Gender equality and the gender gap in sustaining performance

ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Bachelor Thesis Economics and Business Economics

Name student: Iris Wiggerts

Student ID number: 494876

Supervisor: dr. Sacha Kapoor

Second assessor: Anna Baiardi

Date final version: 10 – 08 – 2021

Table of Contents

1 Introduction

In this paper, the paper by Balart and Oosterveen (2009) about the gender gap in sustained performance during test taking is replicated and extended. The ability to sustain performance is relevant in school as well as for a successful career later on. At the workplace, many tasks require multiple hours of work, and being able to sustain one's performance better can impact the quality of the work delivered and the time spent on a task. Knowledge about the source of gender differences in this skill is currently limited, on the individual level as well as on the country level. This is why research on this topic is very important, and besides implications for career advancements, the research on gender gaps in sustaining performance also has large implications for the gender math gap.

It has widely been documented that females and males tend to perform better in different areas of study. Males have an advantage in mathematics, while females tend to perform better in the reading domain (Balart and Oosterveen, 2009; Dee, 2006; Marks, 2008). The gender gap in mathematics performance has been researched extensively in particular, as mathematic skills are needed in order to pursue a career in the science, technology, engineering and mathematics (STEM) fields. Hyde and Mertz (2009) have shown that the gender gap in the STEM fields is related to different socioeconomic outcomes between men and women, which is why the gender gap in mathematics has large implications for gender equality. Balart and Oosterveen (2009) show that the female advantage in sustaining performance could mediate the gender math gap. They documented that a gender gap in sustaining performance exists both in domains favourable to females and in domains favourable to males across countries and time, and tested several hypotheses of potential sources of the gender gap on the micro level. However, they do not interpret the magnitude of the values of the gender gaps found. Nor do they try to explain the difference in the values across countries. This study aims to fill this gap in the current literature on this topic, which is very scarce. Apart from the paper by Balart and Oosterveen (2009), only one other paper, by Borghans and Schils (2019), has been written on this topic. In this paper, the gender gap values will be compared across countries, and potential sources of the differences in the value of the gender gap across countries will be discussed. Specifically, the relationship between the size of the gender gap in sustaining performance and the gender equality in a country will be evaluated. Females tend to benefit more from gender equality when it comes to performance in math tests or cognitive tests, resulting in smaller gender gaps in countries with more gender equality (Riley et. al., 2009). Moreover, the gender gap in reading, which favours females, increases in countries with more gender equality (Guiso et. al, 2008). Thus, the main hypothesis of this study is:

The gender gap in sustaining performance is larger in countries with more gender equality.

To assess the validity of this hypothesis, the data generated by Balart and Oosterveen (2009) from the PISA will be used. The estimates of the gender gap in sustaining performance for each country will be regressed on several indicators of gender equality to determine whether there is a significant correlation between the size of the gender gap in sustaining performance and gender equality across countries. The main finding of this paper is that there does not seem to be a direct correlation between gender equality and the size of the gender gap across countries. None of the measures for gender equality used in this study were statistically significant at the 5% level in the full model, but the gender tertiary education ratio, measuring how many men compared to women followed tertiary education, is statistically significant at the 5% level in 2 out of 5 models. The value of this coefficient is negative, which implies that in countries where more women compared to men follow tertiary education, the gender gap in sustaining performance is larger. This is in line with the main hypothesis, but can only be seen as weak evidence. Overall, the results indicate that if the gender gap in sustaining performance is correlated with social or environmental differences, these differences are likely unrelated to gender equality. Another reason for not finding a significant correlation between gender equality and the gender gap in sustained performance could be that this gender gap mostly originates from biological differences between boys and girls. Balart and Oosterveen (2019) found that in case of higher stakes (the PISA is low-stake in nature), the gender gap in sustaining performance still persisted, even though the value of the gender gap was smaller. It could be the case that when stakes are higher, although the gender gap in sustaining performance is smaller, the variability of this value is larger across countries as a result of gender equality differences. Further research on this topic is needed to evaluate the findings of this research and determine possible sources of the gender gap in sustained performance.

The remainder of this paper is structured as follows. In the next section, the theoretical framework gives the context in which this paper is relevant and analyses previous research done on this topic. Section 3 provides a description of the data used in the replicated paper and in the extension. Section 4 describes the models used in the replicated paper and in the extension. In section 5, the paper by Balart and Oosterveen (2019) is summarized and replicated. In section 6, the replicated paper is extended and lastly in section 7, a discussion of the results is given and conclusions are drawn.

2 Theoretical framework

The current literature on the ability to sustain performance during test taking is limited. Apart from the paper by Balart and Oosterveen (2019) reproduced in this study, only Borghans and Schils (2019) studied this subject. Borghans and Schils (2019) aimed to decompose test scores into measures of cognitive and noncognitive skills, using PISA data. They documented that performance declines during test taking, and that this decline is uncorrelated with initial performance, at the individual level and at the country level. By analysing results of different years they showed that the differences in the observed performance decline were stable over time. Moreover, they analysed whether several cognitive (i.e. IQ and Achievement test) and noncognitive (i.e. personality traits such as openness and need for achievement) measures were related to either performance at the start of the test or performance decline. They found that performance at the start of the test was strongest related to the personality trait openness and to test scores. The performance decline during the test was most strongly related to the personality traits conscientiousness and agreeableness, and need for achievement. Lastly they evaluated whether the performance decline during the test could predict future outcomes. They found that all future outcomes that they considered were more related to the performance at the start of the test than to the performance decline, with performance at the start of the test being the most strongly correlated with labour market outcomes and health outcomes.

Balart and Oosterveen (2019) used the same data as Borghans and Schils (2019), but tested whether there are gender differences in the ability to sustain performance. Moreover, they tested whether these results were stable both in domains favourable to females and domains favourable to males. They then assessed several hypotheses of potential causes of the gap in sustaining performance and evaluated whether longer tests could reduce math gender gaps. Their main findings were that females are better at sustaining their performance, and that this result holds both in the domain favourable to females and in the domain favourable to males. However, none of the potential causes of the gender gap in sustaining performance that they considered could mediate the gender gap. Both studies are in line with the finding that performance decline during test taking is present, but they are not able to identify the sources of the performance decline. Moreover, cross-country comparisons were not made in either paper, leaving much to be researched.

Another way in which continuous performance can be tested, is by conducting the gradual onset continuous performance task (gradCPT). It is designed to measure sustained attentional control, and requires participants either to respond (for instance by pressing a key) when presented with a frequent picture, or to withhold their response when presented with a rare image. Riley et. al. (2016) assessed whether sustained attentional control, as measured by the gradCPT, is related to inequality

across countries. They measured the reaction time, the coefficient of variation (the standard deviation of reaction times divided by the mean reaction time), the commission error rate (responding when one should not respond) and the omission rate (not responding when one should respond). They found that women make more errors of omission and have a greater degree in fluctuation of performance, and that men make more errors of commission. This is in contradiction with Conners et. al. (2003), who found that females made less omission and commission mistakes, but had a longer reaction time. Chan (2000) found no effect of gender on the performance of the Sustained Attention to Response Task (SART), which is a similar task to the gradCPT. Riley et. al. (2016) additionally linked gradCPT to socio-cultural conditions in a country, and they found that gender differences in performance were larger in more unequal countries. In particular, gender equality is significantly correlated to the gender difference. This correlation is mostly driven by fluctuations in the female performance, implying that in a country that has worse socio-cultural conditions, females show less sustained attentional control.

The findings of Riley et. al. (2016) are in line with the predictions of Weber et. al. (2013), who hypothesize that women benefit disproportionately from societal improvements, because they are more disadvantaged than men at the start. They also state that cognitive skills are highly related to visuospatial and mathematical abilities, in which men perform better, and episodic memory and reading literacy, in which females perform better. Their main discovery is that less gender-restricted educational opportunities are associated with gender gap favouring women for some cognitive abilities, and a decreased or eliminated gender gap favouring men for other cognitive abilities. They also look at the cognitive abilities in three separate regions in Europe, and conclude that the size of the gender gap in cognitive abilities varies across regions. Reilly (2012) states that changes in the size or direction of the gender gap implicate that environmental or cultural factors play a large role in determining the gender gap. They also suggest that cognitive abilities of women are influenced by gender stereotypes and gender roles for women in society, which is also known as the gender stratification hypothesis.

Besides the gender gap in cognitive skills, the gender gap in mathematics is also highly discussed. The gap has narrowed over the years and is in some countries even non-existent or reversed (Hyde and Mertz, 2009; Marks, 2008). The source of this gap has been debated extensively, with some researchers claiming it stems from biological differences (Geary ,1998; Kimura, 1999; Baron-Cohen, 2003), while others conclude that environmental and cultural factors also play a large role (Penner, 2008; Dee, 2006; Bharadwaj et. al., 2012). Dee (2006) concludes that a same sex teacher would reduce the gender gap both by lowering the performance of boys and by boosting the performance of girls. He shows that the same-sex effect works similar for the gender gap in reading performance.

Having a male language teacher would eliminate nearly a third of the gender gap in reading by lowering girls performance and increasing boys performance. This corroborates the finding of Marks (2008), who demonstrates that the gender gap in reading and mathematics are highly correlated. He shows that policies designed to close the gender math gap at the same time likely increase the gender gap in reading. Herman and Kopasz (2019) observe that the size of the gender differences in reading and mathematics vary per country. They examine whether this variation can be explained by educational policies, as opposed to cultural factors. They find that the education system does matter for the size of the gender gap. More individualized teaching practices benefit females as well as early tracking, which directly improves the performance of girls relative to boys. Hyde and Mertz (2009) also recognize that the gender gap in mathematics is partially due to changeable sociocultural factors. They find that while there is a gender gap in the high-end performance in mathematics for Caucasian Americans, while the opposite is true for Asian Americans. Guiso et. al. (2008) find a correlation between the Gender Gap Index (GGI) and the size of the mean gender gap per country. They concluded that the gender math gap is smaller in countries with more gender equality. Bharadwaj et. al. (2012) do not find a significant relationship between the math gender gap and classroom environment (including same sex teacher). They do find, however, that girls and boys have different perceptions about their own ability in math, which could be influencing the gender gap.

The gradCPT and mathematics test show similar patterns in gender gaps. For both tests, it seems that female performance is influenced by gender equality and that the gender gap differs per country. Although it can be argued that the gradCPT measures something different than the PISA, the mechanisms through which sociocultural conditions influence gender gaps in sustained attentional control could be the same mechanisms that determine the size of gender gaps in sustained performance. Combining the results of Borghans and Schils (2019) and Balart and Oosterveen (2019) who showed that the performance decline was significantly correlated with cognitive and noncognitive factors and that there is a gender gap in sustaining performance, and of Riley et. al. (2016) and Weber et. al. (2013), who proved that the gender gap in cognitive skills is correlated to gender equality in countries, it seems reasonable to hypothesize that the size of the gender gap in sustaining performance across countries is also linked to gender inequality in those countries. Specifically, as the gender gap favours females, the gender gap can be expected to be larger in more equal countries, and smaller in countries with less gender equality.

3 Data description

3.1 Replication data

The data used in this study originate from the Programme for International Student Assessment (PISA). This is a large scale assessment that measures the academic performance of 15 year old's in the subjects of reading, mathematics and science. The PISA also measures student's abilities to meet real-life challenges, however this study mainly uses the data on academic performance. The PISA is administered by the Organisation for Economic Co-operation and Development (OECD). The test is conducted every three years in 93 countries worldwide including OECD as well as non-OECD countries. In this study, data of the years 2006 to 2015 is used for all the participating countries. This way it can be tested whether the gender gap in sustaining performance is stable over time and across countries. PISA also records student's demographics and noncognitive skills such as motivation and self-confidence.

Each year that the PISA is conducted, the focus lies on one of the domains reading, math and science, such that half of the questions are on that domain. The data in 2009 is used for the baseline results in the study of Balart and Oosterveen (2019), as in that year the main topic of evaluation was reading. As a result, approximately half of the questions were in the domain favourable to females (reading) and half of the questions were in the domain favourable to males (math and science). The PISA uses 20 different booklets that vary in difficulty level and countries can either choose for 13 standard booklets or for 13 easier booklets, where 6 booklets occur in both options. Each booklet then contains four clusters of questions that result in a total of 60 questions. In total there are 13 clusters that are distributed over the 13 booklets rotation wise, thus each cluster appears in each of the four positions once. Moreover, the booklets are randomly assigned to students, so the variation of the position of a question in the test is unrelated to the characteristics of students. Balart and Oosterveen (2019) control for the question difficulty level by including question fixed effects and for school quality by including school fixed effects.

In study 2, an existing dataset from Lindberg et al. (2010) was used. This dataset was created by conducting a meta-analysis on the gender gap in math tests. In total, data from 441 math tests were included. Balart and Oosterveen (2019) then collected information on the number of questions, the stakes of the test and the maximum time given to complete the test. They only included tests that had to be finished within a certain time limit, and this resulted in a final inclusion of 243 tests. For 203 out of these 243 tests, they could collect the number of questions and for 175 tests they could collect the maximum time allowed for the test.

3.2 Extension data

To extend the research of Balart and Oosterveen (2019), additional data was added. In the extension, the estimates of the gender gap in sustaining performance, produced by Balart and Oosterveen (2019) are used. The estimates of each year and every country are used, and together they form the variable $QFemale$. This variable is then regressed on several country and year specific variables. The data of these variables come from the statistical data that is collected yearly by the OECD. Not all countries that are included in the study by Balart and Oosterveen (2019) are included in the data collection by the OECD. For each year, there is data on roughly 42 countries for each variable. As the research by Balart and Oosterveen (2019) focuses on the PISA's conducted in the years 2006, 2009, 2012 and 2015, the additional data collected is also from these years. Besides data from the OECD, data from the World Economic Forum was also collected.

The additional data collected is on the gender gap in unemployment, the gender wage gap, the gender gap in tertiary education, the gender gap in employment per industry and lastly a general gender gap index. The gender gap in unemployment is calculated by dividing the percentage of unemployment of males by the percentage of the unemployment of females. Thus, when males' unemployment increases relative to females' unemployment, the ratio increases and vice versa. The gender wage gap is calculated by taking the difference between men's and women's average earnings, and calculating this as a percentage of male's earnings. The median earnings are taken of full-time employees. The gender gap in tertiary education is calculated by dividing the percentage of males in tertiary education by the percentage of females in tertiary education. Tertiary education is defined by the International Standard Classification of Education (ISCED) 2011 as having a bachelor's, master's, doctoral or equivalent level of education. The gender gap in employment per industry includes the agriculture sector, the industrial sector and the service sector. The estimate is calculated by dividing the percentage of males in each sector by the percentage of females in each sector. The gender gap index is calculated by the World Economic Forum, and takes into account for different indicators for the index. The indicators are economic participation and opportunity, educational attainment, health and survival and political empowerment. Each indicator consists of several ratio's measuring equality between males and females. Supplementary Table 1 shows the descriptive statistics for the extension data. The sector employment ratios show that on average, relatively more females work in the service sector, and in the industrial sector females are the most outnumbered by men. The gender tertiary education ratio has a mean value of 0.9197, indicating that on average more women follow tertiary education. Moreover, the mean gender unemployment ratio of 0.9834 shows that on average, more females than males are unemployed. Supplementary Figure 1 shows the average gender gap estimate for each year per region in Europe, and for countries outside of

Europe. Weber et. al. (2014) found that gender differences of cognitive abilities had different patterns across regions in Europe. A similar pattern cannot be seen for the gender differences in sustaining performance. However, data of more years should be included to conclude with certainty that the gender gap in sustaining performance does not show a similar pattern to the pattern described by Weber et. al. (2014).

4 Methodology

4.1 The replicated model

The first model that accounted for a performance decline in cognitive tests was by Borghans and Schils (2019). Their basic model tested the following specifications:

$$y_{ij} = \alpha_0 + \alpha_1 Q_{ij} + \epsilon_{ij},$$

where the dependent variable y_{ij} is a dummy variable that returns value 1 if student *i* answered question *j* correctly and value 0 if they answered wrongly. Q_{ij} denotes the position of question *j* for student *i* in the test and is normalized between 0 and 1. Coefficient α_1 denotes the probability of choosing the right answer depending of the position of the question in the test. A negative value for α_1 indicates a performance decline. Borghans and Schils (2019) previously estimated the equation above and found the α_1 to be negative in each country that was included. The constant α_0 denotes the probability of choosing the right answer at the beginning of the test.

Balart and Oosterveen (2019) extended this model by including a gender dummy and interacting this dummy with the position of a question in the test:

$$y_{hij} = \beta_0 + \beta_1 Q_{ij} + \beta_2 F_i + \beta_3 Q_{ij} F_i + \boldsymbol{J_j} + \boldsymbol{H_h} + \varepsilon_{hij},$$

where h is a subscript for school, and $\boldsymbol{J_j}$ and $\boldsymbol{H_h}$ are question and school fixed effects, respectively. Coefficient β_3 denotes the effect of being a female on the effect that the position of a question in the test has on the probability of choosing the right answer. This coefficient allows for an estimation of whether females or males are better able to sustain their performance during a test. Balart and Oosterveen (2019) then added topic dummies to determine if the gender gap in sustaining performance is the same in subjects that either females (reading) or males (math and science) perform better on:

$$y_{hij} = \gamma_0^R R_j + \gamma_0^N N_j + \gamma_1^R R_j F_i + \gamma_1^N N_j F_i + \gamma_2^R R_j Q_{ij} + \gamma_2^R N_j Q_{ij} + \gamma_3^R R_j F_i Q_{ij} + \gamma_3^N N_j F_i Q_{ij} + \boldsymbol{J_j} + \boldsymbol{H_h} + \vartheta_{hij}$$

where R_j indicates a question on the topic of reading and N_j indicates a non-reading question, on the topic of either math or science. Q_{ij} is not included separately, as R_j and N_j already include all

questions. Coefficients γ_1^R and γ_1^N indicate the gender difference in performance at the start of the test, separated by topic. Coefficients γ_3^R and γ_3^N indicate the gender difference in sustaining performance, separated by topic. Thus, if F_i takes on a value of 1 and the coefficients γ_3^R and γ_3^N are positive, this means that females can better sustain their performance both in the reading domain and in the math-science domains. As mentioned in the data description, the booklets are distributed randomly across students and the order of the questions differs per booklet. Through inclusion of the question and school fixed effects, the within question variation across students, as a result of the position of a question in the test, is exploited.

Besides the models described above, Balart and Oosterveen (2019) also investigated whether the gender math gap was correlated with the number of questions on a test. The standardized gender math gap was calculated with the following formula: $mgp = \frac{X_{males} - X_{females}}{\sigma_p}$, where X_{males} is the mean performance of males, $X_{females}$ is the mean performance of females and σ_p is the pooled standard deviation. The model is then estimated with the following equation:

$$mgp_i = \theta_0 + \theta_1 noq_i + w_i,$$

Where mgp_i is the standardized gender math gap and noq_i the number of questions on test i. if the coefficient θ_1 takes a negative value, this indicates that when the length of a test increases, the gender math gap consequentially decreases.

4.2 The extended model

This paper extends the model developed by Balart and Oosterveen (2019) by looking at determinants of the gender gap in sustaining performance on a national level. As many countries are included in the study, they most likely also have different characteristics which may impact the gender gap in sustaining performance. To test this hypothesis, the following model is estimated:

$$y_{ct} = \delta_0 X_{ct} + \boldsymbol{C_c} + \boldsymbol{T_t} + \epsilon_{ct},$$

Where the dependent variable is the gender gap in sustaining performance on the country-level. X_{ct} captures the tertiary education gender gap, the Gender Gap Index, the gender gap in agriculture employment, the gender gap in industrial employment, the gender gap in the service sector, the gender gap in unemployment and the gender wage gap. $\boldsymbol{C_c}$ and $\boldsymbol{T_t}$ are country and time fixed effects, respectively. The model is estimated through a multiple regression in OLS.

5 The replication

In this section, the main results of the paper "Females show more sustained performance during test-taking than males" by Balart and Oosterveen (2019) will be replicated. First, an introduction of their research will be given, followed by the replication of the main results and a discussion of the implications.

5.1 Introduction of Balart and Oosterveen: Females show more sustained performance during test-taking than males

The research of Balart and Oosterveen (2019) was on whether there is a gender gap in sustained performance in test taking among 15-year old students. As many complex tasks in the workplace often require a lot of time to finish, it is evident that the ability to sustain performance could largely impact ones success in their professional career. Balart and Oosterveen hypothesized that females are better able to sustain their performance, because of differences in noncognitive skills among females and males. These noncognitive skills could enable female students to better sustain their performance during a cognitive test. Another reason why females might outperform their male peers during test taking is because of different test taking strategies. Balart and Oosterveen (2019) define test taking strategies as "Any reason that leads a student to answer the questions in an order different than the order being administered". Examples of this are skimming through all questions and solving the easiest questions first, or first solving questions which are worth more points. Females have been found to have an advantage in planning (Naglieri & Rojahn, 2001), which could enable them to better finish the test in time or correct mistakes. Lastly, females might be better at sustaining their performance because they can better sustain their effort level. This hypothesis leads from the finding that females exert more effort in non-incentivized tests, and that effort and motivation are highly correlated with test scores (Segal, 2012).

The research done by Balart and Oosterveen (2019) is highly socially relevant because of its implications for gender gaps in different subjects that are tested in school. Females are on average better at verbal and reading tests, while males outperform females on math and science tests (Hyde & Linn, 1988) (Hyde, Fennema & Lamon 1990). Scores obtained for math in high school have been found to impact later in life outcomes on the labour markets (Joensen & Nielsen, 2009). Thus, if females are better at sustaining their performance in test taking, increasing the length of math tests could decrease the gender gap in math scores and possibly also increase the position of females in the labour market. Moreover, previously there has been a focus of males outperforming females on tests with less time, indicating that females perform worse under time pressure. Balart and Oosterveen (2019) test whether longer tests are associated with less time pressure (i.e. more time to solve each exercise), and find that when the amount of questions on a test increases, the testing

time increases less. This indicates that females are actually better at performing under time pressure, and puts the previous findings in a different light.

To test whether there is a gap in sustaining performance during test taking, Balart and Oosterveen (2019) performed two studies. In study 1, they used data from the PISA (more details on this are in the data description). The PISA tests the cognitive skills of 15-year old students in the domains of reading, which females perform better at, and math and science, which favours males. This allows them to test whether the gender gap in sustaining performance persists in the domain favourable for females as well as the domains favourable for males. The baseline results are given for the year 2009. The reason for this is that the PISA this year focused on the reading domain, which led to an even distribution between reading (favourable to females) and math and science (favourable to males). Study 1 also tests some potential determinants of the gender gap in sustaining performance. To test whether test taking strategies could explain the gender gap in sustained effort, data from the PISA in 2015 was used. This is because the test was given on the computer in 58 countries in 2015, and the computer restricted the students from going back and forth between questions. It can thus be ruled out that part of the findings are from test taking strategies instead of the effort level. The PISA in 2015 also recorded the amount of clicks and the amount of time spent per question. Consequently, data form 2015 was used to determine whether sustained effort could be an explanation for the gender gap.

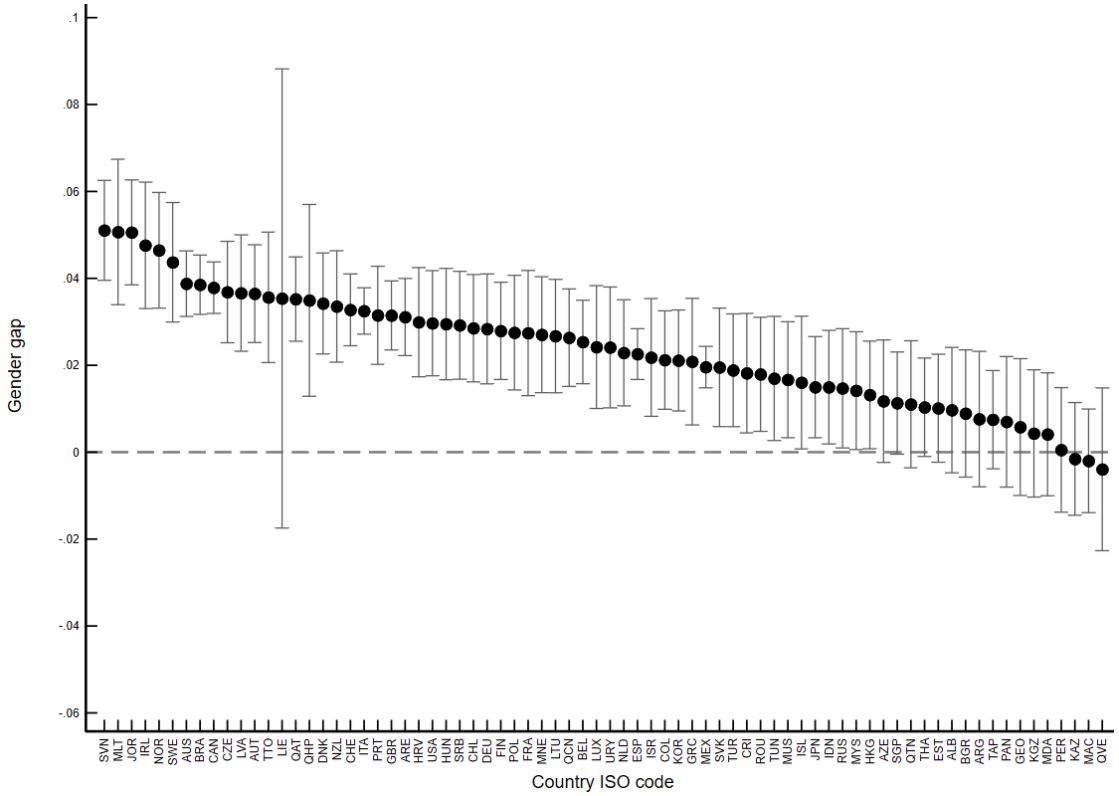
In study 2, data from Lindberg et. al. (2010) was used. This dataset includes female and male performance on around 400 tests, which were conducted worldwide. Balart and Oosterveen (2019) extended this dataset by also including the test length, in terms of questions and time. Study 2 determines the relationship between the amount of questions and the gender gap in performance. For readability, the main results of study 1 are presented in figures. The tables with the estimate of each country as well as the standard error and significance level (p-value obtained through a two-sided t-test) can be found in the appendix.

5.2 The replication process

Balart and Oosterveen (2019) provided their datasets as well as the commands used to get to their results. Consequentially, the replicated results presented in this paper match their results. To make the figures, the datasets generated by the commands of Balart and Oosterveen (2019) were used. The data was then sorted by the gender gap variable in an descending order. Next, a row variable was added, indicating 1 for the first row of data etc., and was given the labels of the country variable. The lower-and upper bound of the 95% confidence interval were calculated with the following formula: $CI = \bar{x} \pm 1.96 * SE$, where $\bar{x}$ is the mean of the gender gap coefficient. After, the 95%

confidence interval and the coefficient estimates for each country were plotted against the row variable.

5.3 Study 1 - part 1



Notes: The figure plots the estimate of the gender difference in sustaining performance during the test for each country participating in the PISA 2009. Positive values indicate countries in which females are better able to sustain their performance during the test than males. Error bars represent the 95% confidence intervals.

Figure 1. Gender differences in sustaining performance.

Figure 1 indicates the gender difference in sustaining performance during test taking in the year 2009. The gender difference in sustaining performance during test taking in the years 2006, 2012 and 2015 in Supplementary Figure 2, 3 and 4 respectively show a similar pattern. The exact estimates of the gender gaps are shown in Supplementary Table 2 in the appendix. The countries are ranked from the largest positive gender gap in sustaining performance to lowest (a negative gender gap, i.e. males better sustain their performance. The coefficient estimates and the 95% confidence intervals were obtained through ordinary least squares (OLS), which determines the difference in the linear slopes that represent the performance decline for males and females. The standard errors are clustered at the student level to account for heteroskedasticity resulting from a binary dependent variable.

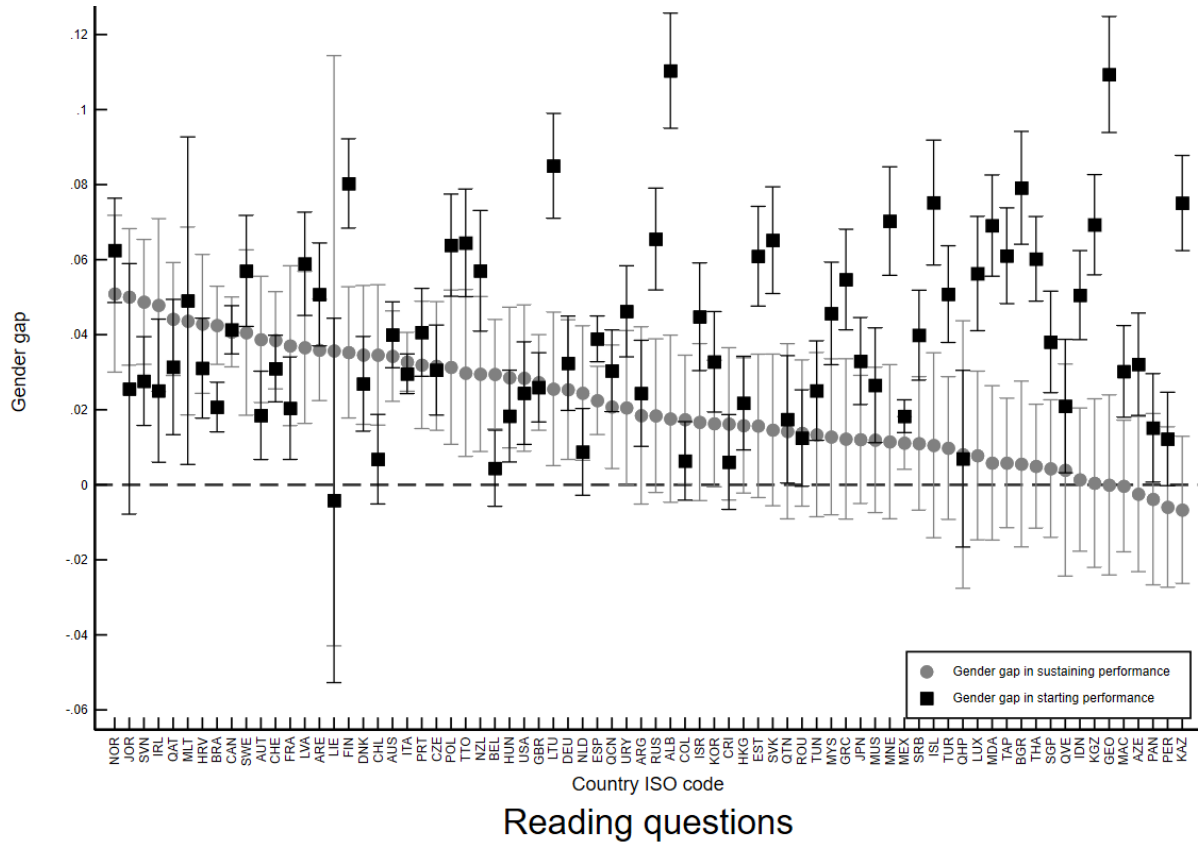
It is remarkable that females were better able to sustain their performance in 71 out of the 74 countries. The estimates of the gender gap were significant at the 5% level for 56 out of the 74 countries. In the countries for which the estimates were negative, Kazakhstan, Miranda and Macao, the estimates were not statistically significant at the 5% level. The point estimate of Slovenia of 0.051

indicates that females are 5.1% more likely to answer the last question of the test correctly than males.

It is notable that the significance of the gender gap in sustaining performance differs across continents. On one hand, in Asia, in 7 out of 21 (8 out of 19 if Russia and Turkey are counted as part of Europe and Georgia as part of Asia) countries the gender gap in sustaining performance is insignificant at the 5% level. This translates to roughly 33% of the countries having an insignificant gender gap. On the other hand, in western continents (Europe, Oceania, North-America), in only 5 out of 40 countries the gender gap in sustaining performance is insignificant at the 5% level. This is equal to 12,5% of the countries not having a significant gender gap. This difference could be the result of a bias in the selection of participating countries, as a larger percentage of western countries participates in the PISA. The Asian countries participating in the PISA could have different characteristics than the Asian countries that are not included in the PISA, and thus the results might be inaccurate. However, a possible explanation for the gender gap being less present in Asian countries is that students have higher intrinsic motivation in test taking. Gneezy et. al. (2019) found this to be true for students in Shanghai compared to students in the USA. Thus, it is reasonable to assume that cultural differences affect the gender gap. Lastly, since only very few countries in Africa or South-America are participating in the PISA it is hard to any statements on the results of those continents.

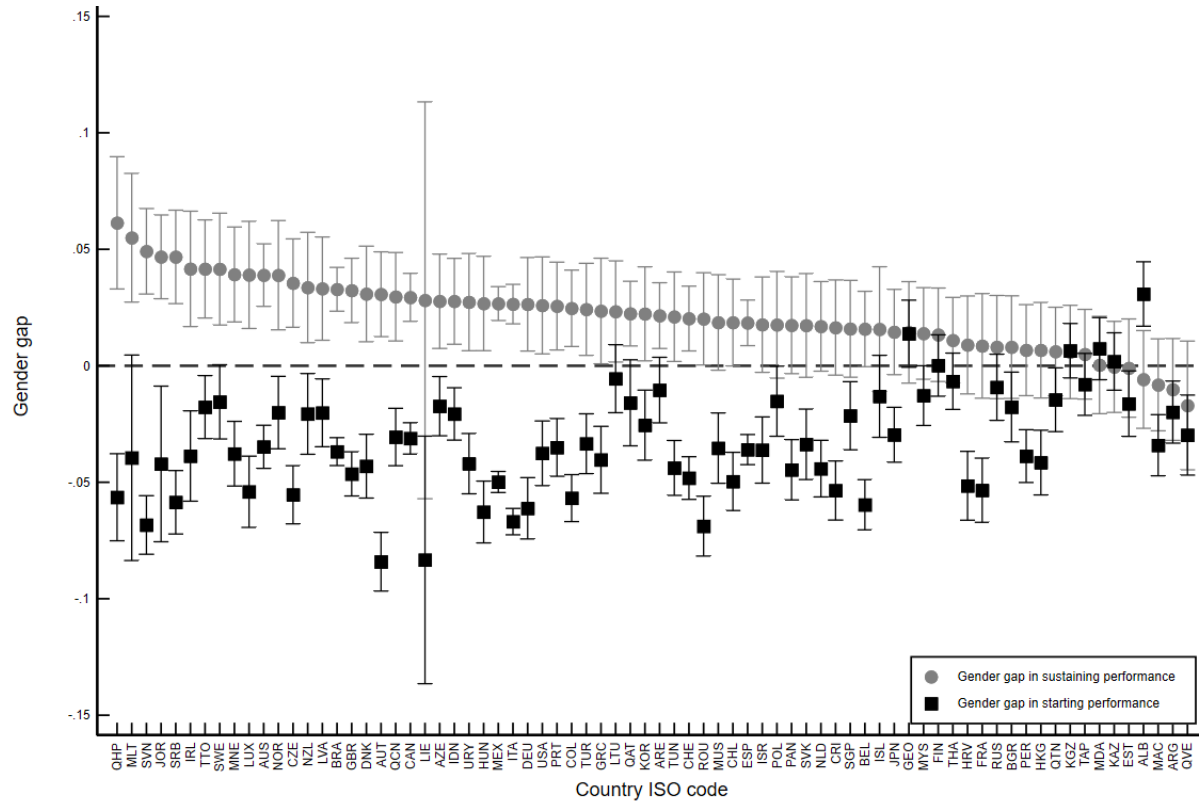
5.4 Study 1 - part 2

A.



Reading questions

B.



Math-science questions

Notes: The panels plot the point estimates of the gender gap in starting performance and in sustaining performance during the test for each country participating in the PISA 2009 for A reading and B math-and-science. Positive values indicate the gender gap favours females. Error bars represent the 95% confidence intervals.

Figure 2. Gender differences in starting performance and in sustaining performance by topic.

The results of the second and most important part of study 1 are shown in Figure 2. Panel A showcases the estimates of the gender gap at the start of the test in grey squares and the gender gap in sustaining performance during the test in black circles on the domain of reading. The grey and black lines represent the corresponding 95% confidence intervals of the estimates. In almost all countries except for one, females score better at the beginning of the test, indicated by positive estimates in the Figure. Moreover, for 64 out of the 74 countries, the 95% confidence interval of the estimate is strictly positive. This corroborates the previous findings that females outperform males on reading tests (Hyde & Linn, 1988). Females are also better at sustaining their performance during test taking in 68 out of the 74 countries. This gender gap is statistically significant at the 5% level in 36 countries.

Panel B showcases the gender difference in performance at the start of the test and the sustained performance on the domain of math and science. The starting performance is indicated by grey squares and the sustained performance is indicated by black circles. The grey and black lines represent the corresponding 95% confidence intervals of the estimates. As was predicted by previous literature, most estimates of starting performance are negative, indicating that males outperform females at the start of the test. For 58 out of the 74 countries, the 95% confidence interval of the estimate is strictly negative. However, most of the black estimates and confidence intervals are positive. This implies that although males have an initial advantage in math or science tests, females are better able to sustain their performance during the test. This gender difference is present in 68 out of the 74 countries, and is statistically significant at the 5% level in 41 countries. When looking at both panels, it seems as if there is a larger variability in the starting performance in reading than in the starting performance in math-science questions. To test this, a two sample variance-comparison test was done. The hypothesis that the standard deviations are the same could not be rejected (p-value of two-sided test = 0.4482, meaning that no evidence was found for a difference in variability for both variables. It is also notable that both figures have different scales, which creates the illusion of a difference. The two panels of Figure 2 imply that the gender gap in sustaining performance is unrelated to the subject that is being tested. The exact estimates of the gender gap in sustaining performance, separated by domain can be found in Supplementary Table 3 in the appendix.

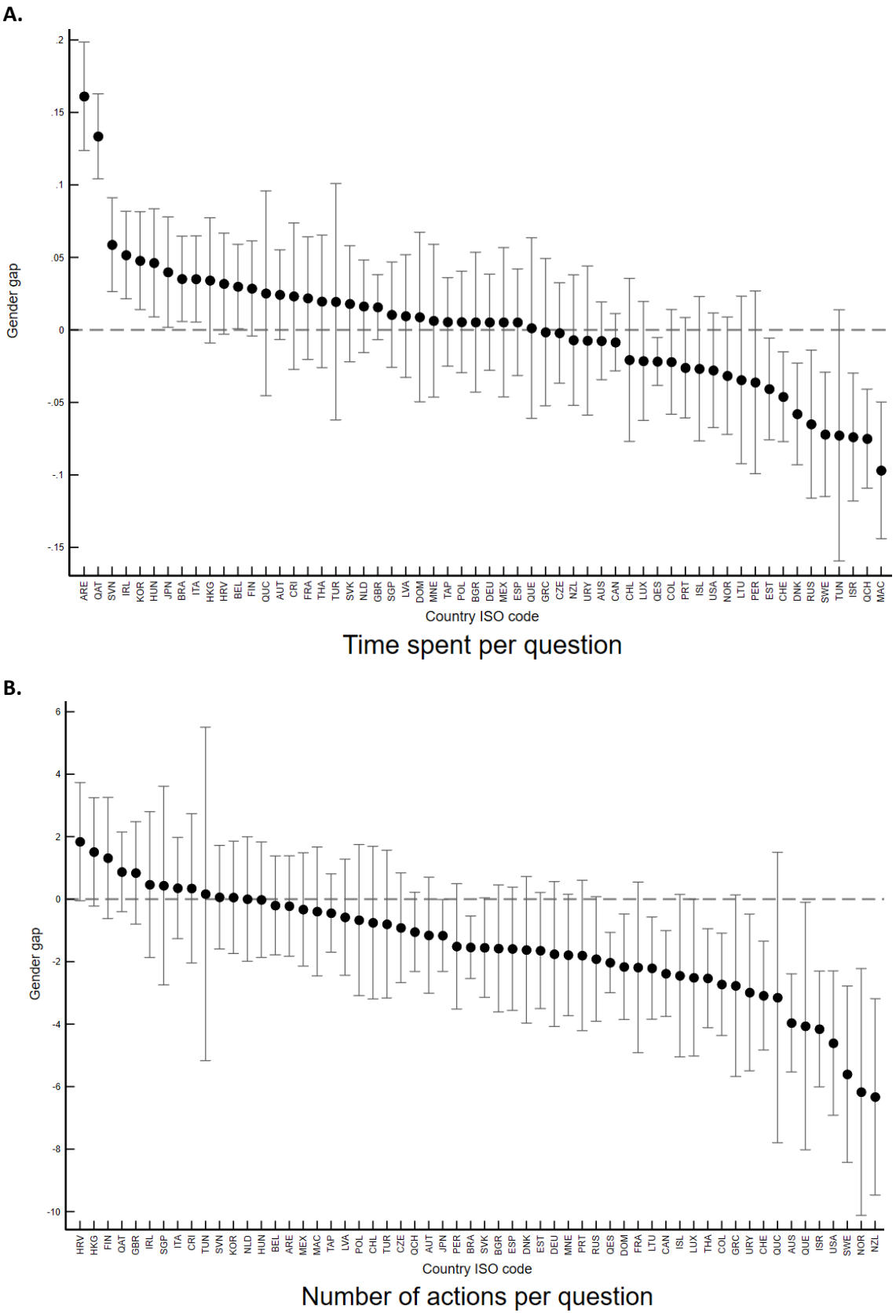
Although not mentioned by Balart and Oosterveen (2019), the values of the gender gaps in sustaining performance in reading and in math-science questions are not significantly different from each other. This was tested by performing a paired t-test. No evidence was found for the means of the gender

gap in sustaining performance in reading and in math-science questions to be different (p-value of two-sided test = 0.8793). Thus, in addition of there being a significant gender gap in both the domains favourable to females and favourable to males, the values of the gender gaps in both categories are not significantly different. This finding excludes the possibility of the gender gap in sustaining performance in the domains favourable to males being smaller than the gender gap in sustaining performance in the domain favourable to females or vice versa. This implies that whether the domain of the test is favourable to males or females affects neither the existence of the gender gap nor the value of the gender gap in sustaining performance.

5.5 Study 1 – potential determinants of the gender difference

After demonstrating the gender gap in sustaining performance in general as well as in specific domains, Balart and Oosterveen (2019) hypothesized three possible causes for the gender gap. Firstly, they consider a difference in noncognitive skills as a possible explanation. Noncognitive skills can be defined as "personality traits, goals, character, motivations, and preferences that are valued in the labour market, in school, and in many other domains" (Kautz et. al., 2014). The noncognitive skills of students were measured through personal questions at the end of the PISA. In each test, different noncognitive skills were collected, such as the interest in a specific domain and the motivation towards the domains math and science. To test whether the noncognitive could mediate the gender gap, they were included in the model. Although the gender differences in noncognitive skills that were found were in accordance with previous literature (Cornwell et. al., 2014), including them in the model did not mediate the gender gap. Besides self-reports, conscientiousness was also tested, by calculating the proportion of questions left unanswered. However, this measure could also not explain the gender gap. Secondly, test taking strategies were considered as a possible explanation for the gender gap. However, the results from the PISA in 2015 in Supplementary Figure 4 in the appendix still show that there was a gender gap in sustaining performance, so this explanation can be disregarded.

Lastly, Balart and Oosterveen (2019) considered sustained effort as a possible explanation for the gender gap. The PISA recorded the amount of clicks per question, which they used as a measure for action. In 48 out of the 58 countries a statistically significant positive correlation was found between the number of actions and having a correct answer to a question. Another measure for effort that was recorded, is the amount of time spent on each question. It has been proven that students that score better on tests also generally take more time (OECD, 2015), so if females sustain the amount of time spent per question, this could explain the gender gap. Figure 3 below plots the results of both measures.



Notes: The panels plot the estimates of the gender gap in sustaining A time spent per question and B the number of actions per question for each country participating in the PISA 2015. Positive values indicate the gender gap favours females. Error bars represent the 95% confidence intervals.

Figure 3. Gender differences in sustaining time spent per question and number of actions per question.

Panel A plots the estimates of the gender gap in the time spent per question and the corresponding 95% confidence intervals. The estimates do not show a clear pattern, as in roughly half of the countries, females decrease the amount of time spent per question more quickly, while in the other half males decrease the amount of time spent per question more quickly. It can thus be concluded that there is no international gender gap in the sustained time spent per question, and thus it is not an explanation for the gender gap in sustained performance. Panel B plots the estimates of the gender gap in the number of actions per question and the corresponding 95% confidence intervals. From the Figure it becomes clear that females actually decrease the number of actions per question more quickly than males in most of the countries. Consequently, sustaining the number of actions per question can also not explain the gender gap. Note that the scale of panel A is much smaller than the scale of panel B.

As none of the three hypotheses could provide an explanation for the gender gap in sustaining performance, Balart and Oosterveen (2019) conclude that the gender gap is not caused by a difference in inputs, such as effort noncognitive skills or test taking strategies. Rather, the gender gap exists because of different abilities to transform inputs into results. Males have been found to experience more boredom during activities that take more time (Vodanovich & Kass1990). A possible explanation could thus be that males are bored more quickly during test taking, which negatively impacts their performance. Supplementary Table 4 provides some suggestive evidence for this hypothesis. However, there was no data available to test this hypothesis directly, so conclusions on this can only be drawn in further research.

5.6 Study 2

Study 1 provides evidence for a gender gap in sustaining performance during test taking. The implication of this finding is that the length of a test could decrease or increase the gender gap in performance in the given domain. As stated before, females on average perform worse than males in math tests. In study 2, Balart and Oosterveen (2019) analyse whether the gender gap in math is smaller when more questions are included in the test. To test this, they regressed the gender gap in math on the number of questions of a test. The results of this regression can be seen in Table 1 below. The coefficient “number of questions” is statistically significant at the 5% level and has a negative value. This is in line with the hypothesis that longer tests decrease the math gender gap. the coefficient value of -0.00159 means that when the number of questions is increased by one, the gender gap in math decreases by 0.00159. The result is robust for excluding an outlier, a test with 240 questions. Balart and Oosterveen (2019) also recalculated the gender gap and reduced the weight by 50% for observations that they coded differently than Lindberg et. al. (2010), whom the data was originally from. As can be seen in columns 3 and 4, the estimates are robust to both of

these changes, although they are now only significant at the 10% level. The results of study 2 cannot be interpreted causally, as it does not exploit exogenous variation in the test length (i.e. no random distribution of test lengths).

Table 1: The relationship between the gender gap in math and the number of questions

	Whole sample (1)	**Exclude outlier (2)**	**Recalculated gender gap (3)**	**Weighted regression (4)**
Number of questions	-0.00159**	-0.00188**	-0.00152*	-0.00149*
	(-2.06)	(-2.10)	(-1.97)	(-1.94)
Constant	0.200***	0.210***	0.194***	0.205***
	(4.59)	(4.48)	(4.40)	(4.33)
N	203	202	203	203
Adjusted R^2	0.012	0.015	0.011	0.10

Notes: t statistics in parentheses, heteroskedasticity robust standard errors *p < 0.10, **p < 0.05, ***p < 0.01. The standardized math gender gap (mgp) is measured by subtracting the mean performance of girls from the mean performance of boys and dividing this by the pooled standard deviation. The equations estimated are as follows: $mgp_i = \delta_0 + \delta_1 noq_i + w_i$, where i is a subscript for test i and noq_i denotes the number of questions on the test.

6 The extension

Although Balart and Oosterveen (2019) showed that the gender gap in sustaining performance was consistent across countries and time, they did not provide a possible explanation for why the magnitude of the gender gap might change per country or per year. This study aims to fill this gap in the current literature on gender gaps in sustaining performance. In table 2 below, the correlations between the gender equality indicators are given. It is notable that many of the indicators are highly correlated with each other. This indicates that even though the indicators are different, they measure similar principles. Moreover, it is a sign of multicollinearity, the phenomenon that variables are not independent from each other.

Table 2: Correlations between measures of gender equity

	Tertiary education ratio	**GGI**	**Agriculture employment ratio**	**Industrial employment ratio**	**Service employment ratio**	**Gender wage gap**	**Unemployment ratio**
Tertiary education ratio	1.000						
GGI	-0.380***	1.000					
Agriculture employment ratio	-0.197**	0.351***	1.000				
Industrial employment ratio	-0.0597	0.613***	0.1149	1.000			
Service employment ratio	0.381***	-0.208**	-0.331***	0.114	1.000		
Gender wage gap	0.400***	-0.280***	-0.358***	-0.197	0.274***	1.000	
Unemployment ratio	-0.253***	0.303***	0.135	0.0912	-0.0343	0.398***	1.000

*p < 0.10, **p < 0.05, ***p < 0.01.

Below in Table 3, the gender gap in sustaining performance is regressed on several variables that measure the gender equality within a country. From left to right in the columns, variables are added to the regression model, while time and country fixed effects are included in each regression. In the full model in the rightmost column, none of the gender equality indicators are statistically significant at the 5% level. However, when not including the variables GGI and the Gender unemployment gap ratio which are highly insignificant (t-statistics of 0.04 and 0.12 in the full model respectively), the gender tertiary education ratio becomes statistically significant at the 5% level. It is also notable that the agriculture sector ratio and the service sector ratio remain statistically significant at the 10% level in models 3 to 5. The values of the gender tertiary education ratio and the agriculture and service sector ratio also remain relatively unchanged in models 3 to 5. The gender tertiary education ratio is negative, indicating that when the ratio increases (i.e. there are relatively less females in tertiary education), the gender gap decreases. The agriculture, industrial and service sector ratios are positive, indicating that when there are relatively more males employed in these sectors, the gender gap in sustaining performance increases. This is unexpected, as especially the agriculture and industrial sectors are male-dominated (see Supplementary Table 1), so a decrease in these ratios would mean more equality and thus a hypothesized increase in the gender gap in sustaining performance. It is notable that when excluding the gender wage gap variable, the values of the other coefficients change drastically. The gender wage gap is significantly correlated with the tertiary education ratio and the agriculture and service sector employment ratios, which is seemingly the main cause of the changes in the coefficients.

The remaining variables do not all have the expected signs. The coefficient of the gender unemployment gap ratio is positive, meaning that when the ratio increases (i.e. relatively more men that women are unemployed), the gender gap in sustaining performance increases. The mean value for the gender unemployment gap ratio is 0.9834, as can be seen in Supplementary Table 1. An increase in this value would indicate more equality, so the value of the gender unemployment gap is also as expected. The value of the gender unemployment gap ratio remains stable in both models. The natural logarithm of the gender wage gap coefficient is very slightly positive. It can be interpreted as follows: when the gender wage gap ratio increases by 10%, the gender gap in sustaining performance increases by $\log(1.1) \times 0.00561 \approx 0.0002$. This indicates that when the gender wage gap is larger, the gender gap in sustaining performance is slightly larger, which is unexpected. The value of the gender wage gap coefficient also stays constant across models. The GGI is negative, indicating that when the GGI increases (i.e. more gender equality), the gender gap in sustaining performance decreases. This is contrary to what can be expected based on relevant literature.

Table 3: The relationship between the gender gap in sustaining performance and country and year specific variables.

	(1)	**(2)**	**(3)**	**(4)**	**(5)**
Gender tertiary education ratio	-0.0140	-0.00950	-0.0422**	-0.0419**	-0.0416*
	(-1.18)	(-0.65)	(-2.08)	(-2.04)	(-1.71)
Agriculture sector ratio		0.00175	0.00656*	0.00645*	0.00646*
		(0.67)	(1.88)	(1.87)	(1.87)
Industrial sector ratio		0.00198	0.0108	0.0110	0.0109
		(0.39)	(1.27)	(1.25)	(1.21)
Service sector ratio		-0.0165	0.155*	0.151*	0.151*
		(-0.32)	(1.84)	(1.78)	(1.76)
Gender wage gap			0.00567	0.00559	0.00561
			(1.62)	(1.60)	(1.66)
Gender unemployment gap ratio				0.00164	0.00164
				(0.13)	(0.12)
GGI					-0.00341
					(0.04)
Constant	0.0373***	0.0347	-0.111	-0.111	-0.113
	(3.35)	(0.74)	(-1.35)	(-1.33)	(-1.03)
R^2	0.6183	0.6252	0.6130	0.6129	0.6129

Notes: obtained by OLS estimation. The natural log is taken of the Gender wage gap indicator. T statistic in parentheses, heteroskedasticity robust standard errors *p < 0.10, **p < 0.05, ***p < 0.01.

There are three possible explanations for most of the coefficients not being statistically significant. The first possible explanation is that the wrong indicators for gender equality were used. Guiso et. al. (2008) found the gender math gap to be significantly correlated with the GGI, which is why this indicator was used in this study. Moreover, Riley et. al. (2016) used the male/female ratio of labour participants in their study and found that gender differences of commission and omission errors were correlated with this measure. In this study, the female labour participation ratio was separated into three employment sectors, and an unemployment gender ratio was added additionally. Baker and Jones (1993) found a correlation between the gender math gap and the percentage of females in higher education, which is why the tertiary education gender ratio was used as an indicator for gender equality. Lastly, Marks (2008) found the gender wage gap to be associated with the gender gap in reading but not with the gender gap in mathematics, which was corroborated by Reilly (2012). All in all, the measures used in this study were based on relevant literature, however gender

inequality indicators that are relevant for the gender math gap could be different from gender inequality indicators that are relevant for the gender gap in sustaining performance. Moreover, most of the measures are highly correlated, which reduces the accuracy of the values of the coefficients. The second potential explanation is that the sample size used in this study is too small, leading to a small statistical power. The data of 42 countries was used for 4 years. To increase the sample size, the data could be collected for more years, or more countries should be included. In future studies, including more countries could also lead to a better balance between countries in Europe and countries in other continents. This will also lead to a more accurate estimate as cultural differences are large between continents. The last potential reason could be that the gender gap in sustaining performance is mostly due to a biological difference between males and females, as opposed to sociocultural differences. It is reasonable to assume that the gender difference in sustaining performance is not entirely determined by biological differences or environmental differences, but by a combination of the two. However, biological differences could be the strongest determinant, or sociocultural and environmental differences other than gender equality could play a larger role.

7 Discussion and conclusion

The aim of this paper was to evaluate whether the size of the gender gap in sustaining performance is related to gender equality across countries. It was hypothesized that the gender gap in sustaining performance is larger in countries with more gender equality and smaller in countries with less gender equality. Gender equality was measured by several indicators, among which the GGI and the gender tertiary education ratio. No real evidence was found in support of this hypothesis, as most estimates were insignificant, with the exception of the gender tertiary education ratio in some of the models. Moreover, the agriculture sector ratio and the service sector ratio were significant at the 10% level in 3 out of 5 models. Although this study could not provide any concrete evidence for the hypothesis:

The gender gap in sustaining performance is larger in countries with more gender equality,

it does highlight that more research needs to be conducted on this subject. The results are very relevant for the debate about gender equality across countries. The fact that no evidence was found for a relationship between gender equality and the gender gap in sustaining performance was unexpected, as many researchers have shown that there is a link between student performance and gender inequality (Guiso et. al., 2008; Hyde and Mertz, 2009; Reilly, 2012). Although this study should be extended with different data and different indicators for gender equality to be able to draw any conclusions, the results suggest that the gender gap in sustaining performance is impacted through a different mechanism than the gender gap in performance.

The main limitation of this study was the sample size used in the extension. Only the PISA data collected in the years that Balart and Oosterveen (2019) used, were used in the extension. If the calculation done by Balart and Oosterveen (2019) would have been extended to the other years that the PISA was conducted, more datapoints could have been used. In addition, the PISA contains only a limited set of countries, and the distribution of countries is not even across continents. By including more countries, it can be ensured that the results are not biased because a sub selection of countries of a certain continent has different characteristics than other countries within that continent. Another limitation of this study is that the indicators used to measure gender equality are highly correlated, which makes it hard to separate individual effects of the indicators on the gender gap in sustaining performance across countries. Using different variables in future studies or centering the variables that are highly correlated could help solve multicollinearity.

The results of this research confirm the importance of more research on this topic. Specifically, the importance of the gender ratio in tertiary education and the gender ratios in employment in different sectors for the gender gap in sustaining performance could be a direction for future research. Including a larger sample of countries, with a more balanced distribution between continents, or more years can provide more accurate results. Moreover, since the PISA is a low-stake test in nature, future research could also focus on cross country differences in gender gaps in sustaining performance in relation to gender equality for tests with higher stakes. Besides gender equality as an explanation of the gender gap on the macro level, other potential explanations such as educational policies, which influence gender gaps in performance (Hermann & Kopasz, 2019), should also be looked into. In addition, future research should also focus on possible determinants of the gender gap in sustaining performance on the micro level, such as upbringing. One last direction for future research, could be to determine if there are differences in the gender gap in sustaining performance in different regions of a country. If differences are found on the meso level, this could point to cultural or socioeconomic differences within countries as opposed to between countries as causes of the differences in the gender gap in sustaining performance.

Bibliography

Baker, D. P., & Jones, D. P. (1993). Creating gender equality: Cross-national gender stratification and mathematical performance. *Sociology of education*, 91-103.

Balart, P., & Oosterveen, M. (2019). Females show more sustained performance during test-taking than males. *Nature communications*, 10(1), 1-11.

Baron-Cohen, S. (2004). *The essential difference*. Penguin UK.

Bharadwaj, P., De Giorgi, G., Hansen, D., & Neilson, C. (2012). The gender gap in mathematics: Evidence from low-and middle-income countries (No. w18464). *National Bureau of Economic Research*.

Borghans, L., & Schils, T. (2019). Decomposing achievement test scores into measures of cognitive and noncognitive skills. *Available at SSRN 3414156*.

Chan, R. C. (2001). A further study on the sustained attention response to task (SART): the effect of age, gender and education. *Brain Injury*, 15(9), 819-829.

Conners, C. K., Epstein, J. N., Angold, A., & Klaric, J. (2003). Continuous performance test performance in a normative epidemiological sample. *Journal of abnormal child psychology*, 31(5), 555-562.

Cornwell, C., Mustard, D. B., & Van Parys, J. (2013). Noncognitive skills and the gender disparities in test scores and teacher assessments: Evidence from primary school. *Journal of Human resources*, 48(1), 236-264.

Geary, D. C. (1996). Sexual selection and sex differences in mathematical abilities. *Behavioral and brain sciences*, 19(2), 229-247.

Gneezy, U., List, J. A., Livingston, J. A., Qin, X., Sadoff, S., & Xu, Y. (2019). Measuring success in education: the role of effort on the test itself. *American Economic Review: Insights*, 1(3), 291-308.

Guiso, L., Monte, F., Sapienza, P., & Zingales, L. (2008). Culture, gender, and math. *SCIENCE-NEW YORK THEN WASHINGTON-*, 320(5880), 1164.

Hermann, Z., & Kopasz, M. (2019). Educational policies and the gender gap in test scores: a cross-country analysis. *Research Papers in Education*, 1-22.

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin*, 104(1), 53–69.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis*. Psychological Bulletin*, 107(2), 139–155.

Joensen, J. S., & Nielsen, H. S. (2009). Is there a causal effect of high school math on labor market outcomes?. *Journal of Human Resources*, 44(1), 171-198.

Kautz, T., Heckman, J. J., Diris, R., Ter Weel, B., & Borghans, L. (2014). Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success.

Kimura, D. (1999). *Sex and cognition*. MIT press.

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin*, 136(6), 1123–1135.

Marks, G. N. (2008). Accounting for the gender gaps in student performance in reading and mathematics: evidence from 31 countries. *Oxford Review of Education*, 34(1), 89-109.

Naglieri, J. A., & Rojahn, J. (2001). Gender differences in planning, attention, simultaneous, and successive (PASS) cognitive processes and achievement. *Journal of Educational Psychology, 93*(2), 430–437.

Organization for Economic Co-operation and Development PISA 2015 Technical Report (OECD, Paris, 2015).

Penner, A. M. (2008). Gender differences in extreme mathematical achievement: An international perspective on biological and social factors. *American Journal of Sociology*, 114(S1), S138-S170.

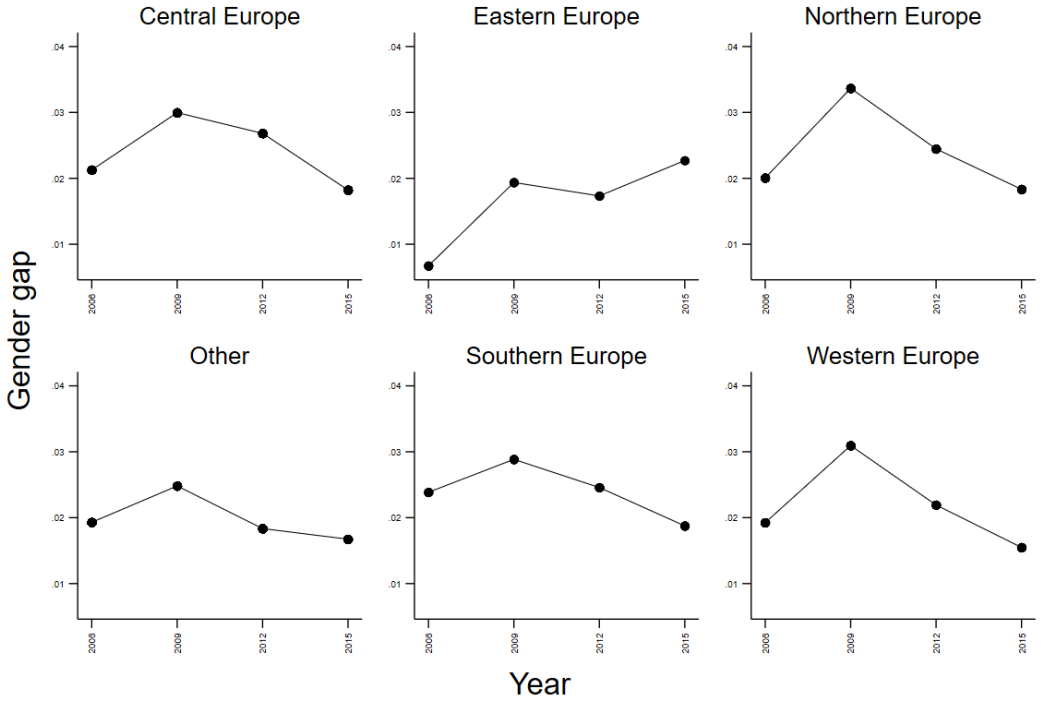
Reilly, D. (2012). Gender, culture, and sex-typed cognitive abilities. *PloS one*, 7(7), e39904.

Riley, E., Okabe, H., Germine, L., Wilmer, J., Esterman, M., & DeGutis, J. (2016). Gender differences in sustained attentional control relate to gender inequality across countries. *PloS one*, 11(11), e0165100.

Segal, C. (2012). Working when no one is watching: Motivation, test scores, and economic success. *Management Science*, 58(8), 1438-1457.

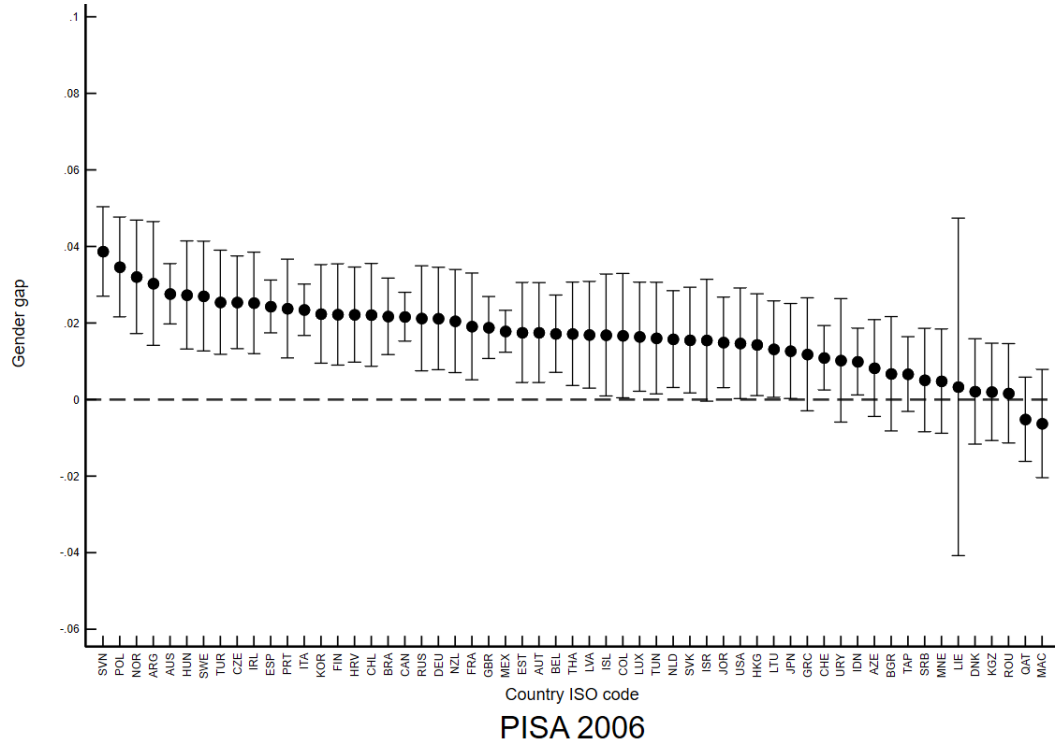
Vodanovich, S. J. & Kass, S. J. (1990). Age and gender differences in boredom proneness. J. *Soc. Behav. Pers.* 5, 297–307.

Appendix



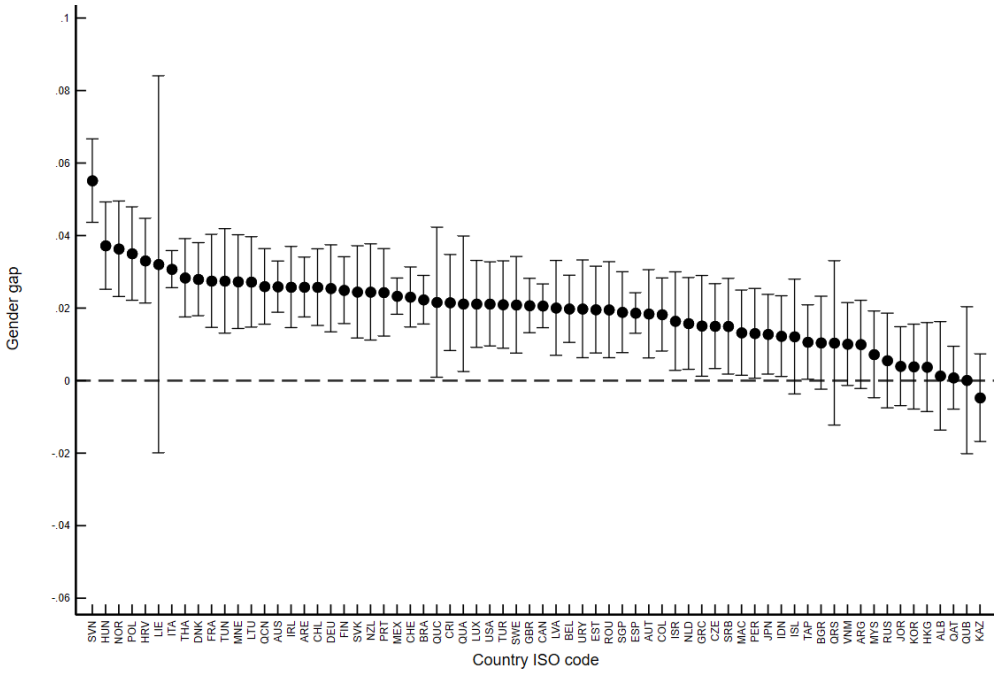
Notes: The mean gender gap in sustaining performance for Central, Eastern, Northern, Western and Southern Europe, as well as for countries outside Europe.

Supplementary Figure 1. Mean gender gap in sustaining performance per region



Notes: The Figure plots the estimate of the gender difference in sustaining performance during the test for each country participating in the PISA 2009. Positive values indicate countries in which females are better able to sustain their performance during the test than males. Error bars represent the 95% confidence intervals.

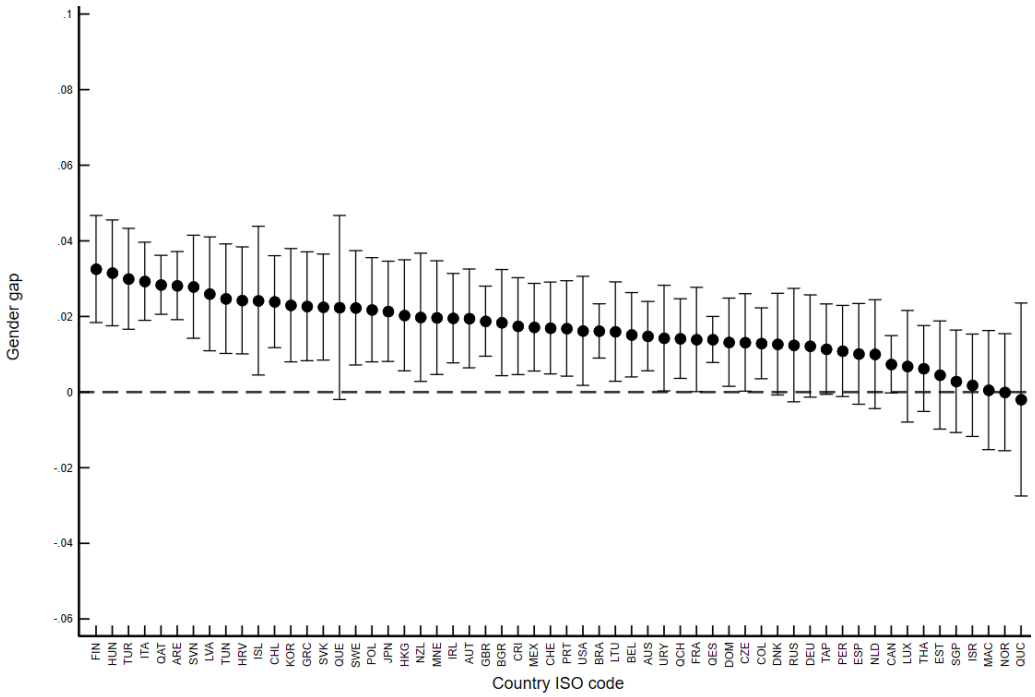
Supplementary Figure 2. Gender differences in sustaining performance



PISA 2012

Notes: The figure plots the estimate of the gender difference in sustaining performance during the test for each country participating in the PISA 2009. Positive values indicate countries in which females are better able to sustain their performance during the test than males. Error bars represent the 95% confidence intervals.

Supplementary Figure 3. Gender differences in sustaining performance



PISA 2015

Notes: The Figure plots the estimate of the gender difference in sustaining performance during the test for each country participating in the PISA 2009. Positive values indicate countries in which females are better able to sustain their performance during the test than males. Error bars represent the 95% confidence intervals.

Supplementary Figure 4. Gender differences in sustaining performance

Supplementary Table 1: Descriptive statistics of country-level variables

	N	Minimum	Maximum	Mean	Standard deviation
Tertiary education ratio	150	0.5558	1.8158	0.9197	0.2185
Gender gap index	165	0.5828	0.881	0.7186	0.0552
Agriculture employment ratio	144	0.4547	7.2121	2.3117	1.1732
Industrial employment ratio	142	1.3245	5.2971	2.8707	0.7901
Service employment ratio	142	0.5793	1.1460	0.7134	0.0775
Gender wage gap	110	0.4	39.8	14.8791	7.4891
Unemployment ratio	160	0.4130	1.6286	0.9834	0.2262

Supplementary Table 2: Gender differences in sustaining performance during the test. The PISA 2009.

CNT	Coefficient	Standard error	CNT	Coefficient	Standard error	CNT	Coefficient	Standard error
ALB	0.0097	0.0074	HRV	0.0299	0.0064***	NZL	0.0335	.0335***
ARE	0.0311	0.0045***	HUN	0.0295	0.0065***	PAN	0.0070	.0077
ARG	0.0076	0.0080	IDN	0.0150	0.0067**	PER	0.0005	.0073
AUS	0.0388	0.0038***	IRL	0.0476	0.0074***	POL	0.0275	.0067***
AUT	0.0365	0.0057***	ISL	0.0160	0.0078**	PRT	0.0315	.0058***
AZE	0.0117	0.0072	ISR	0.0218	0.0069***	QAT	0.0352	.0049***
BEL	0.0254	0.0049***	ITA	0.0325	0.0027***	QCN	0.0263	.0057***
BGR	0.0089	0.0075	JOR	0.0506	0.0062***	QHP	0.0349	.0113***
BRA	0.0385	0.0035***	JPN	0.0150	0.0059**	QTN	0.0110	.0075
CAN	0.0379	0.0030***	KAZ	-0.0016	0.0066	QVE	-0.0039	.0096
CHE	0.0328	0.0042***	KGZ	0.0043	0.0075	ROU	0.0179	.0067***
CHL	0.0285	0.0063***	KOR	0.0211	0.0059***	RUS	0.0147	.0070**
COL	0.0212	0.0058***	LIE	0.0354	0.0270	SGP	0.0113	.0060*
CRI	0.0182	0.0070***	LTU	0.0267	0.0067***	SRB	0.0292	.0063***
CZE	0.0368	0.0060***	LUX	0.0242	0.0072***	SVK	0.0195	.0070***
DEU	0.0284	0.0065***	LVA	0.0366	0.0068***	SVN	0.0510	.0059***
DNK	0.0342	0.0059***	MAC	-0.0020	0.0061	SWE	0.0437	.0070***
ESP	0.0226	0.0030***	MDA	0.0041	0.0072	TAP	0.0075	.0058
EST	0.0101	0.0063	MEX	0.0196	0.0024***	THA	0.0103	.0058*
FIN	0.0279	0.0057***	MLT	0.0507	0.0085***	TTO	0.0356	.0077***
FRA	0.0274	0.0074***	MNE	0.0270	0.0068***	TUN	0.0170	.0073**
GBR	0.0315	0.0041***	MUS	0.0167	0.0068**	TUR	0.0188	.0066***
GEO	0.0058	0.0080	MYS	0.0142	0.0069**	URY	0.0241	.0071***
GRC	0.0208	0.0074***	NLD	0.0229	0.0062***	USA	0.0297	.0062***
HKG	0.0132	0.0063**	NOR	0.0465	0.0068***			

1. *Notes*: obtained by OLS estimation. Standard errors are clustered at the student level.
2. *** Significant at the 1 percent level
3. ** Significant at the 5 percent level
4. * Significant at the 10 percent level

Supplementary Table 3: Gender differences in starting performance and in sustaining performance during the test by topic. The PISA 2009.

CNT	Diff. in reading starting level	Standard error	Diff. in reading during the test	Standard error	Diff. in science and math starting level	Standard error	Diff in science and math during the test	Standard error
ALB	0.1104	0.0078***	0.0176	0.0114	0.0308	0.0071***	-0.0059	0.0107
ARE	0.0507	0.0070***	0.0359	0.0069***	-0.0104	0.0072	0.0215	0.0072***

ARG	0.0244	0.0072***	0.0185	0.0121	-0.0199	0.0068***	-0.0102	0.0111
AUS	0.0400	0.0045***	0.0343	0.0061***	-0.0348	0.0047***	0.0389	0.0068***
AUT	0.0185	0.0060***	0.0387	0.0086***	-0.0841	0.0064***	0.0307	0.0093***
AZE	0.0321	0.0070***	-0.0025	0.0106	-0.0173	0.0065***	0.0277	0.0103***
BEL	0.0044	0.0052	0.0295	0.0074***	-0.0597	0.0055***	0.0158	0.0083*
BGR	0.0791	0.0077***	0.0055	0.0113	-0.0177	0.0076**	0.0080	0.0112
BRA	0.0207	0.0034***	0.0425	0.0053***	-0.0368	0.0031***	0.0328	0.0048***
CAN	0.0413	0.0033***	0.0407	0.0047***	-0.0312	0.0034***	0.0294	0.0053***
CHE	0.0310	0.0045***	0.0385	0.0066***	-0.0482	0.0047***	0.0203	0.0071***
CHL	0.0068	0.0061	0.0346	0.0095***	-0.0496	0.0064***	0.0186	0.0095*
COL	0.0064	0.0053	0.0174	0.0087**	-0.0568	0.0051***	0.0247	0.0084***
CRI	0.0061	0.0064	0.0162	0.0104	-0.0535	0.0065***	0.0164	0.0104
CZE	0.0306	0.0061***	0.0317	0.0087***	-0.0554	0.0064***	0.0355	0.0097***
DEU	0.0324	0.0064***	0.0254	0.0095***	-0.0612	0.0067***	0.0264	0.0102***
DNK	0.0269	0.0064***	0.0346	0.0094***	-0.0431	0.0070***	0.0308	0.0105***
ESP	0.0389	0.0031***	0.0225	0.0046***	-0.0360	0.0033***	0.0184	0.0050***
EST	0.0609	0.0068***	0.0157	0.0098	-0.0163	0.0072**	-0.0010	0.0108
FIN	0.0803	0.0061***	0.0353	0.0089***	0.0001	0.0067	0.0133	0.0102
FRA	0.0204	0.0070***	0.0371	0.0109***	-0.0534	0.0070***	0.0085	0.0115
GBR	0.0260	0.0047***	0.0273	0.0065***	-0.0464	0.0048***	0.0323	0.0070***
GEO	0.1093	0.0079***	-0.0000	0.0122	0.0138	0.0073*	0.0143	0.0111
GRC	0.0547	0.0068***	0.0122	0.0109	-0.0404	0.0073***	0.0235	0.0116**
HKG	0.0218	0.0064***	0.0158	0.0092*	-0.0415	0.0071***	0.0067	0.0105
HRV	0.0311	0.0068***	0.0429	0.0094***	-0.0515	0.0075***	0.0089	0.0107
HUN	0.0183	0.0062***	0.0286	0.0096***	-0.0628	0.0068***	0.0267	0.0103***
IDN	0.0506	0.0061***	0.0014	0.0097	-0.0207	0.0057***	0.0277	0.0094***
IRL	0.0251	0.0097***	0.0479	0.0118***	-0.0387	0.0099***	0.0416	0.0126***
ISL	0.0752	0.0085***	0.0105	0.0126	-0.0131	0.0090	0.0157	0.0137
ISR	0.0448	0.0073***	0.0167	0.0107	-0.0362	0.0073***	0.0176	0.0104*
ITA	0.0296	0.0027***	0.0328	0.0040***	-0.0669	0.0029***	0.0265	0.0043***
JOR	0.0256	0.0170	0.0501	0.0093***	-0.0421	0.0170**	0.0468	0.0092***
JPN	0.0330	0.0059***	0.0121	0.0087	-0.0296	0.0060***	0.0145	0.0145
KAZ	0.0751	0.0065***	-0.0067	0.0100	0.0018	0.0063	-0.0005	0.0100
KGZ	0.0693	0.0068***	0.0005	0.0115	0.0065	0.0060	0.0059	0.0102
KOR	0.0328	0.0068***	0.0163	0.0086*	-0.0255	0.0077***	0.0223	0.0103**
LIE	-0.0042	0.0248	0.0357	0.0401	-0.0833	0.0271***	0.0281	0.0435
LTU	0.0850	0.0071***	0.0256	0.0104**	-0.0055	0.0074	0.0233	0.0111**
LUX	0.0563	0.0078***	0.0078	0.0115	-0.0541	0.0078***	0.0390	0.0117***
LVA	0.0589	0.0070***	0.0366	0.0103***	-0.0202	0.0074***	0.0331	0.0113***
MAC	0.0302	0.0062***	-0.0003	0.0089	-0.0341	0.0067***	-0.0081	0.0101
MDA	0.0691	0.0069***	0.0058	0.0105	0.0073	0.0068	0.0003	0.0107
MEX	0.0183	0.0022***	0.0111	0.0036***	-0.0499	0.0023***	0.0267	0.0037***
MLT	0.0491	0.0223**	0.0437	0.0128***	-0.0395	0.0225*	0.0550	0.0141***
MNE	0.0703	0.0074***	0.0115	0.0105	-0.0377	0.0071***	0.0392	0.0104***
MUS	0.0265	0.0078***	0.0120	0.0099	-0.0353	0.0077***	0.0186	0.0105*
MYS	0.0457	0.0070***	0.0128	0.0106	-0.0128	0.0065*	0.0139	0.0100
NLD	0.0088	0.0059	0.0245	0.0091***	-0.0441	0.0062***	0.0169	0.0098*
NOR	0.0625	0.0071***	0.0509	0.0107***	-0.0201	0.0079**	0.0389	0.0120***
NZL	0.0570	0.0082***	0.0295	0.0295***	-0.0207	0.0088**	0.0336	0.0121***
PAN	0.0152	0.0074**	-0.0038	0.0117	-0.0446	0.0066***	0.0174	0.01060
PER	0.0122	0.0064*	-0.0059	0.0109	-0.0388	0.0058***	0.0067	0.0100
POL	0.0639	0.0070***	0.0313	0.0105***	-0.0152	0.0077**	0.0176	0.0117
PRT	0.0406	0.0060***	0.0320	0.0087***	-0.0350	0.0063***	0.0256	0.0096***
QAT	0.0314	0.0092***	0.0442	0.0077***	-0.0159	0.0094*	0.0224	0.0071***
QCN	0.0304	0.0056***	0.0208	0.0084**	-0.0306	0.0063***	0.0296	0.0097***
QHP	0.0070	0.0120	0.0081	0.0182	-0.0564	0.0095***	0.0614	0.0145***
QTN	0.0174	0.0087**	0.0143	0.0119	-0.0146	0.0070**	0.0061	0.0097
QVE	0.0210	0.0091**	0.0039	0.0144	-0.0297	0.0088***	-0.0170	0.0141
ROU	0.0125	0.0066*	0.0138	0.0099	-0.0688	0.0066***	0.0201	0.0101**
RUS	0.0655	0.0069***	0.0184	0.0104*	-0.0092	0.0072	0.0080	0.0113
SGP	0.0381	0.0069***	0.0043	0.0094	-0.0215	0.0075***	0.0159	0.0106
SRB	0.0399	0.0061***	0.0110	0.0091	-0.0586	0.0069***	0.0467	0.0102***
SVK	0.0652	0.0073***	0.0146	0.0103	-0.0337	0.0077***	0.0173	0.0114
SVN	0.0276	0.0060***	0.0488	0.0085***	-0.0683	0.0064***	0.0491	0.0094***
SWE	0.0570	0.0076***	0.0406	0.0112***	-0.0155	0.0081*	0.0415	0.0123***
TAP	0.0610	0.0065***	0.0058	0.0088	-0.0080	0.0068	0.0050	0.0098
THA	0.0602	0.0058***	0.0050	0.0084	-0.0066	0.0062	0.0109	0.0094
TTO	0.0645	0.0073***	0.0298	0.0113***	-0.0177	0.0069**	0.0416	0.0107***
TUN	0.0251	0.0068***	0.0134	0.0112	-0.0438	0.0060***	0.0210	0.0098**
TUR	0.0508	0.0066***	0.0098	0.0097	-0.0334	0.0066***	0.0242	0.0101**
URY	0.0462	0.0062***	0.0205	0.0105*	-0.0420	0.0066***	0.0273	0.0106**

USA	0.0244	0.0070***	0.0284	0.0100***	-0.0376	0.0071	0.0259	0.0106**

Notes: obtained by OLS estimation. Standard errors are clustered at the student level.

*** Significant at the 1 percent level

** Significant at the 5 percent level

* Significant at the 10 percent level

Supplementary Table 4: Regression of the gender difference in sustaining performance during test on the gender difference in dynamic inputs during test.

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Gender difference in time during the test	0.0843***	0.109***			0.0666***	0.0892***
	(4.44)	(3.28)			(2.88)	(2.81)
Gender difference in actions during test			0.00174**	0.00398***	0.000986	0.00282**
			(2.62)	(3.38)	(1.30)	(2.12)
Gender difference in time at the start		0.0134				0.0251
		(0.31)				(0.67)
Gender difference in actions at the start				0.000162		-0.0000534
				(0.44)		(-0.17)
Its interaction for time		-0.258				-0.176
		(-0.57)				(-0.47)
Its interaction for actions				-0.000182		-0.000163
				(-1.52)		(-1.31)
Constant	0.0167***	0.0153***	0.0194***	0.0192***	0.0182***	0.0170***
	(17.29)	(5.54)	(14.36)	(8.28)	(12.87)	(5.48)
N	58	58	58	58	58	58
Adjusted R^2	0.203	0.183	0.132	0.150	0.227	0.205

Notes: obtained by OLS estimation. Standard errors are clustered at the student level.

*** Significant at the 1 percent level

** Significant at the 5 percent level

* Significant at the 10 percent level