gender match effects analysis does therefore depend on gender interactions between teachers and students, which might sometimes not be representative of their self-identification.

For this reason, it would be very useful to have more information on the actual gender identification of students and teachers in order to conduct more accurate research. Studies collecting such data are rather rare for the moment but could emerge in the near future.[24] The possibility of elaborating data with non-binary gender options still involves many challenges.

For instance, there could be inconsistency between schools on the different options presented, but also difficulties of recognizing all students' gender identities that would vary strongly depending on the level of inclusiveness in schools. Moreover, it seems probable that an

[22] The teacher self-perception variables will be added in a different column as these control could themselves be outcomes of the treatment variable and thus be considered as bad controls. However, the teacher perception of disruption relies on the whole class behavior so the gender match variable would have at best only a minor effect on the teacher perceptions of the class. I therefore consider relevant to include them in my tables.

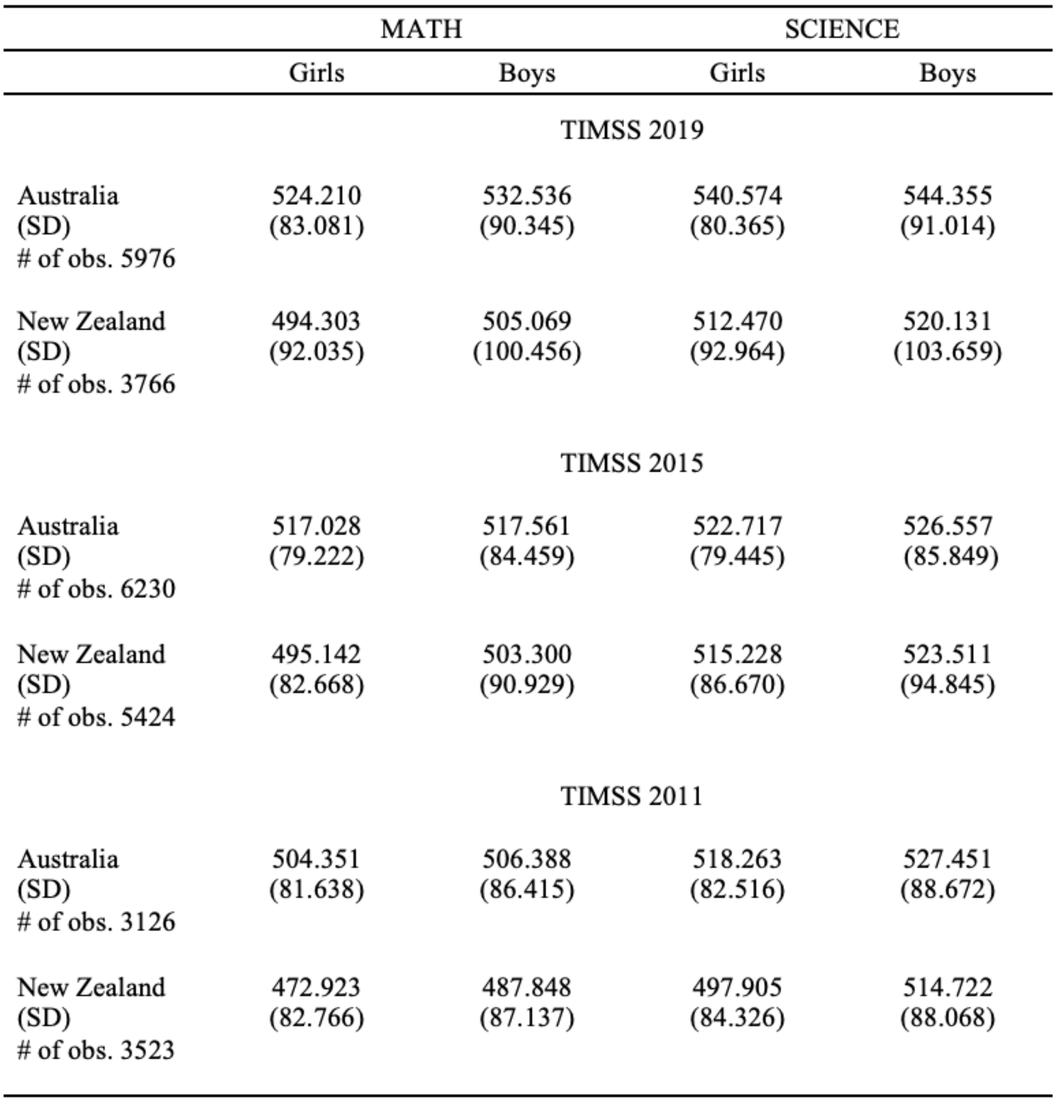
[23] For instance, to measure the home resources index questions such as whether the student has his or her own room, a study desk, or a computer tablet are asked. Therefore, I consider that home educational resources can actually be interpreted as a good proxy for students' socio-economic background.

[24] Herold (2022) and Ujifusa (2021) in their respective article illustrate well the gain of interest in this topic in Pennsylvania and the challenges associated with it.

international study like TIMSS would face political reluctance in various parts of the world with such data collection.

Therefore, my analysis based simply on binary genders matching independent on self-identification appears to be the best approach at this time. After all, there are still major differences in outcomes between both of these genders. Many researchers also discovered significant effects of gender match which provide valuable lessons. However, it must be recognized that the results I find ignore the participants' own possible sense of non-affiliation to the attributed gender.

Table 1: Average Scores By Country, Subject And Gender For TIMSS 2011, 2015 And 2019 (Using PV1)



	MATH		SCIENCE	
	Girls	Boys	Girls	Boys
		TIMSS 2019		
Australia	524.210	532.536	540.574	544.355
(SD)	(83.081)	(90.345)	(80.365)	(91.014)
# of obs. 5976				
New Zealand	494.303	505.069	512.470	520.131
(SD)	(92.035)	(100.456)	(92.964)	(103.659)
# of obs. 3766				
		TIMSS 2015		
Australia	517.028	517.561	522.717	526.557
(SD)	(79.222)	(84.459)	(79.445)	(85.849)
# of obs. 6230				
New Zealand	495.142	503.300	515.228	523.511
(SD)	(82.668)	(90.929)	(86.670)	(94.845)
# of obs. 5424				
		TIMSS 2011		
Australia	504.351	506.388	518.263	527.451
(SD)	(81.638)	(86.415)	(82.516)	(88.672)
# of obs. 3126				
New Zealand	472.923	487.848	497.905	514.722
(SD)	(82.766)	(87.137)	(84.326)	(88.068)
# of obs. 3523				

Note : The means of test scores are measured for each country, subject and gender in all study waves using only one plausible value.[25] The standard deviations are lower than when all plausible values are used because the uncertainty from the distribution is ignored here.

[25] In the Appendix, Table A4 presents the results found with all 5 plausible values.

Table 1 exhibits a clear gender gap in test scores, with boys performing better on average in both countries and subjects for all study waves. This gap does not necessarily disappear over time. Males test scores seem to persistently have greater standard deviations which is consistent with the results of Hedges and Nowell (1995) who find an overrepresentation of boys in the extremes of the distribution.[26] A substantial overlap in the distribution of female and male test scores is still observed.

5 Results

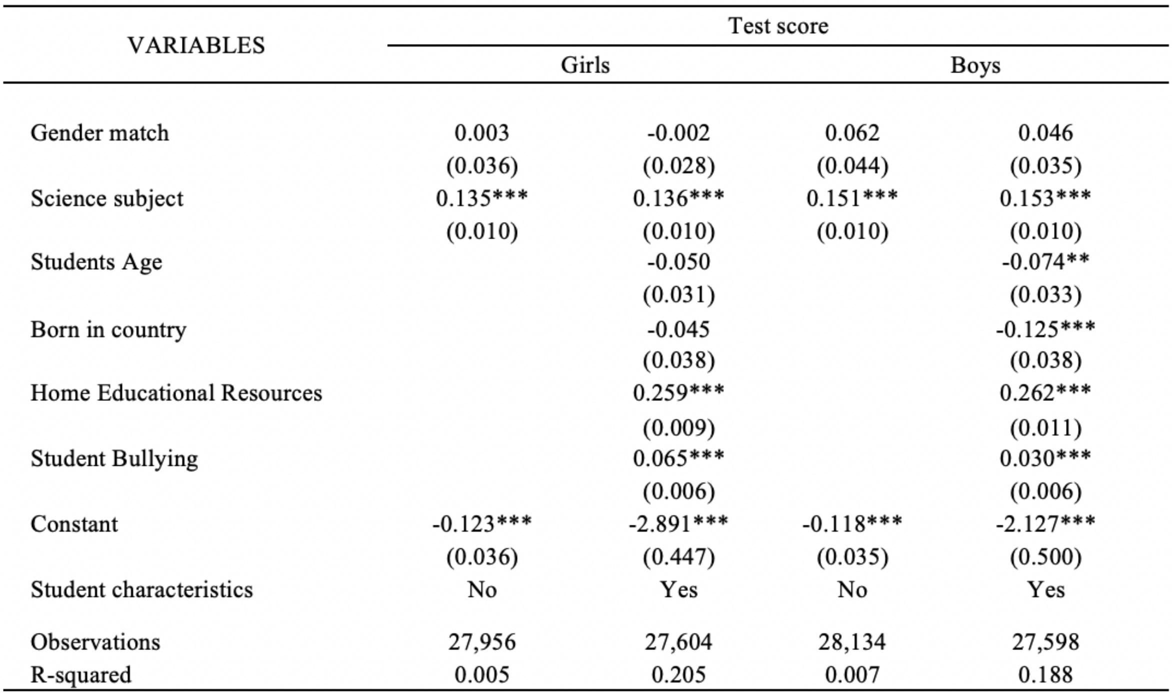
5.1 Pooled OLS Regression

As discussed earlier, when performing a simple analysis using OLS regressions, the estimates could be biased because of omitted variables. This method is unable to control for unobservable student characteristics, ηi, from equation (1) which can be correlated to the variable of interest (GM_{im} - GM_{is}). It is nevertheless worthwhile to observe the results produced with this method and especially the impact of adding student characteristics on the estimated gender match coefficients.

Therefore, I pool all test scores together as if there was a unique assessment and run an ordinary least squares regression. I include the subject as a control in the regression with a dummy variable. The model analyzes how the assessment scores are impacted on average if the teacher is of the same gender. Some student characteristics are added as controls to evaluate the extent of possible omitted variable bias. Included in this regression are the age of the student, a dummy variable determining whether the student was born in the country or not and two indices calculating student's home educational resources and the bullying experienced at school. In order to better interpret the magnitude of these effects to facilitate comparison with previous research, I standardized the test scores in all my tables. I therefore subtract the mean from the educational outcomes and divide over the standard deviation so the mean is equal to 0 and the standard deviation equal to 1. The following table presents the results of this regression.

[26] In the Appendix, Table A5 presents kernel densities of student's performance in math and science by gender. There, we can easily observe that the curve is more concentrated at the center for girls.

Table 2: Pooled OLS Regression Of Test Scores On Gender Match (Using PV1)



VARIABLES	Test score			
	Girls		Boys	
Gender match	0.003	-0.002	0.062	0.046
	(0.036)	(0.028)	(0.044)	(0.035)
Science subject	0.135***	0.136***	0.151***	0.153***
	(0.010)	(0.010)	(0.010)	(0.010)
Students Age		-0.050		-0.074**
		(0.031)		(0.033)
Born in country		-0.045		-0.125***
		(0.038)		(0.038)
Home Educational Resources		0.259***		0.262***
		(0.009)		(0.011)
Student Bullying		0.065***		0.030***
		(0.006)		(0.006)
Constant	-0.123***	-2.891***	-0.118***	-2.127***
	(0.036)	(0.447)	(0.035)	(0.500)
Student characteristics	No	Yes	No	Yes
Observations	27,956	27,604	28,134	27,598
R-squared	0.005	0.205	0.007	0.188

Note: The science subject variable is equal to 1 if the test score is in science and 0 if in math. Robust standard errors clustering at school level are reported in parentheses. The test scores are standardized as to have a 0 mean and a standard deviation of 1. Sampling weights specific to each student are also used. Results are achieved by using of the first plausible value.[27]

* Statistically significant at the 10-percent level.
** Statistically significant at the 5-percent level.
*** Statistically significant at the 1-percent level.

Table 2 shows small and insignificant coefficients of gender match, although to boys it is close to being significant at the 10-percent level when regressed without controls. Both for girls and boys, the size of the coefficient becomes smaller when controls are added. This shift is especially notable for boys where the coefficient falls from 0.062 SD to 0.046 SD.[28]

The non-negligible influence of these controls on the estimates confirms that the assignment is not perfectly random. It is reasonable to suspect that other student characteristic variables that I have omitted from the regression or that are simply not observable could bias the results due to their correlation with the variable of interest. Worth noting in this table is also the expected importance of the home educational resources on test scores: a one-unit increase in the home-resource index is associated with a 0.25 SD increase in test scores.[29]

[27] Estimates found using all plausible values can be seen in Table A6 from the Appendix.

[28] It is precisely results of this magnitude that Dee (2007) finds in his paper, hence the importance of seriously considering these effects, even though insignificant in my case. He finds in fact that same-gender teacher assignment improves girls test scores in history by 0.075 standard deviations and decrease math test scores by 0.066 standard deviations. While for boys he finds that the gender match improves both math and science test scores by 0.078 and by 0.048 standard deviations respectively.

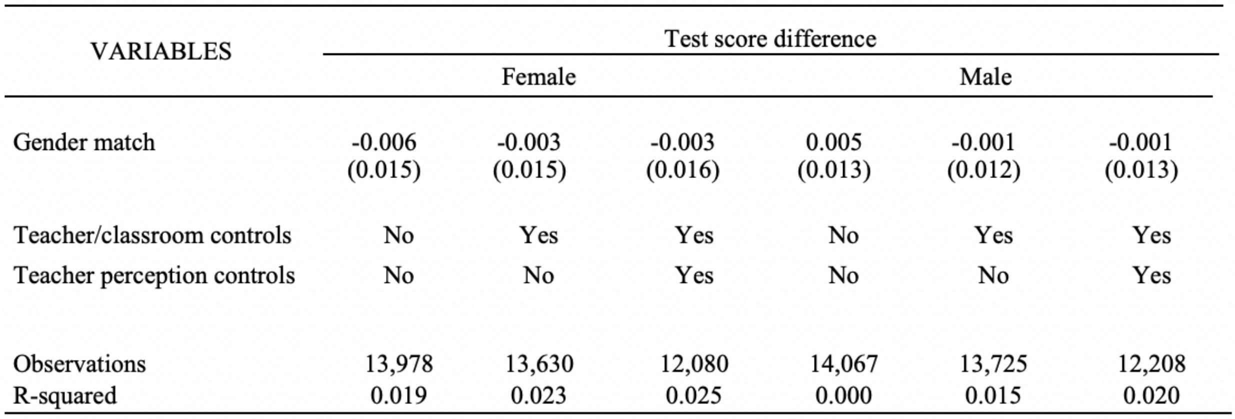
[29] This index has a mean of 11.105 and goes from 4.323 to 14.018, as showed in Table A2.

5.2 Fixed effects model

To overcome the endogeneity of the OLS model, I now employ a first difference method (eq.(3)) to control for all subject invariant student characteristics. Differencing out over mathematics and science removes a potential important source of omitted variable bias.

The following teacher and classroom characteristics are included in the model: Teacher age, experience, formal education level, perception of classroom disturbance and disinterest as well as school safety, satisfaction and emphasis on academic success, class size, classroom female ratio and classroom native born ratio. Each math teacher variable is thus subtracted by the equivalent science teacher variable. When there is a dummy variable such as teacher age (6 categorical dummy variables), each categorical dummy variable in math is subtracted by the corresponding one in science. The controls based on teacher self-perception will be added to a third column over concerns that these are bad controls. The variables could actually be outcomes of same-gender teacher assignment. For example, if the gender match would cause the teacher to judge that the classroom is more disruptive. However, since the teacher's perceptions are formed at the classroom or school level, the individual effects of same-gender assignment are likely to be small. The cautious inclusion of these variables can therefore be of interest.

Table 3 : First Difference Regression With And Without Controls (Using PV1)



VARIABLES	Test score difference					
	Female			Male		
Gender match	-0.006 (0.015)	-0.003 (0.015)	-0.003 (0.016)	0.005 (0.013)	-0.001 (0.012)	-0.001 (0.013)
Teacher/classroom controls	No	Yes	Yes	No	Yes	Yes
Teacher perception controls	No	No	Yes	No	No	Yes
Observations	13,978	13,630	12,080	14,067	13,725	12,208
R-squared	0.019	0.023	0.025	0.000	0.015	0.020

Note: Teacher and classroom controls listed in the data section are used in this model. Those resulting from the teacher's perception of the classroom (disruption, disinterest) or school (focus on academic achievement, safety, satisfaction) are added in an additional column because of the risk of bad control problem. Country x subject x year dummies are included for time and country trends. Robust standard errors clustering at school level are reported in parentheses. Only the first plausible value is used and test scores are standardized. Sampling weights are also applied.
* Statistically significant at the 10-percent level.
** Statistically significant at the 5-percent level.
*** Statistically significant at the 1-percent level.

The results of this estimation are shown in Table 3. It must be remembered that the gender match effects are only identified for students with two teachers of different genders. This is due to the design of the model that uses within-student variation in gender match across subjects. These results are thus driven by 7023 observations for girls and 6625 for boys. Moreover, this method only removes subject-invariant effects and therefore does not control for possible bias coming from student abilities that differ across subjects.

In Table 3, the coefficients of gender match are small and insignificant when student fixed effects are accounted for. The inclusion of teacher and classroom controls prevents a possible correlation between these and the gender match variable to bias the results. In this case, these controls have only a minor effect on the coefficients.

However, it would be reasonable to consider the possibility that the effects of teacher gender depend on the subject. If this were the case, the insignificant estimates observed in the table could actually hide some interesting heterogenous effects. For example, a female teacher might have a positive impact on girls' math scores but an equivalent negative impact in science. In the following, I explore this potential heterogeneity.

5.3 Subject specific effects

Table 4 reports the estimated effects of gender match on test scores that allow for subject specific effects. The same method as previously is employed, but this time it is assumed that the coefficients for same-sex teachers are different between subjects when differencing out both equations (eq.(6)). If the effect of teacher gender varies by the subject being taught it should be observed that the coefficients β_m and β_s are significantly different.

The estimates in Table 4 do differ slightly depending on the subject. Student-teacher gender match appear to have a positive effect on math test scores and a negative effect on science test scores for both genders. However, these effects remain small and not significant. These opposite coefficient signs explain why the effects were even weaker in the previous table, which did not account for subject specific effects.

To evaluate if there is a significant difference between estimates, I conduct an F-test with the hypothesis that both coefficients are equal. I do find that this difference is significant at the 10-percent level for girls, but not for boys since the effects are smaller for them. This exemplifies how relevant it can be to control for the subject specific effects.

It is also noteworthy to observe that the controls do not have a strong impact on the estimates which could suggest that there would not be a strong bias coming from omitted teacher and classroom characteristics. This is encouraging as it might also apply to unobservable characteristics that we cannot control with this model.

Table 4: First Difference Regression Allowing For Subject Specific Effect Of Gender Match (Using PV1)

VARIABLES	Test score difference					
	Female			Male		
Gender match in Math	-0.025	-0.021	-0.023	-0.013	-0.016	-0.017
	(0.019)	(0.019)	(0.021)	(0.018)	(0.018)	(0.019)
Gender match in Science	0.011	0.016	0.018	0.012	0.013	0.014
	(0.018)	(0.018)	(0.019)	(0.018)	(0.018)	(0.019)
Teacher/classroom controls	No	Yes	Yes	No	Yes	Yes
Teacher perception controls	No	No	Yes	No	No	Yes
P-value (H_0: $B_m = B_s$)	0.096	0.097	0.083	0.329	0.278	0.256
Observations	13,978	13,630	12,080	14,067	13,725	12,208
R-squared	0.020	0.023	0.026	0.012	0.015	0.020

Note: This model allows gender match coefficients to be different for math and science. Teacher and classroom controls listed in the data section are again used. Country x subject x year dummies are included for time and country trends. Robust standard errors clustering at school level are reported in parentheses. Only the first plausible value is used and test scores are standardized. Sampling weights are also applied. The p-value is calculated with an F-test testing that the two coefficients are equal.

* Statistically significant at the 10-percent level.
** Statistically significant at the 5-percent level.
*** Statistically significant at the 1-percent level

6 Robustness check

6.1 All plausible values

As previously discussed, students do not answer to all assessment items. That is why this analysis uses plausible value of test scores as a dependent variable. Although the use of a single plausible value may be considered satisfactory in large samples as it leads to unbiased estimates, it is still recommended to consider all of the plausible values to improve the precision of the estimates. The method to adopt when results with all plausible values are desired is simple since it requires only to perform the estimations separately for each plausible value and average them together. The calculation of the standard errors takes into account the variation of the estimates found with all the different plausible values, which will typically result in larger standard errors. I use the Stata module proposed by Macdonald (2019) to conduct my estimations with all plausible values. In this program standard errors are calculated with a Jackknife procedure specific to TIMSS data.

Table 5 : First Difference Regression With And Without Subject Specific Effect Of Gender Match (Using All PVs)

VARIABLES	Test score difference					
	Female			Male		
	Without Subject Specific Effect Of Gender Match					
Gender match	-0.004 (0.014)	0.001 (0.016)	0.001 (0.019)	-0.001 (0.013)	-0.001 (0.013)	0.002 (0.013)
Teacher and class controls	No	Yes	Yes	No	Yes	Yes
Teacher perception controls	No	No	Yes	No	No	Yes
Observations	13,978	13,630	12,080	14,067	13,725	12,208
R-squared	0.017	0.021	0.024	0.012	0.014	0.020
	With Subject Specific Effect Of Gender Match					
Gender match in Math	-0.019 (0.020)	-0.017 (0.021)	-0.021 (0.023)	-0.019 (0.015)	-0.019 (0.015)	-0.016 (0.017)
Gender match in Science	0.011 (0.017)	0.019 (0.019)	0.023 (0.023)	0.018 (0.019)	0.017 (0.017)	0.020 (0.017)
Teacher/classroom controls	No	Yes	Yes	No	Yes	Yes
Teacher perception controls	No	No	Yes	No	No	Yes
Observations	13,978	13,630	12,080	14,067	13,725	12,208
R-squared	0.018	0.022	0.025	0.012	0.015	0.021

Note: A Jackknife bootstrapping method specific to TIMSS data is applied to calculate standard errors. All plausible values are used after being standardized. This model is otherwise identical to the previous tables.
* Statistically significant at the 10-percent level.
** Statistically significant at the 5-percent level.
*** Statistically significant at the 1-percent level.

Table 5 presents the results of this estimation using all five plausible values. The reported estimates are closely similar to those seen previously in Table 3 and 4. The differences are minor and do not reverse my conclusions. This confirms that the use of a single plausible value would be sufficient to perform this analysis with reliability. In the following, I will therefore present my findings using only one plausible value while still providing a table using all plausible values as a robustness test in the Appendix. It will be noted that each time the estimates are similar.

6.2 The mean comparison between two subgroups

Since the main student fixed effects model employs within-student variation across subjects, the gender match estimate is actually identified for students that have two teachers of different genders. It is fundamental to explore if those students are representative of the whole sample. For this purpose, I compare the means of student characteristics to examine potential subgroup differences. In Table 6, the means of the following variables are compared for each gender separately: student age, home resources, bullying experienced, math and science liking, valuation and confidence, native born, as well as mother and father native born.

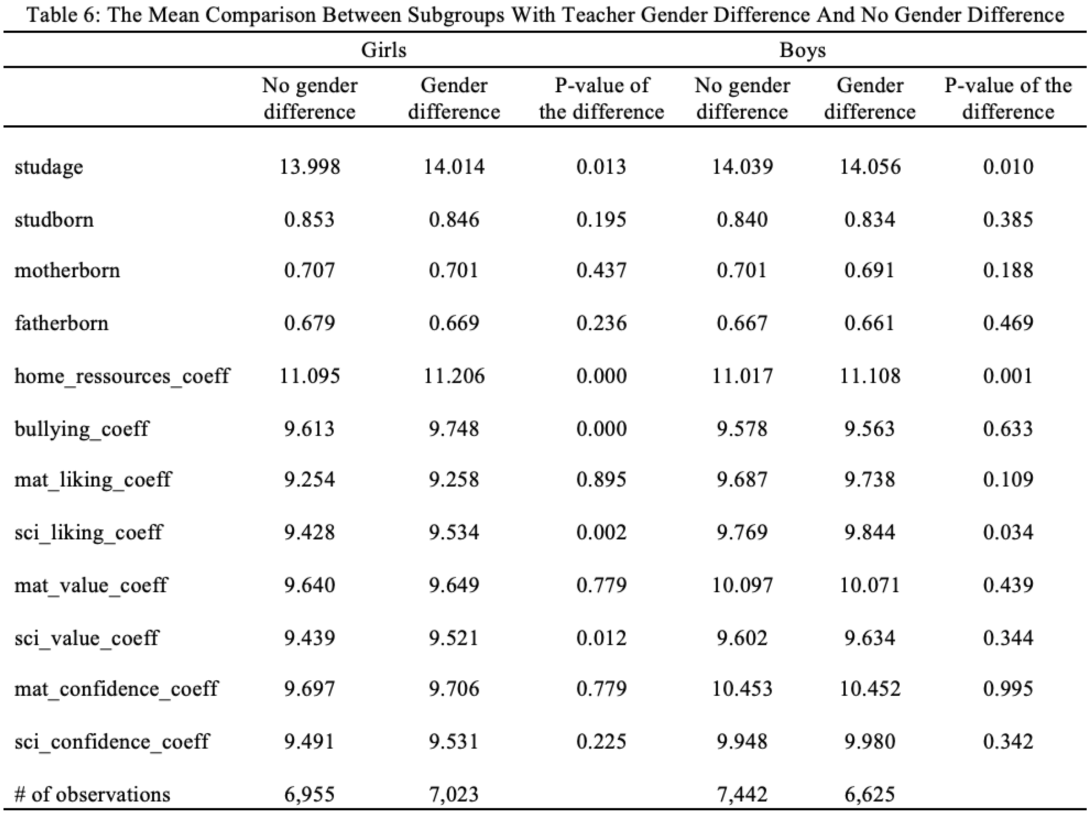


Table 6: The Mean Comparison Between Subgroups With Teacher Gender Difference And No Gender Difference

	Girls			Boys		
	No gender difference	Gender difference	P-value of the difference	No gender difference	Gender difference	P-value of the difference
studage	13.998	14.014	0.013	14.039	14.056	0.010
studborn	0.853	0.846	0.195	0.840	0.834	0.385
motherborn	0.707	0.701	0.437	0.701	0.691	0.188
fatherborn	0.679	0.669	0.236	0.667	0.661	0.469
home_ressources_coeff	11.095	11.206	0.000	11.017	11.108	0.001
bullying_coeff	9.613	9.748	0.000	9.578	9.563	0.633
mat_liking_coeff	9.254	9.258	0.895	9.687	9.738	0.109
sci_liking_coeff	9.428	9.534	0.002	9.769	9.844	0.034
mat_value_coeff	9.640	9.649	0.779	10.097	10.071	0.439
sci_value_coeff	9.439	9.521	0.012	9.602	9.634	0.344
mat_confidence_coeff	9.697	9.706	0.779	10.453	10.452	0.995
sci_confidence_coeff	9.491	9.531	0.225	9.948	9.980	0.342
# of observations	6,955	7,023		7,442	6,625	

Note: This Table presents means for both boys and girls of selected student characteristics by two subgroups determined by whether the teacher gender changes across subjects or not. The third and sixth column shows the p-value testing the difference between means.

Table 6 shows that for girls the variable measuring home resources, student age, bullying, science liking, and science valuation are all significantly different at the 5-percent level at least, while this is only the case of student age, home resources, and science liking for boys. This may appear to threaten the possibility of results generalization. It would be problematic if the effects of gender match were to vary for students with diverse home resources as this variable's subgroup means are significantly different for both genders.[30] However, these differences remain quite small. For example, the student age mean difference for boys is 0.016 which is equivalent to less than 6 days. Overall, while there are a few significant differences between the students driving the variation and the general sample, these differences are small, so the results can be considered to be generalizable. It must be remembered that the data has been considerably restricted due to lack of observations or gender segregation in the schools, which

[30] This is actually explored later in the section on heterogeneous effects.

may have altered the students' characteristics slightly between subgroups. Moreover, although the number of student observations is large, the number of schools and teachers is much smaller. This facilitates the creation of observable differences between students.

6.3 Knowing cognitive domain test scores

Considering the connections between math and science, potential spillovers could as previously mentioned further complicate the identification of student-teacher gender match effects. If the skills learned can be transferred across subjects, then the gender of the math teacher could, for example, affect the students' science test score. I make the assumption that the instruction of theory application and development of sophisticated reasoning is more frequently shared between disciplines than are subject-specific facts. Table 7 displays the estimates I find when only considering student responses to the assessment's knowing cognitive domain section. These test scores are also measured with plausible values that I standardize prior to the analysis. Except for this modification, the model remains identical.

The economic magnitude of the estimates shown in Table 7 without subject-specific effects is stronger than observed before (in Table 3), with female teachers improving both boys and girls test score on average. However, these effects remain minor since the coefficients do not exceed 0.017 SD. Still, this change may indicate that the previous results did not precisely identify the influence of female teachers on test scores due to spillovers. If we assume that female teachers improve learning acquisition for both male and female students on average, then the assignment to a female teacher could have positive effects not only on her discipline but also on the discipline of the second teacher, who may be male. If this is the case the positive effects of female teacher on both girls and boys will be underestimated. Thus, these findings based on questions about subject-specific facts provide greater reliability.

The results found with subject specific effects are even more interesting since I find that the assignment to a female teacher improves girls' science test scores by 0.042 SD on average with teacher and classrooms controls.[31] It is the only effect that is statistically significant at the 10-

[31] Possible heterogeneity across countries and years was explored for this effect. I found no significant heterogeneity across countries, as the effect is present in both Australia and New Zealand. However, this effect appear to be driven by the 2011 and 2015 study waves.
Interestingly, when analyzing New Zealand with TIMSS study waves from 1995 to 2007, Cho (2012) already finds that the gender match effects are strongest for girls in science, with an estimate of 0.030 SD. This could indicate that this effect is persisting over time.

percent level.[32] The difference between math and science estimates (β_m - β_s) is also significant for girls. It is noticeable that the impact of controls on the estimates is stronger in this table, although no drastic change is observed.

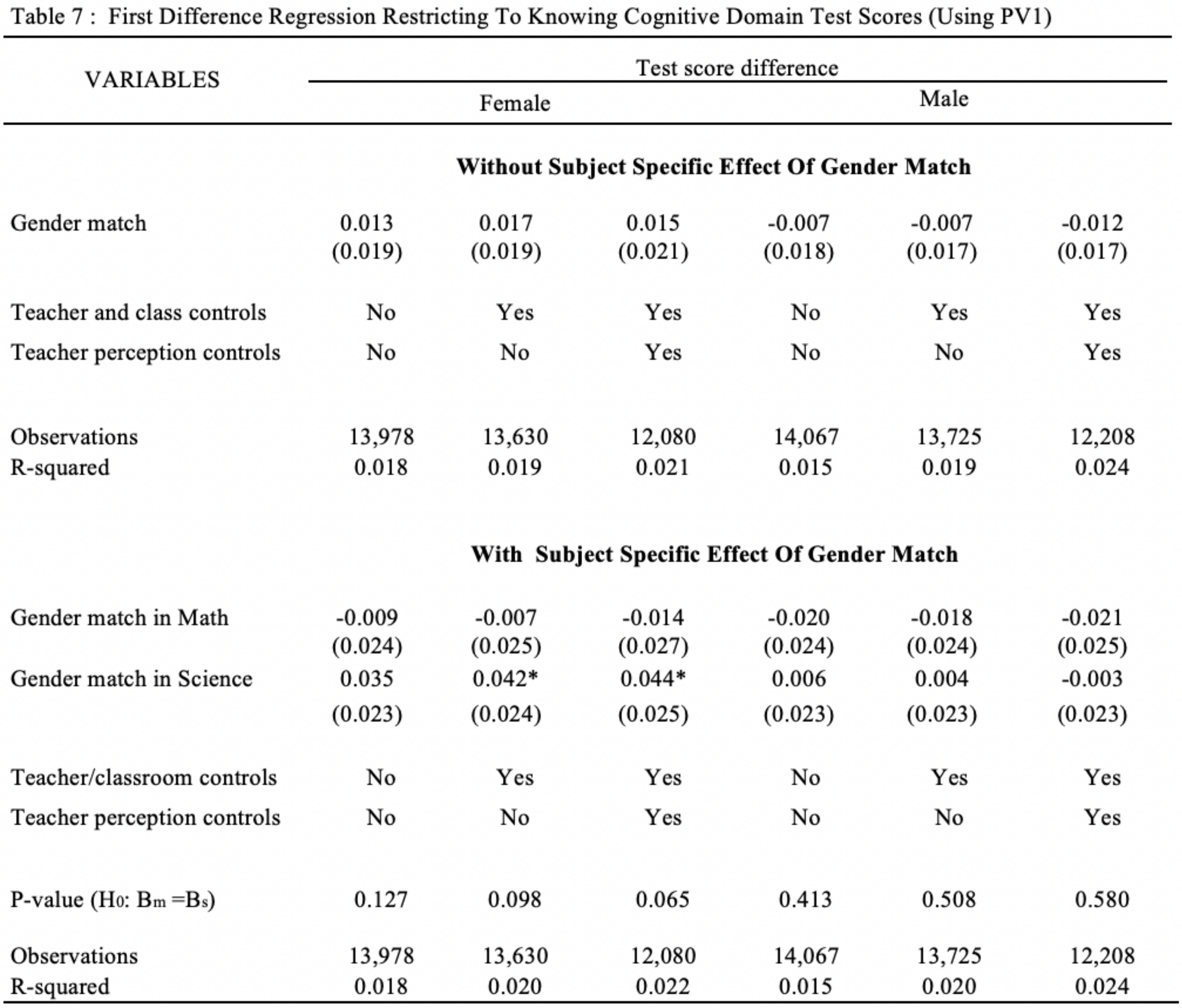


Table 7 : First Difference Regression Restricting To Knowing Cognitive Domain Test Scores (Using PV1)

VARIABLES	Test score difference					
	Female			Male		
	Without Subject Specific Effect Of Gender Match					
Gender match	0.013 (0.019)	0.017 (0.019)	0.015 (0.021)	-0.007 (0.018)	-0.007 (0.017)	-0.012 (0.017)
Teacher and class controls	No	Yes	Yes	No	Yes	Yes
Teacher perception controls	No	No	Yes	No	No	Yes
Observations	13,978	13,630	12,080	14,067	13,725	12,208
R-squared	0.018	0.019	0.021	0.015	0.019	0.024
	With Subject Specific Effect Of Gender Match					
Gender match in Math	-0.009 (0.024)	-0.007 (0.025)	-0.014 (0.027)	-0.020 (0.024)	-0.018 (0.024)	-0.021 (0.025)
Gender match in Science	0.035 (0.023)	0.042* (0.024)	0.044* (0.025)	0.006 (0.023)	0.004 (0.023)	-0.003 (0.023)
Teacher/classroom controls	No	Yes	Yes	No	Yes	Yes
Teacher perception controls	No	No	Yes	No	No	Yes
P-value (H_0: $B_m = B_s$)	0.127	0.098	0.065	0.413	0.508	0.580
Observations	13,978	13,630	12,080	14,067	13,725	12,208
R-squared	0.018	0.020	0.022	0.015	0.020	0.024

Note: This model uses the student test scores on the knowing cognitive area of the assessment. These test scores are also evaluated with plausible values that I standardize. This model is otherwise identical to the previous tables.
* Statistically significant at the 10-percent level.
**Statistically significant at the 5-percent level.
*** Statistically significant at the 1-percent level.

These results can be compared with those found with standard test scores (in Table 4) to examine the impact of spillover effects. The estimates in Table 7 indicate that a female teacher in math is better for boys. As the skills acquired in math can improve science test scores, we can assume that in Table 4 the gender match coefficient for boys in science is biased by the female math teacher positive spillovers. When focusing only on the knowing cognitive domain, we observe an estimate reporting smaller effects of gender match. This is therefore consistent

[32] In Table A7 from the Appendix, these results are even more noteworthy when using all plausible values, 0.046 SD with the first set of controls and 0.054 with teacher perception controls.

with my interpretation of spillover effects. This logic, however, does not seem to apply to girls. Although Table 4 shows that male math teachers are better for girls, their gender match effects in science increase when positive spillovers are removed.

Still, these interactions remain difficult to interpret as it is also plausible that a transfer of learned abilities is easier from one subject to another. For instance, skills from math could be more transferable to science than vice versa. This could explain why the results are consistent with my interpretation when the gender effects in math are larger.[33]

7 Heterogenous effects

I believe the estimates found with knowing test scores capture the effects of gender match with less bias as they limit spillovers disturbance. Therefore, I will explore heterogenous effects using these achievement outcomes.

The effects observed so far are small and of little significance except for girls' assignment to a female science teacher. I explore now if certain categories of the population experience effects of different intensity and direction. According to the role model hypothesis, it could be assumed that girls from lower socioeconomic backgrounds with mostly poorly educated mothers would be most affected by a female role model teacher. In order to test this hypothesis, I investigate the effects of gender match on students from low and high socioeconomic backgrounds. Since I do not have any variable measuring students' socioeconomic background directly, I use the home resources index calculated with multiple answers to the survey as a proxy. I would therefore expect students with the lower home resources to have stronger positive effects of same-sex teacher assignment. I apply the same methodology but restricting the observations to those with a home resources index below 10.5 and then above 11.5.[34] In Table 8, the results are shown directly with subject specific effects.

[33] The difference of these results compared to Table 3 and 4 might also be caused by different gender match effects between cognitive domains. For example, female teachers could be better at improving students' knowledge for both genders, whereas male teachers would be better at improving reasoning skills. This would explain why, when focusing solely on the knowledge domain, I find more positive gender match effects for girls and more negative effects for boys. When conducting the same analysis for the applying and reasoning domains separately I find no significant effects and the estimates are closely similar for both domains. This seems to confirm that the results do not vary depending on cognitive domains but rather due to spillover effects. Jan Bietenbeck (2014) finds that traditional teaching practices increase students' knowledge domain and applying domain test scores, whereas modern teaching practices improve their reasoning skills. Assuming that male teachers are more likely to employ modern teaching than female teachers (or the other way around), I would expect to observe different estimates between the applying and reasoning domain which is not the case here.

[34] This is equivalent to analyzing the effects on two groups with a little more than 10'000 observations. These groups are not equally sized due the fact that many students share the same home resources index number. The choice of cutoff points is set as close as possible to the 35th percentile and the 65th percentile.

Surprisingly, Table 8 does not indicate that students with lower home resources experience any positive gender match effects. On the contrary, female math teachers seem to significantly decrease disadvantaged girls' test scores by 0.059 SD with the first set of controls and 0.073 when adding teacher self-perception controls. The only exceptions are disadvantaged boys in math where a small and insignificant positive effect remains.

Table 8 : Heterogeneity Analysis For Home Resources, Subject Specific And Knowing Domain Scores (Using PV1)

VARIABLES	Test score difference					
	Female			Male		
			A. Low Home Resources			
Gender match in Math	-0.060*	-0.059*	-0.073**	0.030	0.031	0.039
	(0.032)	(0.032)	(0.036)	(0.037)	(0.038)	(0.034)
Gender match in Science	-0.012	-0.019	-0.020	-0.014	-0.012	-0.020
	(0.031)	(0.032)	(0.033)	(0.038)	(0.037)	(0.036)
Teacher and class controls	No	Yes	Yes	No	Yes	Yes
Teacher perception controls	No	No	Yes	No	No	Yes
P-value (H$_0$: B$_m$ =B$_s$)	0.249	0.333	0.230	0.421	0.428	0.271
Observations	4,907	4,770	4,220	5,336	5,202	4,620
R-squared	0.029	0.033	0.035	0.014	0.018	0.025
			B. High Home Resources			
Gender match in Math	0.024	0.029	0.019	-0.065**	-0.059*	-0.066**
	(0.033)	(0.034)	(0.036)	(0.032)	(0.031)	(0.032)
Gender match in Science	0.080**	0.095***	0.089**	-0.003	-0.009	-0.007
	(0.032)	(0.034)	(0.036)	(0.031)	(0.031)	(0.032)
Teacher/classroom controls	No	Yes	Yes	No	Yes	Yes
Teacher perception controls	No	No	Yes	No	No	Yes
P-value (H$_0$: B$_m$ =B$_s$)	0.1722	0.107	0.106	0.119	0.215	0.161
Observations	5,477	5,354	4,712	5,218	5,086	4,519
R-squared	0.020	0.025	0.034	0.029	0.037	0.039

Note: This model allows subject specific effects and uses the first plausible value of student knowing test scores. Panel A evaluates the estimates for students with low home resources (index lower than 10.5) and panel B for student with high home resources (index higher than 11.5). This model is otherwise identical to the previous tables.
* Statistically significant at the 10-percent level.
** Statistically significant at the 5-percent level.
*** Statistically significant at the 1-percent level.

Significant positive effects of a same-gender teacher are observed on girls with high home resources, with an estimate in the second column of 0.095 SD in science significant at the 1-percent level. Such positive effects do not occur on boys with high home resources, a significant decline in their academic performance can be observed when they are paired with a male math teacher.[35] Moreover, this time there are no significant differences between math and science estimates as the effects across subjects seem to follow the same direction when a significant estimate is found in one of the disciplines. Although this difference is close to being significant for advantaged girls. Overall, female teachers appear to be beneficial for students with a high socioeconomic background while mostly detrimental for low socioeconomic background students.

The decrease in the number of observations on which our results are based makes it more relevant to use all plausible values to provide the most accurate estimates. Table A8 reports the estimates found using the five plausible values. In this table, the previously observed significant effects are slightly weaker, causing the estimates to become insignificant.[36] Only the positive effects of gender match for girls in science remains strongly significant. Moreover, the positive estimate of disadvantaged boys in math disappears, showing even more clearly that there are no positive gender match effects for low home resources students.

This undermines my original hypothesis that girls from disadvantaged backgrounds are more positively influenced by the assignment to a female teacher due to a lack of female role models in their environment. However, it would be possible to imagine an alternative hypothesis consistent with my findings. Perhaps since privileged girls are more likely to pursue long academic careers, the influence of a female role model teacher could help deconstruct the stereotypes of a male dominated field, which would enhance the motivation of those students who see themselves persisting to study that subject. Given the age of students approaching the end of compulsory school, this hypothesis would eventually make just as much sense.[37]

8 Student non-cognitive outcomes

In this study, I extend student outcomes beyond the classical standardized test scores to also include students' attitudes toward the discipline. It is likely that these non-cognitive outcomes are correlated to academic achievements. For example, students who perform well are more

[35] In the heterogeneity analysis, the same effects are captured with standard test scores as dependent variable although these effects are of lower intensity. The results are presented in Table A9 from the Appendix.

[36] This is also partially due to an increase in standard errors caused by the variation in estimates found with all the different plausible values.

[37] Further heterogeneous effects were explored to examine other hypotheses, such as the possible positive effect of a same-gender role model when a student is victim of bullying at school. No significant effects were found.

prone to enjoy the subject. However, a teacher could still greatly inspire his or her students without causing an immediate improvement in their performance. Without examining students' perceptions of the subject these effects would not be captured.

To investigate the possible effects on students' non-cognitive outcomes, I apply the same methodology as in the previous models, but I use students' self-perceptions of subject liking, valuation, and confidence as dependent variables. These variables are an aggregated measure generated by multiple survey items. Only the student characteristics fixed across subjects that affect perceptions in the same way will be removed. One positive feature is that since all students respond to the entire set of questions, I no longer have to rely on plausible values. However, the number of items used to measure these indices is now much smaller than that of the test scores.[38] The variables also depend on students' own feelings toward the subjects, which could potentially lead to measurement errors if not reported representatively.

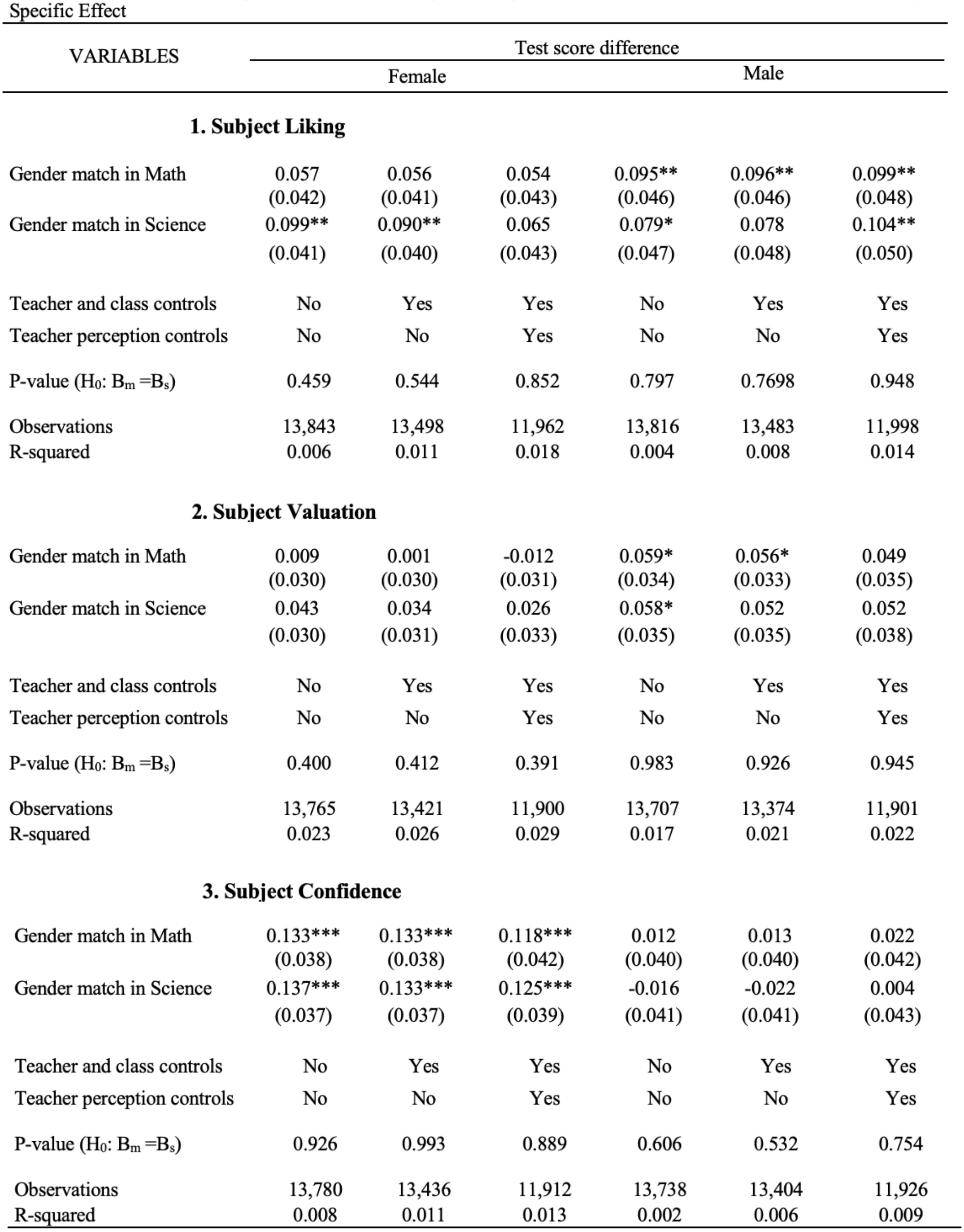
Table 9 presents the estimated effects of gender match on students' perceptions that allow for subject specific effects. The results depicted in this table indicate that the assignment to a same-gender teacher improves students' non-cognitive outcomes as most estimates are positive. The most notable findings are again observed on girls, with the subject confidence index increasing by just over 0.130 SD in both math and science. These estimates are significant at the 1-percent level. For boys the gender match effects on subject confidence are small. However, they do show higher estimates than girls on subject valuation, with coefficients close to 0.050 SD. The magnitude of the estimates evaluating the effects on subject liking is notable for both boys and girls. Girls' science liking and boys' math liking indices increase by around 0.090 SD with gender match.

There are no statistically significant differences between the estimates for math and science, although girls appear to benefit from stronger gender match effects in science, while the gender match effects for boys are stronger in math. This is not surprising for girls considering the positive effects of gender match on science test scores observed in the previous tables. However, this was less expected for boys since they have lower estimates in math than in science when using academic achievement as dependent variable. This might indicate that the effects of same-gender teacher assignment on educational achievement are therefore not solely driven by a change in students' attitudes towards disciplines. These effects are also of greater magnitude than those previously detected on educational achievement which suggest that

[38] The principal purpose of TIMSS study is to asses student achievement in science and mathematics. The indices measuring students' perceptions of the subjects should not be considered as precise as the test scores.

teacher gender has a larger impact on students' non-cognitive outcomes than on their cognitive outcomes.

Table 9 : First Difference Regression Of Student Subject Liking, Valuation And Confidence Coefficient With Subject Specific Effect



VARIABLES	Test score difference					
	Female			Male		
1. Subject Liking						
Gender match in Math	0.057	0.056	0.054	0.095**	0.096**	0.099**
	(0.042)	(0.041)	(0.043)	(0.046)	(0.046)	(0.048)
Gender match in Science	0.099**	0.090**	0.065	0.079*	0.078	0.104**
	(0.041)	(0.040)	(0.043)	(0.047)	(0.048)	(0.050)
Teacher and class controls	No	Yes	Yes	No	Yes	Yes
Teacher perception controls	No	No	Yes	No	No	Yes
P-value (H_0: $B_m = B_s$)	0.459	0.544	0.852	0.797	0.7698	0.948
Observations	13,843	13,498	11,962	13,816	13,483	11,998
R-squared	0.006	0.011	0.018	0.004	0.008	0.014
2. Subject Valuation						
Gender match in Math	0.009	0.001	-0.012	0.059*	0.056*	0.049
	(0.030)	(0.030)	(0.031)	(0.034)	(0.033)	(0.035)
Gender match in Science	0.043	0.034	0.026	0.058*	0.052	0.052
	(0.030)	(0.031)	(0.033)	(0.035)	(0.035)	(0.038)
Teacher and class controls	No	Yes	Yes	No	Yes	Yes
Teacher perception controls	No	No	Yes	No	No	Yes
P-value (H_0: $B_m = B_s$)	0.400	0.412	0.391	0.983	0.926	0.945
Observations	13,765	13,421	11,900	13,707	13,374	11,901
R-squared	0.023	0.026	0.029	0.017	0.021	0.022
3. Subject Confidence						
Gender match in Math	0.133***	0.133***	0.118***	0.012	0.013	0.022
	(0.038)	(0.038)	(0.042)	(0.040)	(0.040)	(0.042)
Gender match in Science	0.137***	0.133***	0.125***	-0.016	-0.022	0.004
	(0.037)	(0.037)	(0.039)	(0.041)	(0.041)	(0.043)
Teacher and class controls	No	Yes	Yes	No	Yes	Yes
Teacher perception controls	No	No	Yes	No	No	Yes
P-value (H_0: $B_m = B_s$)	0.926	0.993	0.889	0.606	0.532	0.754
Observations	13,780	13,436	11,912	13,738	13,404	11,926
R-squared	0.008	0.011	0.013	0.002	0.006	0.009

Note: This model uses student perceptions of math and science produced by multiple answers to the survey items. The three indices for students liking, valuation, and confidence are standardized and used as dependent variables. This model is otherwise identical to the previous tables.

* Statistically significant at the 10-percent level.
** Statistically significant at the 5-percent level.
*** Statistically significant at the 1-percent level.

I also conduct a heterogeneity analysis for home resources to compare to the results observed in Table 8 with academic achievement. Table 10 presents the estimates found with a model allowing for subject specific effects that I estimate separately for students with lower and higher home resources.

It is noteworthy that disadvantaged girls have larger gender match effects than advantaged girls in science for subject liking, value, and confidence. In particular, they experience the most notable effect with the subject confidence index increasing significantly by 0.151 SD in science when using controls. This may seem rather curious when reconsidering my findings in Table 8. In that table, the significant positive impact of female teacher on science test scores is detected on girls with higher home resources, while negative estimates are observed on girls with lower home resources. This confirms that the effects of same-gender teacher assignment on educational performances are therefore not entirely driven by a change in students' attitudes towards disciplines.

For boys, students from higher socioeconomic backgrounds appear to have stronger gender match effects in science than math, while the opposite is true for students from lower socioeconomic backgrounds. However, this does not explain the significant negative coefficient observed in math for advantaged boys in Table 8. Especially since the estimates with students' perceptions remain mostly positive in this subject. Therefore, the results for boys are also inconsistent with the hypothesis that academic achievement is driven by non-cognitive outcomes.

Interestingly, in this table the controls appear to have a greater influence on the coefficients than in previous tables. However, this seems more likely due to the number of observations eliminated because of variables' missing values. This was also occurring before, but here the observations are already strongly reduced at the time of the analysis due to the exploration of heterogenous effects.

It is worth noting that the results reported with this model could be biased by a correlation with the gender match variable and teacher and classroom characteristics, although I try to mitigate these effects with controls. Moreover, there could be possible spillovers, where for example students would start liking both subjects because of one teacher.

However, these results do offer important insights into the possible effects that occur in the classroom. Although I do not have information on long-term impacts, it is quite possible to imagine that the observed patterns could further influence students' achievements and choices of career in the future (e.g., girls' aspirations in STEM).

Table 10 : First Difference Regression Of Student Subject Liking, Valuation And Confidence Coefficient With Subject Specific Effect And Heterogeneity Analysis For Home Resources

VARIABLES	Test score difference					
	Female			Male		
1. Subject Liking						
			A. Low Home Resources			
Gender match in Math	0.018 (0.059)	0.022 (0.058)	0.010 (0.061)	0.070 (0.062)	0.071 (0.062)	0.096 (0.065)
Gender match in Science	0.138** (0.058)	0.139** (0.056)	0.120** (0.060)	-0.063 (0.062)	-0.056 (0.066)	-0.049 (0.070)
Teacher and class controls	No	Yes	Yes	No	Yes	Yes
Teacher perception controls	No	No	Yes	No	No	Yes
P-value (H_0: B_m =B_s)	0.119	0.115	0.175	0.111	0.153	0.124
Observations	4,839	4,703	4,158	5,219	5,093	4,523
R-squared	0.013	0.022	0.029	0.007	0.010	0.013
			B. High Home Resources			
Gender match in Math	0.095 (0.063)	0.111* (0.060)	0.123** (0.062)	0.086 (0.069)	0.103 (0.063)	0.067 (0.066)
Gender match in Science	0.113* (0.062)	0.107* (0.058)	0.073 (0.060)	0.128* (0.068)	0.115* (0.065)	0.155** (0.068)
Teacher/classroom controls	No	Yes	Yes	No	Yes	Yes
Teacher perception controls	No	No	Yes	No	No	Yes
P-value (H_0: B_m =B_s)	0.836	0.961	0.575	0.647	0.889	0.344
Observations	4,531	4,434	3,931	4,363	4,252	3,797
R-squared	0.004	0.017	0.033	0.004	0.019	0.031
2. Subject Valuation						
			A. Low Home Resources			
Gender match in Math	0.035 (0.048)	0.022 (0.049)	0.001 (0.052)	0.065 (0.051)	0.062 (0.051)	0.040 (0.052)
Gender match in Science	0.086* (0.048)	0.083* (0.049)	0.054 (0.052)	0.024 (0.051)	0.015 (0.054)	0.013 (0.056)
Teacher and class controls	No	Yes	Yes	No	Yes	Yes
Teacher perception controls	No	No	Yes	No	No	Yes
P-value (H_0: B_m =B_s)	0.418	0.341	0.451	0.542	0.509	0.713
Observations	4,799	4,664	4,124	5,168	5,039	4,477
R-squared	0.028	0.033	0.040	0.029	0.032	0.034

B. High Home Resources

Gender match in Math	-0.037	-0.058	-0.064	-0.004	0.015	0.002
	(0.047)	(0.044)	(0.044)	(0.053)	(0.049)	(0.052)
Gender match in Science	0.017	-0.008	-0.004	0.082	0.063	0.070
	(0.046)	(0.046)	(0.046)	(0.053)	(0.050)	(0.054)
Teacher/classroom controls	No	Yes	Yes	No	Yes	Yes
Teacher perception controls	No	No	Yes	No	No	Yes
P-value (H_0: B_m =B_s)	0.389	0.390	0.324	0.258	0.499	0.376
Observations	4,508	4,410	3,915	4,333	4,225	3,775
R-squared	0.001	0.036	0.047	0.001	0.017	0.024

3. Subject Confidence

A. Low Home Resources

Gender match in Math	0.126*	0.118*	0.081	0.018	0.015	0.034
	(0.065)	(0.062)	(0.065)	(0.057)	(0.056)	(0.058)
Gender match in Science	0.155**	0.151**	0.151**	-0.089	-0.085	-0.070
	(0.063)	(0.060)	(0.064)	(0.057)	(0.058)	(0.061)
Teacher and class controls	No	Yes	Yes	No	Yes	Yes
Teacher perception controls	No	No	Yes	No	No	Yes
P-value (H_0: B_m =B_s)	0.692	0.662	0.379	0.211	0.254	0.239
Observations	4,809	4,674	4,133	5,178	5,049	4,486
R-squared	0.007	0.016	0.022	0.007	0.015	0.017

B. High Home Resources

Gender match in Math	0.126**	0.142***	0.133**	0.016	0.022	0.004
	(0.054)	(0.054)	(0.056)	(0.066)	(0.065)	(0.066)
Gender match in Science	0.109**	0.105**	0.086*	0.019	0.013	0.056
	(0.053)	(0.050)	(0.052)	(0.064)	(0.066)	(0.069)
Teacher/classroom controls	No	Yes	Yes	No	Yes	Yes
Teacher perception controls	No	No	Yes	No	No	Yes
P-value (H_0: B_m =B_s)	0.818	0.610	0.533	0.969	0.917	0.551
Observations	4,516	4,418	3,921	4,338	4,228	3,778
R-squared	0.005	0.015	0.019	0.000	0.009	0.016

Note: This model uses student perceptions of math and science produced by multiple answers to the survey items. The three indices for students liking, valuation, and confidence are standardized and used as dependent variables. The estimations are separated for students with low home resources (index lower than 10.5) and student with high home resources (index higher than 11.5). This model is otherwise identical to the previous tables.

* Statistically significant at the 10-percent level.
** Statistically significant at the 5-percent level.
*** Statistically significant at the 1-percent level.

9 Conclusion

In this paper, I examine the effects of same-sex teacher assignment on students' academic performance and non-cognitive outcomes in Australia and New Zealand using TIMSS data. I find that girls' knowing test scores increase significantly by around 0.040 SD when assigned to a female teacher. This effect is less pronounced when standard test scores are used as dependent variable. However, these outcomes are more likely to cause estimates to be biased due to spillovers, and thus misidentify the true causal effect.

Compared to previous findings, the economic magnitude of this effect remains rather moderate. For example, Dee (2007) finds significant gender match effects on test scores that are nearly twice as large. Moreover, in his paper, significant effects are noted for both boys and girls, as well as in different subjects, whereas all of my estimates, with the exception of girls in science, are small and insignificant.

The gender match effects I find on students' subject perceptions are larger. These effects are mostly positive and observed for both boys and girls, although the most notable effects are again observed on girls, with subject confidence increasing by just over 0.130 SD in both classes. The magnitude of these effects is more in line with previous research finding a significant impact of same-gender teacher assignment. This suggests that a same-sex teacher could more easily influence their students' attitudes toward the discipline than their academic performance.

Overall, the positive impacts of female teachers on girls' cognitive and non-cognitive outcomes are consistent with previous literature, as positive effects tend to be reported for girls. This may indicate that having a same-gender teacher role model is relatively more important for girls.

The heterogeneity analysis results suggest that students with lower socioeconomic background do not experience positive gender match effects on their test scores for both genders, despite showing mostly positive effects on their subject enjoyment, valuation, and confidence. In particular, it is noteworthy that in science, disadvantaged girls have larger gender match effects on subject perceptions than advantaged girls, although the previously mentioned significant positive effect of female teacher on girls' test scores is driven by high home resources girls. These results are therefore inconsistent with the hypothesis that the effects of same-gender teacher assignment on academic achievement are caused mainly by a change in students' attitudes toward the disciplines.

This contrasts with previous studies that generally find similar gender match effects on both cognitive and non-cognitive outcomes. For instance, Gong, Lun, and Song (2018) find that a same-sex teacher increases girls' test scores by nearly 0.140 SD, while improving mental stress with a similar magnitude.

These results are found using a fixed effects model depending on the following identifying assumption. Unobservable student specific characteristics are identical for math and science. Considering the similarities between the subjects, this could be regarded as conceivable. However, my findings could still be biased due to the teacher and classroom characteristics correlation with teacher's gender. I try to limit these effects by including some controls in my first difference regressions, which does not eliminate the possible bias caused by unobservables.

Policies to increase the proportion of male teachers have been justified on the grounds that boys may suffer from a lack of male role models in school. My results provide little support for this theory as boys do not experience any strong and significant positive effects of gender match on academic achievement and that the effects on students' subject perceptions are also observed predominantly on girls.

Therefore, this argument does not appear to be a legitimate reason for recruiting more male teachers in Australia and New Zealand. Such recruitment policy would also most likely result in a decrease in teacher quality due to the shortage of qualified male teachers. However, other reasons can be suggested. For instance, gender parity in the teaching profession could help students deconstruct gender-based stereotypes, which could have a long-lasting positive impact on society.

It is worth recalling that this paper investigates the effects of student-teacher gender match in math and science. In these subjects, there is a gender gap in test scores that favors boys. The gender interactions between students and their teachers may vary depending on the subjects. Further research could examine the effects of gender match in subjects in which girls outperform boys (e.g., reading and language subjects) to provide more clarity on these gender interactions in schools.

Moreover, my study is limited to Australia and New Zealand while the student-teacher gender interactions may differ by country. For instance, Alfa and Hermann (2017) and Cho (2012) conduct their analysis on multiple countries and find no universal teacher–student gender match effect. In addition, I only observe 8th grade students, meaning that my research focuses on the impacts of teacher gender in the lower secondary level. It would therefore be wise not to over-generalize these results.

Nevertheless, this study provides valuable insight into the influence of teacher gender on students’ outcomes. Although my estimation only shows the effects on students over the course of one year, it is reasonable to assume that these impacts would not fade over time, resulting in long-term effects. The inclusion of longitudinal data in the TIMSS dataset in the future would allow to estimate such effects. The survey could also include more gender identification options to ensure the accurate collection of all students' and teachers’ gender identity. We might then observe estimates of greater economic magnitude when students are paired with a teacher who has the same gender self-identification.

References

Almquist, Elizabeth M. and Shirley S. Angrist. "Role Model Influences on College Women's Career Aspirations." *Merrill-palmer Quarterly* (1971): n. pag.

Alfa, Diallo and Zoltán Hermann. "Does teacher gender matter in Europe? Evidence from TIMSS data = Számít a tanár neme Európában? Eredmények a TIMSS adatok alapján." (2017).

Ammermüller, Andreas and Peter J. Dolton. "Pupil-Teacher Gender Interaction Effects on Scholastic Outcomes in England and the Usa." (2006).

Antecol, Heather, Ozkan Eren and Serkan Ozbeklik. "The Effect of Teacher Gender on Student Achievement in Primary School." *Journal of Labor Economics* 33 (2014): 63 - 89.

Aparicio, Juan, José Manuel Cordero and Lidia Ortiz. "Efficiency Analysis with Educational Data: How to Deal with Plausible Values from International Large-Scale Assessments." *Mathematics* (2021): n. pag.

Ashenfelter, Orley, and Alan Krueger. "Estimates of the Economic Return to Schooling from a New Sample of Twins." *The American Economic Review* 84, no. 5 (1994): 1157–73.

Ashenfelter, Orley C. and Cecilia Elena Rouse. "Income, Schooling, and Ability: Evidence from a New Sample of Identical Twins." *Labor: Human Capital* (1997): n. pag.

Baker, David P. and Deborah Jones. "Creating Gender Equality: Cross-National Gender Stratification and Mathematical Performance." *Sociology Of Education* 66 (1993): 91-103.

Basow, Susan A. and Karen G. Howe. "Role-Model Influence: Effects of Sex and Sex-Role Attitude in College Students." *Psychology of Women Quarterly* 4 (1980): 558 - 572.

Bettinger, Eric and Bridget Terry Long. "Do Faculty Serve as Role Models? The Impact of Instructor Gender on Female Students." *The American Economic Review* 95 (2005): 152-157.

Bibby, Yan. Plausible Values: How Many for Plausible Results. Diss. University of Melbourne, 2020.

Bietenbeck, Jan. "Teaching practices and cognitive skills." *Labour Economics* 30 (2014): 143-153.

Boston College. "TIMSS & PIRLS International Study Center. Methods and Procedures in TIMSS 2015." Available online: https://timssandpirls.bc.edu/publications/timss/2015-methods.html (accessed on 14th May 2022).

Boston College. "TIMSS & PIRLS International Study Center. TIMSS 2015 Assessment Frameworks." Available online: https://timssandpirls.bc.edu/timss2015/frameworks.html (accessed on 14th May 2022).

Boston College. “TIMSS & PIRLS International Study Center. TIMSS 2015 Fourth Grade Almanacs and TIMSS 2015 Eighth Grade Almanacs.” Available online: https://timssandpirls.bc.edu/timss2015/international-database/index.html (accessed on 14th May 2022).

Carrell, Scott E., Marianne E. Page and James E. West. “Sex and Science: How Professor Gender Perpetuates the Gender Gap.” *NBER Working Paper Series* (2009): n. pag.

Cho, Insook. “The effect of teacher–student gender matching: Evidence from OECD countries.” *Economics of Education Review* 31 (2012): 54-67.

Dee, Thomas S.. “A Teacher Like Me: Does Race, Ethnicity, or Gender Matter?” *The American Economic Review* 95 (2005): 158-165.

Dee, Thomas S.. “Teachers and the Gender Gaps in Student Achievement.” *The Journal of Human Resources* XLII (2007): 528 - 554.

Education Queensland. “Male teachers’ strategy : strategic plan for the attraction, recruitment and retention of male teachers in Queensland state schools 2002-2005.” (2002). Available online: https://www.bgnelson.com/workshops/queensland-strategy.pdf (accessed on 14th May 2022).

Ehrenberg, Ronald G., Dan Goldhaber and Dominic J. Brewer. “Do Teachers’ Race, Gender, and Ethnicity Matter? Evidence from the National Educational Longitudinal Study of 1988.” *Industrial & Labor Relations Review* 48 (1995): 547 - 561.

Gardner, Howard and Thomas Hatch. “Multiple Intelligences Go to School: Educational Implications of the Theory of Multiple Intelligences. Technical Report No. 4.” (1989).

Gong, Jie, Yi Lu and Hong Song. “The Effect of Teacher Gender on Students’ Academic and Noncognitive Outcomes.” *Journal of Labor Economics* 36 (2018): 743 - 778.

González-Pérez, Susana, Ruth Mateos de Cabo and Milagros Sáinz. “Girls in STEM: Is It a Female Role-Model Thing?” *Frontiers in Psychology* 11 (2020): n. pag.

Griliches, Zvi and Jerry Hausman. “Errors in Variables in Panel Data.” *Econometrics: Econometric & Statistical Methods - Special Topics eJournal* (1984): n. pag.

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano J., M. Lagos, P. Norris, E. Ponarin & B. Puranen. “World Values Survey: Round Seven - Country-Pooled Datafile Version 4.0.” *Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat* (2022).

Hanushek, Eric Alan. “Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro Data.” *The American Economic Review* 61 (1971): 280-288.

Hedges, Larry V and Amy Nowell. “Sex differences in mental test scores, variability, and numbers of high-scoring individuals.” *Science* 269 5220 (1995): 41-5.

Herold, Benjamin. "Students Embrace a Wide Range of Gender Identities. Most School Data Systems Don't." *Education week* (2022). Available online: https://www.edweek.org/leadership/students-are-embracing-a-wide-range-of-gender-identities-most-school-data-systems-dont/2022/01 (accessed on 14th May 2022).

Hoffmann, Florian and Philip Oreopoulos. "A Professor Like Me: The Influence of Instructor Gender on College Achievement." *Journal of Human Resources* 44 (2007): 479 - 494.

Holmlund, Helena and Krister Sund. "Is the Gender Gap in School Performance Affected by the Sex of the Teacher." *Labour Economics* 15 (2008): 37-53.

Laveist, Thomas A. and Amani M. Nuru-Jeter. "Is doctor-patient race concordance associated with greater satisfaction with care?" *Journal of health and social behavior* 43 3 (2002): 296-306 .

Lim, Jaegeum and Jonathan Meer. "Persistent Effects of Teacher–Student Gender Matches." *The Journal of Human Resources* 55 (2020): 809 - 835.

Ma, Xin. "Within-School Gender Gaps in Reading, Mathematics, and Science Literacy." *Comparative Education Review* 52 (2008): 437 - 460.

Macdonald, Kevin. "PV: Stata module to perform estimation with plausible values." *Statistical Software Components* (2008): n. pag.

Martin, Carol Lynn and Diane N. Ruble. "Children's Search for Gender Cues." *Current Directions in Psychological Science* 13 (2004): 67 - 70.

Meinck, Sabine and Falk Brese. "Trends in gender gaps: using 20 years of evidence from TIMSS." *Large-scale Assessments in Education* 7 (2019): 1-23.

Neugebauer, Martin, Marcel Helbig and Andreas Landmann. "Unmasking the Myth of the Same-Sex Teacher Advantage." *European Sociological Review* 27 (2011): 669-689.

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

Rosenthal, Robert W. and Lenore Jacobson. "Pygmalion in the Classroom: Teacher Expectation and Pupils' Intellectual Development." (1968).

Rouse, Cecilia Elena. "Further Estimates of the Economic Return to Schooling from a New Sample of Twins." *Economics of Education Review* 18 (1999): 149-157.

Rubin, Donald B.. "Multiple imputation for nonresponse in surveys." (1987).

Sadker, Myra and David Miller Sadker. "Failing at fairness : how our schools cheat girls." (1995).

Spencer, Steven John, Claude M. Steele and Diane M. Quinn. "Stereotype Threat and Women's Math Performance." *Journal of Experimental Social Psychology* 35 (1999): 4-28.

Sommers, Christina Hoff. “The War Against Boys: How Misguided Feminism Is Harming Our Young Men.” (2000).

Steele, Claude M.. “A threat in the air. How stereotypes shape intellectual identity and performance.” *The American psychologist* 52 6 (1997): 613-29 .

Strumpf, Erin. “Racial/Ethnic Disparities in Primary Care: The Role of Physician-Patient Concordance.” *Medical Care* 49 (2011): 496–503.

Ujifusa, Andrew. “ Schools Could Count Nonbinary Students Under Biden Proposal.” *Education week* (2021). Available online: https://www.edweek.org/policy-politics/schools-could-count-nonbinary-students-in-biden-proposal/2021/11 (accessed on 14th May 2022).

Von Davier, Matthias, Eugenio Gonzalez, and Robert Mislevy. "What are plausible values and why are they useful." *IERI monograph series* 2.1 (2009): 9-36.

Appendix

Table A1 : Previous research summary

Paper	Data	Method	Results
Lim and Meer (2020)	Secondary school level: Seoul Education Longitudinal Study of 2010 (SELS2010)	Linear regression with students and teachers randomly assigned in classrooms	Positive effects of female teachers on female students' standardized test scores (even 5 years later) and their aspiration to a STEM degree.
Carrell, Page and West (2009)	College level: U.S. Air Force Academy (USAFA) 2007	OLS with random assignment	Positive effects of gender interaction for females. Strong effects on their math and science performance and on their decision to pursue courses in those fields as well as their aspiration to a STEM degree.
Antecol, Eren and Ozbeklik (2014)	Primary school level: The Mathematica Policy Research, Incorporated (MPR), National Evaluation of Teach for America (NETFA) 2001-2003	OLS with random assignment	Negative effects of female teachers on female students in disadvantaged neighborhoods' math test scores.
Gong, Lu and Song (2018)	Secondary school level: 2014 China Education Panel Survey (CEPS)	OLS with random assignment	Positive effects for girls that have a female teacher on their academic achievement and non-cognitive outcomes (e.g., mental stress, school satisfaction).
Bettinger and Long (2005)	College level: 12 public four-year colleges in Ohio 1998-1999	Instrumental variable method	Positive effects of same-gender faculty members on students' interest on a field and choice of major.
Ehrenberg, Goldhaber and Brewer (1995)	Secondary school level: National Education Longitudinal Study of 1988 (NELS:88)	Linear regression with controls	No effects of gender interaction on students' test scores, but a positive impact on teachers' subjective evaluations.
Dee (2007)	Secondary school level: National Education Longitudinal Study of 1988 (NELS:88)	Student Fixed effects model (first difference across two academic subjects) to differences out unobservable student traits	Positive effects of same-gender teacher for both boys' and girls' academic achievement and teachers' perception of the student.
Ammermueller and Dolton 2006	Primary and secondary school level: PIRLS 2001 and TIMMS data from 1995, 1999, 2003	Similar Fixed effects method	Positive effects of student-teacher gender interaction in England for math but not in the United States.
Hoffmann and Oreopoulos (2007)	College level: University of Toronto's Arts and Science Faculty 1996 to 2005	Similar Fixed effects method (also obtaining results with within instructor variation)	Little or no effects of gender interactions on students' academic achievement and choice of courses. Those minor effects still indicate that gender match matter for some students.

Holmlund and Sund (2008)	Secondary school level: Municipality of Stockholm 2002-2003	Similar fixed effects method (using measures of the student two years apart instead than with two academic subjects the same year)	No effects of same-sex teacher on students' outcomes are detected.
Neugebauer, Helbig and Landmann (2011)	Primary school level: Large-scale data from IGLU-E in 2001, an expansion of PIRLS in Germany (with more precision on the assignment of teachers and their teaching durations)	Similar Fixed effects method	No effects of same-sex teacher on students' objective test scores and subjective teacher's grades for both boys and girls.
Cho (2012)	Secondary school level: Fifteen OECD countries using data from TIMSS in 1995, 1999, 2003 and 2007	Similar Fixed effects method (using different subject pairs with TIMSS data, here math and science are compared instead of English/history with math/science)	No effects of gender matching on students' achievement in eight countries, positive impact for boys in four countries and for girls in three countries. It is also found that those positive effects may be driven by differences in teacher quality.
Paredes (2014)	Secondary school level: Chile Education Quality Measurement System (SIMCE) 1998	Similar Fixed effects method (adding a theoretical framework to explore whether the positive effects is caused by role model or teacher bias effect)	Positive effects of gender matching on girl's test scores. The paper also provides some evidence that it is due to role model effects and not teacher bias effects
Alfa and Hermann (2017)	Secondary school level: Twenty European countries using data from TIMSS in 2003, 2007 and 2011	Similar Fixed effects method (also adding interaction terms and detecting specific class level with hours spent teaching)	Positive effects of gender interaction on student's test scores in only half of the countries. These effects are more observed with girls and in Western Europe countries. It is also found that the female teacher effects may be driven by selection into the teaching profession.

Note: This table present the data used in the previous research with indication of the students' educational level, the methods employed and their various results.

Table A2: Descriptive statistics

Variables	Definition	Mean	SD	Min/Max
Student characteristics:				
# of students		28,045		
studsex	1 if student gender is female, 0 otherwise	0.498	0.500	0/1
studage	Student age with 2 decimals precision	14.027	0.394	10.42 /17.833
studborn	1 if student born in country, 0 otherwise	0.843	0.363	0/1
motherborn	1 if mother born in country, 0 otherwise	0.700	0.458	0/1
fatherborn	1 if father born in country, 0 otherwise	0.668	0.471	0/1
Parents valuation of Mathematics				
Parentsmatvalue1	1 if parents agree a lot, 0 otherwise	0.633	0.482	0/1
Parentsmatvalue2	1 if parents agree a little, 0 otherwise	0.292	0.455	0/1
Parentsmatvalue3	1 if parents disagree a little, 0 otherwise	0.058	0.234	0/1
Parentsmatvalue4	1 if parents disagree a lot, 0 otherwise	0.017	0.129	0/1
Parents valuation of Science				
Parentsscitvalue1	1 if parents agree a lot, 0 otherwise	0.413	0.492	0/1
Parentsscivalue2	1 if parents agree a little, 0 otherwise	0.378	0.485	0/1
Parentsscivalue3	1 if parents disagree a little, 0 otherwise	0.159	0.366	0/1
Parentsscivalue4	1 if parents disagree a lot, 0 otherwise	0.050	0.218	0/1
home_ressources_coeff	Index increasing with student home resources	11.105	1.580	4.323/14.018
bullying_coeff	Index decreasing with student bullying experienced	9.626	1.884	1.953/13.040
mat_liking_coeff	Index increasing with student Math liking	9.484	1.851	4.968/13.978
mat_confidence_coeff	Index increasing with student confidence in Math	10.077	2.061	3.178/15.925
mat_value_coeff	Index increasing with student valuation of Math	9.865	1.906	2.999/13.707
sci_liking_coeff	Index increasing with student Science liking	9.643	2.032	3.771/13.621
sci_confidence_coeff	Index increasing with student confidence in Science	9.737	1.950	2.821/15.296
sci_value_coeff	Index increasing with student valuation of Science	9.549	1.946	4.136/13.158
Teacher characteristics:				
# of teachers		3256		
Mathematics				
mat_yearsteach	Number of years teaching	16.093	11.630	0/53
mat_sex	1 if teacher gender is female, 0 otherwise	0.524	0.499	0/1
mat_FF	1 if female student and teacher , 0 otherwise	0.257	0.437	0/1
mat_MM	1 if male student and teacher , 0 otherwise	0.234	0.424	0/1
Teacher's age				
mat_age1	1 if teacher age is under 25, 0 otherwise	0.030	0.171	0/1
mat_age2	1 if teacher age is between 25-29, 0 otherwise	0.115	0.319	0/1
mat_age3	1 if teacher age is between 30-39, 0 otherwise	0.233	0.423	0/1
mat_age4	1 if teacher age is between 40–49, 0 otherwise	0.248	0.432	0/1
mat_age5	1 if teacher age is between 50–59, 0 otherwise	0.281	0.450	0/1
mat_age6	1 if teacher age is 60 or more, 0 otherwise	0.093	0.290	0/1
Teacher's formal education				
mat_educ1	1 if some tertiary completed, 0 otherwise	0.041	0.199	0/1
mat_educ2	1 if Bachelor completed, 0 otherwise	0.616	0.486	0/1
mat_educ3	1 if Master completed, 0 otherwise	0.328	0.469	0/1
mat_educ4	1 if Doctor completed, 0 otherwise	0.015	0.123	0/1
Disruptive students in class				
mat_disruptive1	1 if teacher considers not at all, 0 otherwise	0.271	0.444	0/1

mat_disruptive2	1 if teacher considers some, 0 otherwise	0.560	0.496	0/1
mat_disruptive3	1 if teacher considers a lot, 0 otherwise	0.170	0.375	0/1
Uninterested students in class				
mat_uninterested1	1 if teacher considers not at all, 0 otherwise	0.213	0.410	0/1
mat_uninterested2	1 if teacher considers some, 0 otherwise	0.654	0.476	0/1
mat_uninterested3	1 if teacher considers a lot, 0 otherwise	0.132	0.339	0/1
mat_emphasucc_coeff	Index increasing with teacher perception of school emphasis on academic success	10.098	2.129	0.626/18.095
mat_safe_coeff	Index increasing with teacher perception of school safety and order	10.525	2.302	4.214/14.062
mat_satisfac_coeff	Index increasing with teacher job satisfaction	9.726	2.111	1.819/13.802
Science				
sci_yearsteach	Number of years teaching	13.801	10.819	0/48
sci_sex	1 if teacher gender is female, 0 otherwise	0.549	0.498	0/1
sci_FF	1 if female student and teacher , 0 otherwise	0.269	0.444	0/1
sci_MM	1 if male student and teacher , 0 otherwise	0.221	0.415	0/1
Teacher's age				
sci_age1	1 if teacher age is under 25, 0 otherwise	0.037	0.188	0/1
sci_age2	1 if teacher age is between 25-29, 0 otherwise	0.136	0.343	0/1
sci_age3	1 if teacher age is between 30-39, 0 otherwise	0.256	0.437	0/1
sci_age4	1 if teacher age is between 40–49, 0 otherwise	0.271	0.444	0/1
sci_age5	1 if teacher age is between 50–59, 0 otherwise	0.220	0.414	0/1
sci_age6	1 if teacher age is 60 or more, 0 otherwise	0.080	0.271	0/1
Teacher's formal education				
sci_educ1	1 if some tertiary completed, 0 otherwise	0.013	0.112	0/1
sci_educ2	1 if Bachelor completed, 0 otherwise	0.527	0.499	0/1
sci_educ3	1 if Master completed, 0 otherwise	0.426	0.495	0/1
sci_educ4	1 if Doctor completed, 0 otherwise	0.034	0.181	0/1
Disruptive students in class				
sci_disruptive1	1 if teacher considers not at all, 0 otherwise	0.249	0.432	0/1
sci_disruptive2	1 if teacher considers some, 0 otherwise	0.558	0.497	0/1
sci_disruptive3	1 if teacher considers a lot, 0 otherwise	0.193	0.395	0/1
Uninterested students in class				
sci_uninterested1	1 if teacher considers not at all, 0 otherwise	0.199	0.399	0/1
sci_uninterested2	1 if teacher considers some, 0 otherwise	0.664	0.472	0/1
sci_uninterested3	1 if teacher considers a lot, 0 otherwise	0.137	0.344	0/1
sci_emphasucc_coeff	Index increasing with teacher perception of school emphasis on academic success	10.013	2.067	3.305/18.095
sci_safe_coeff	Index increasing with teacher perception of school safety and order	10.309	2.273	4.214/14.062
sci_satisfac_coeff	Index increasing with teacher job satisfaction	9.634	2.065	1.819/13.802
Classroom characteristics:				
mat_csize	Number of students in Math class	22.262	6.152	6/59
sci_csize	Number of students in Science class	21.083	6.910	6/61
mat_cfemaleratio	Proportion of girls in Math class	0.497	0.208	0/1
sci_cfemaleratio	Proportion of girls in Science class	0.497	0.211	0/1
mat_cbornratio	Proportion of students born in country in Math class	0.843	0.131	0/1
sci_cbornration	Proportion of students born in country in Science class	0.843	0.133	0/1

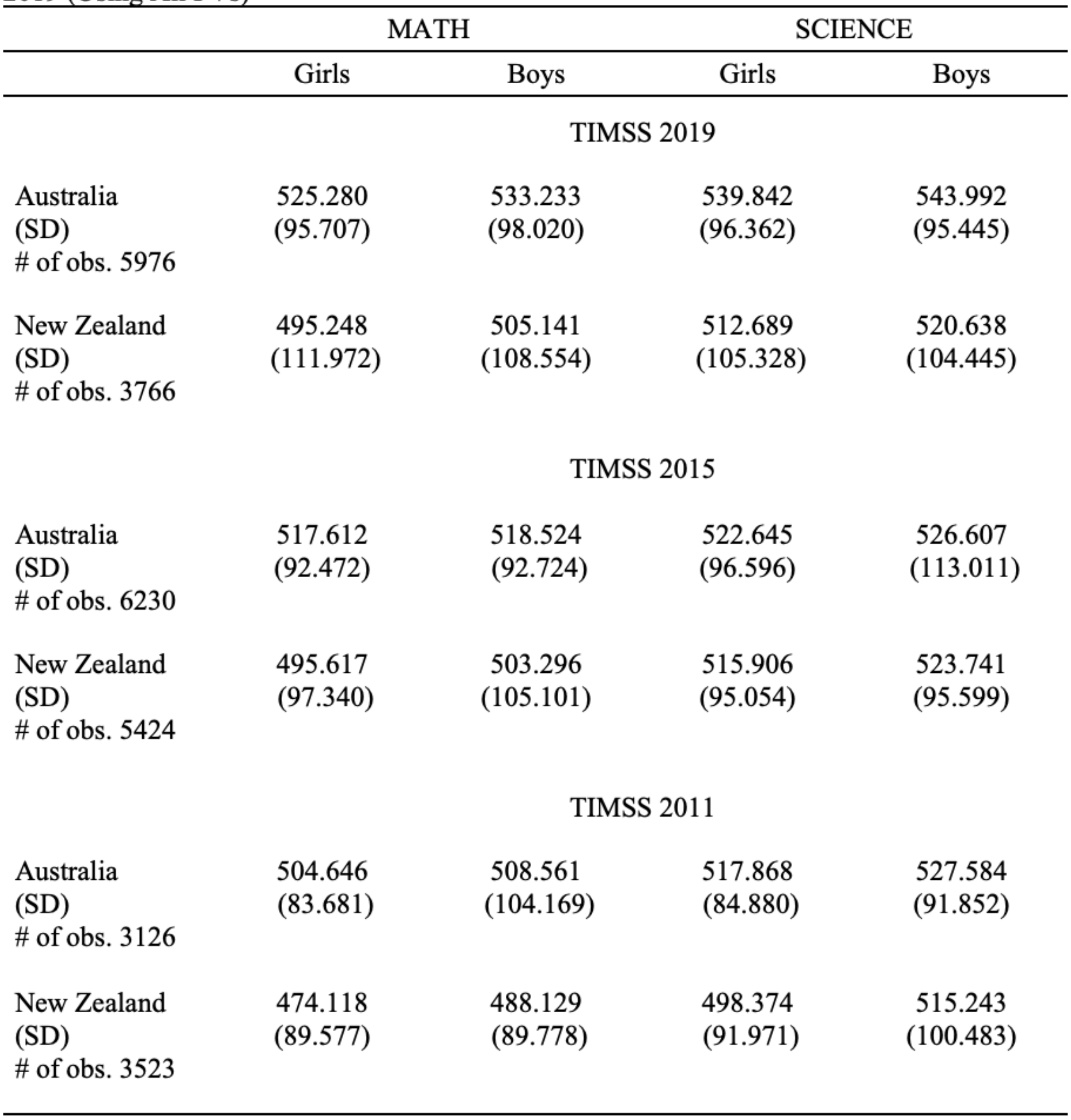
Note: those summary statistics are calculated based on the remaining observations (28045) that meet the restrictions mentioned in the data section.

Table A3: Teacher Characteristics By Gender And Subject

	MATH		SCIENCE	
	Female	Male	Female	Male
Years teaching	16.116 (11.553)	16.765 (11.931)	13.420 (10.284)	14.640 (11.340)
Age (dummy = 1 if at least 50 years old)	0.386 (0.487)	0.386 (0.487)	0.288 (0.453)	0.326 (0.469)
Education (dummy = 1 if at least a Master degree)	0.355 (0.479)	0.340 (0.474)	0.477 (0.499)	0.470 (0.499)
Disruptive (dummy = 1 if some or a lot in class)	0.714 (0.452)	0.724 (0.447)	0.745 (0.436)	0.743 (0.437)
Uninterested (dummy = 1 if some or a lot in class)	0.774 (0.418)	0.780 (0.414)	0.794 (0.404)	0.798 (0.401)
Emphasis on success	10.183 (2.107)	10.162 (2.121)	10.049 (2.035)	10.016 (2.071)
Safety and order	10.543 (2.270)	10.643 (2.304)	10.316 (2.300)	10.401 (2.239)
Job satisfaction	9.796 (2.088)	9.736 (2.125)	9.711 (2.065)	9.521 (2.043)

Note: The variables Age, Education, Disruptive and Uninterested were all transformed into a dummy variable taking the value 1 if the condition mentioned is respected and 0 otherwise.

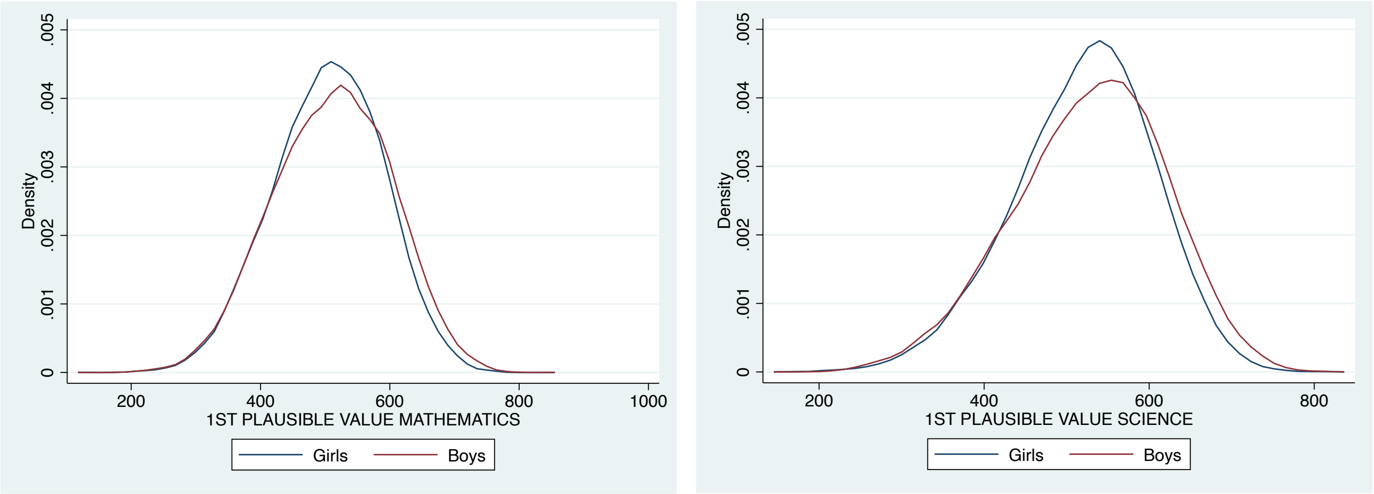
Table A4: Average Scores By Country, Subject And Gender For TIMSS 2011, 2015 And 2019 (Using All PVs)



	MATH		SCIENCE	
	Girls	Boys	Girls	Boys
		TIMSS 2019		
Australia	525.280	533.233	539.842	543.992
(SD)	(95.707)	(98.020)	(96.362)	(95.445)
# of obs. 5976				
New Zealand	495.248	505.141	512.689	520.638
(SD)	(111.972)	(108.554)	(105.328)	(104.445)
# of obs. 3766				
		TIMSS 2015		
Australia	517.612	518.524	522.645	526.607
(SD)	(92.472)	(92.724)	(96.596)	(113.011)
# of obs. 6230				
New Zealand	495.617	503.296	515.906	523.741
(SD)	(97.340)	(105.101)	(95.054)	(95.599)
# of obs. 5424				
		TIMSS 2011		
Australia	504.646	508.561	517.868	527.584
(SD)	(83.681)	(104.169)	(84.880)	(91.852)
# of obs. 3126				
New Zealand	474.118	488.129	498.374	515.243
(SD)	(89.577)	(89.778)	(91.971)	(100.483)
# of obs. 3523				

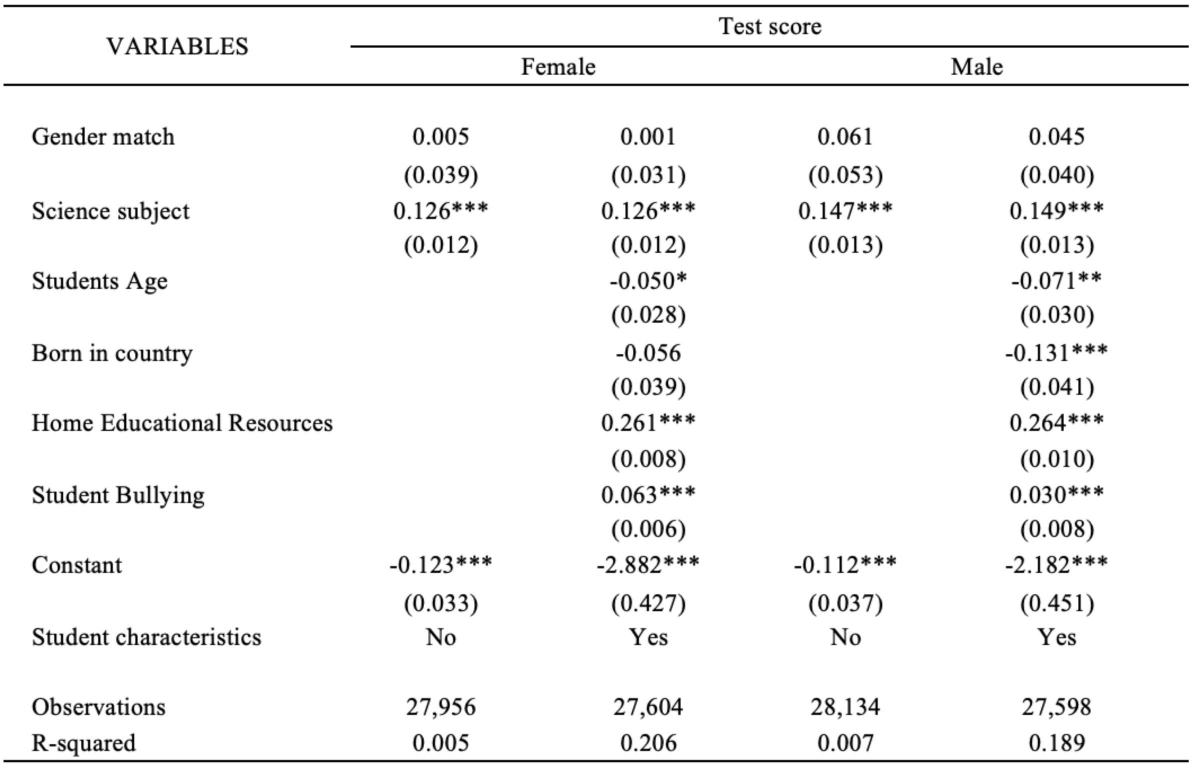
Note : The standard deviations are higher than when only one plausible value is used because the variation between the various estimates is taken into account here.

Table A5: Kernel Density Of Math And Science Test Scores By Gender



Note : The kernel density is measured using only the 1st plausible value.

Table A6: Pooled OLS Regression Of Test Scores On Gender Match (Using All PVs)



VARIABLES	Test score			
	Female		Male	
Gender match	0.005	0.001	0.061	0.045
	(0.039)	(0.031)	(0.053)	(0.040)
Science subject	0.126***	0.126***	0.147***	0.149***
	(0.012)	(0.012)	(0.013)	(0.013)
Students Age		-0.050*		-0.071**
		(0.028)		(0.030)
Born in country		-0.056		-0.131***
		(0.039)		(0.041)
Home Educational Resources		0.261***		0.264***
		(0.008)		(0.010)
Student Bullying		0.063***		0.030***
		(0.006)		(0.008)
Constant	-0.123***	-2.882***	-0.112***	-2.182***
	(0.033)	(0.427)	(0.037)	(0.451)
Student characteristics	No	Yes	No	Yes
Observations	27,956	27,604	28,134	27,598
R-squared	0.005	0.206	0.007	0.189

Note: This table presents results using all plausible values. A Jackknife bootstrapping method specific to TIMSS data is applied to calculate standard errors. The test scores are standardized as to have a 0 mean and a standard deviation of 1. Sampling weights specific to each student are also used.

* Statistically significant at the 10-percent level.
** Statistically significant at the 5-percent level.
*** Statistically significant at the 1-percent level.

Table A7: First Difference Regression Restricting To Knowing Cognitive Domain Test Scores (Using All PVs)

VARIABLES	Test score difference					
	Female			Male		
	Without Subject Specific Effect Of Gender Match					
Gender match	0.011 (0.018)	0.017 (0.020)	0.019 (0.022)	-0.009 (0.018)	-0.010 (0.017)	-0.011 (0.017)
Teacher and class controls	No	Yes	Yes	No	Yes	Yes
Teacher perception controls	No	No	Yes	No	No	Yes
Observations	13,978	13,630	12,080	14,067	13,725	12,208
R-squared	0.015	0.019	0.020	0.012	0.015	0.019
	With Subject Specific Effect Of Gender Match					
Gender match in Math	-0.016 (0.027)	-0.011 (0.028)	-0.015 (0.033)	-0.022 (0.021)	-0.022 (0.021)	-0.025 (0.021)
Gender match in Science	0.037 (0.024)	0.046* (0.026)	0.054* (0.028)	0.005 (0.028)	0.002 (0.027)	0.002 (0.026)
Teacher/classroom controls	No	Yes	Yes	No	Yes	Yes
Teacher perception controls	No	No	Yes	No	No	Yes
Observations	13,978	13,630	12,080	14,067	13,725	12,208
R-squared	0.016	0.020	0.022	0.012	0.015	0.020

Note: This model uses the student test scores on the knowing cognitive area of the assessment. These test scores are also evaluated with plausible values that I standardize. This model is otherwise identical to the previous tables.

* Statistically significant at the 10-percent level.

** Statistically significant at the 5-percent level.

*** Statistically significant at the 1-percent level.

Table A8 : Heterogeneity Analysis For Home Resources, Subject Specific And Knowing Domain Scores (Using all PVs)

VARIABLES	Test score difference					
	Female			Male		
			Low Home Resources			
Gender match in Math	-0.056	-0.049	-0.060	-0.003	-0.004	0.002
	(0.046)	(0.047)	(0.053)	(0.043)	(0.043)	(0.043)
Gender match in Science	-0.006	-0.009	0.001	-0.033	-0.023	-0.022
	(0.035)	(0.038)	(0.041)	(0.043)	(0.041)	(0.040)
Teacher and class controls	No	Yes	Yes	No	Yes	Yes
Teacher perception controls	No	No	Yes	No	No	Yes
Observations	4,907	4,770	4,220	5,336	5,202	4,620
R-squared	0.024	0.030	0.034	0.011	0.015	0.022
			High Home Resources			
Gender match in Math	0.022	0.028	0.019	-0.040	-0.034	-0.045
	(0.032)	(0.032)	(0.039)	(0.035)	(0.034)	(0.034)
Gender match in Science	0.070**	0.089**	0.090**	0.025	0.010	0.012
	(0.034)	(0.036)	(0.040)	(0.037)	(0.034)	(0.035)
Teacher/classroom controls	No	Yes	Yes	No	Yes	Yes
Teacher perception controls	No	No	Yes	No	No	Yes
Observations	4,907	4,770	4,220	5,336	5,202	4,620
R-squared	0.029	0.033	0.035	0.014	0.018	0.025

Note: This model allows subject specific effects and use all plausible values of student knowing test scores. Panel A evaluates the estimates for students with low home resources (index lower than 10.5) and panel B for student with high home resources (index higher than 11.5). This model is otherwise identical to the previous tables.

* Statistically significant at the 10-percent level.

** Statistically significant at the 5-percent level.

*** Statistically significant at the 1-percent level.

Table A9 : Heterogeneity Analysis For Home Resources, Subject Specific And Standard Scores (Using PV1)

VARIABLES	Test score difference					
	Female			Male		
			Low Home Resources			
Gender match in Math	-0.046*	-0.040	-0.049*	0.021	0.022	0.020
	(0.026)	(0.026)	(0.028)	(0.027)	(0.027)	(0.027)
Gender match in Science	-0.009	-0.016	-0.010	-0.017	-0.008	-0.006
	(0.025)	(0.025)	(0.026)	(0.028)	(0.029)	(0.028)
Teacher and class controls	No	Yes	Yes	No	Yes	Yes
Teacher perception controls	No	No	Yes	No	No	Yes
P-value (H_0: $B_m = B_s$)	0.291	0.487	0.264	0.381	0.498	0.537
Observations	4,907	4,770	4,220	5,336	5,202	4,620
R-squared	0.027	0.033	0.038	0.013	0.015	0.021
			High Home Resources			
Gender match in Math	-0.001	0.007	0.002	-0.040	-0.048*	-0.052*
	(0.026)	(0.025)	(0.028)	(0.029)	(0.028)	(0.029)
Gender match in Science	0.043*	0.055**	0.050*	0.021	0.012	0.012
	(0.024)	(0.025)	(0.027)	(0.027)	(0.027)	(0.027)
Teacher/classroom controls	No	Yes	Yes	No	Yes	Yes
Teacher perception controls	No	No	Yes	No	No	Yes
P-value (H_0: $B_m = B_s$)	0.190	0.151	0.193	0.104	0.113	0.100
Observations	5,477	5,354	4,712	5,218	5,086	4,519
R-squared	0.033	0.042	0.048	0.021	0.034	0.039

Note: This model allows subject specific effects and use the first plausible value of student standard test scores. Panel A evaluates the estimates for students with low home resources (index lower than 10.5) and panel B for student with high home resources (index higher than 11.5). This model is otherwise identical to the previous tables.

* Statistically significant at the 10-percent level.

** Statistically significant at the 5-percent level.

*** Statistically significant at the 1-percent level.