The Gender Gap in Math: What are we Measuring?

Silvia Griselda
Bocconi University*

December 30, 2022

Job Market Paper
Latest version available here

Abstract

Standardized tests are widely used to compare and select students and candidates, and by policy-makers as measures of human capital. But, what if the testing technologies used in these tests are not gender-neutral? Can the formats of these tests end up reinforcing education inequalities? I employ data from the largest standardized test in the world, and reveal that the gender gaps in performance largely depend on the format of exams students are randomly allocated to. Exams with an additional 10 percentage points of multiple-choice questions inflate women's under-performance in mathematics by 0.025 standard deviations and male under-performance in reading by 0.035 standard deviations. I document that multiple-choice questions create a cognitive load on students with a low level of self-efficacy, like women in math, which affect students' level of effort and performance in subsequent parts of the exam.

Keywords: human capital, achievement gap, gender gap, mathematics, standardized assessments, multiple-choice
JEL Classification: I21, I24, J24

*AXA Research Lab on Gender Equality, Bocconi University. I am also thankful for the invaluable comments from Anna Adamecz, Ghazala Azmat, Francesca Borgonovi, David Byrne, Caterina Calsamiglia, Alexia Delfino, Marina Della Giusta, Maria Recalde, Nicolas Salamanca, Sofia Trommlerova, Yanos Zylberberg, and all the participants at seminars and conferences that provided useful comments for this project. This paper was previously circulating with the title "Different Questions, Different Gender Gap: Can the Format of Questions Change the Gender Gap in Mathematics". I acknowledge financial support from the University of Melbourne's FBE Doctoral Program Scholarship for this research. All errors are my own. Email: silvia.griselda@unibocconi.it.

1 Introduction

Performances on standardized tests have long-lasting consequences for students and reveal important information to policymakers. Each year, students around the globe take standardized exams to enter colleges or obtain scholarships that determine their career trajectories and lifetime earnings (Doty et al., 2022). The ACT, SAT[1], GRE, and GMAT are only a few examples of standardized tests which shape students' educational and economic opportunities. Similarly, governments and policymakers often use standardized assessments to measure educational outputs and compare student performance across countries and time periods (Hanushek and Kimko, 2000; Schoellman, 2012).

On these tests, large educational inequality persists. On the one hand, women underperform men in mathematics and science, resulting in lower female representation in STEM disciplines and occupations, and contributing to the gender wage gap (Kahn and Ginther, 2017; Dickerson et al., 2015; McNally, 2020).[2] On the other hand, men underperform women in reading and humanities, a trend that seems to be contributing to the higher school dropout rate among males (Lundberg, 2020).

Can the testing process itself explain, at least partially, these large gender differences in test performance? Although standardized tests are widely used, a recent debate is arising regarding their effectiveness at measuring individual differences in knowledge. Indeed, differences in test performance by socioeconomic status seem to be driven by inequities in the testing process, rather than by underlying differences in competencies or knowledge (Miller et al., 2014; Goodman et al., 2020; Duquennois, 2022). This paper provides new insight into this growing debate by studying the effectiveness of standardized tests in measuring differences in performance across a new dimension: gender. Considering the key role that standardized tests play for students and educators, it is crucial to understand if there are stable gender gaps across different testing methods.

This paper examines whether gender differences in performance can be explained by the common use of multiple-choice questions in standardized tests. Multiple-choice questions are largely employed in tests, as they are often seen as objective, low-cost, and easy to implement

[1]Recently, the Board of Regents at the University of California voted against the use of SAT exams for granting admission. This decision followed criticism regarding how standardized tests may be biased against certain groups.

[2]STEM disciplines are increasingly demanded and better paid compared to other occupations, therefore, women's lower entrance into these occupations contributes to the earnings difference across genders. In addition, women's lower entrance rate in STEM disciplines can have important consequences for the level of productivity of the economy, as greater gender balance has been associated with a higher level of innovation and productivity (Hong and Page, 2004).

on a large scale (Frederiksen, 1984).[3] Yet, multiple-choice questions require students to choose the correct answer and rule out incorrect ones which are often referred to as *distractors*. A student's ability to rule out distractors depends not only on his level of competence but also on their level of self-efficacy, namely how confident he is on his knowledge in a particular topic (Lindner et al., 2014).[4] Answering multiple-choice questions is more cognitively costly for students with lower levels of self-efficacy than answering closed-response questions, since the presence of distractors misleads respondents from their preferred option, negatively affecting their current and future performance (Gierl et al., 2017).

I reveal that the use of multiple-choice questions in testing reinforces existing educational inequalities and provides an upper bound of gender differences in performance. Using data from one of the largest standardized assessments worldwide, the PISA test, I document that females' under-performance in mathematics and males' under-performance in reading are inflated among students who sit an exam with a larger proportion of multiple-choice questions.[5] I exploit the random variation in the format of tests assigned to students sitting the PISA test. Within countries and schools, PISA assigns different exam booklets to different students. These booklets aim to assess the same level of proficiency but contain a different proportion of multiple-choice (with no penalty for wrong-response) and closed-response questions (which require a short answer, a number of a word). Consequently , students can be assigned to tests with multiple-choice questions composing as low as 20% or as high as 70% of the test. I show that this variation in the format of the test plays a crucial role in students' overall performances and has spillovers on all questions. In mathematics, a 10 percentage point increase in the proportion of multiple-choice questions increases the gender gap in performance in favor of males by 0.025 of a standard deviation, about one-quarter of the overall mathematics gender gap.[6] On the contrary, in reading, a 10 percentage point increase in the proportion of multiple-choice questions increases males' under-performance in

[3]Multiple-choice questions are particularly used in mathematics or quantitative exams, where they often represent the large majority of questions. Several universities in the US and around the world employ Scholastic Aptitude Tests (SATs) and Graduate Record Exams (GREs) to determine students' admission to undergraduate and graduate programs. The mathematics and quantitative sections of these tests contain more than 75% and 50% of multiple-choice questions respectively. See https://collegereadiness.collegeboard.org/pdf/official-sat-study-guide-about-math-test.pdf and https://e-gmat.com/blogs/gre-exam-pattern.

[4]In finance-related multiple-choice questions, female students are more likely than males to answer the "I do not know" option, even when they knew the correct answer (Bucher-Koenen et al., 2016).

[5]Previous literature has focused on multiple-choice tests with different penalties for wrong responses or different stakes (Coffman and Klinowski, 2020). In this paper, I analyze exams that use multiple-choice and other formats of questions for the same assessment.

[6]The effect of reducing the share of multiple-choice questions in mathematics on female performance is comparable to an increase in teacher quality of one-quarter of a standard deviation (Rivkin et al., 2005), or a decrease in a class size of one student (Angrist and Lavy, 1999).

reading by 0.035 of a standard deviation, about one-fifth of the overall gender gap in reading. Following the estimates of Hampf et al. (2017) a 10 percentage points increase in the share of mathematics multiple-choice questions could potentially results in a decreasing in earning for female by 0.6%.

I document that answering multiple-choice questions creates a cognitive load on students with low levels of topic specific self-efficacy, such as female students in mathematics. This cognitive load affects both students' engagement levels, and their subsequent performance. Using time data and information on omission rates, I construct a proxy for *inattentive students*, by looking at those who omit questions even if they have enough time left to answer, and/or answer questions too rapidly.[7] I show that mathematics exams that rely heavily on multiple-choice questions affect the level of effort of male and female students. For exams with a larger share of multiple-choice questions, females, who on average have lower mathematics self-efficacy become more disengaged in the test, omitting answers or answering too-rapidly.

To investigate the cognitive load caused by the presence of distractors on multiple-choice questions, I look at the performance of students on mathematics sections randomly placed after a section with a larger share of multiple-choice questions. I observe that gender differences in mathematics performance in a given section increases (decreases) in favor of males, among students who previously faced a section of the exam with a greater share of multiple-choice (closed-response) questions. In particular, a 10 percentage point increase in the proportion of mathematics multiple-choice questions increases gender differences in performance in the following sections by 0.013 SD.

The relationship between extensive sets of options and cognitive load has been investigated by the marketing and financial literature (Schwartz and Ward, 2004; Kida et al., 2010). This relationship is known in the literature as the *"Paradox of Choice"* and refers to the idea that for individuals who are uncertain about their choice, choosing among a set of alternatives can result in confusion, frustration, and anxiety. The higher the level of confidence, the easier it is to eliminate alternative hypotheses with a lower cognitive effort. Therefore, in people who lack self-efficacy and who are more likely to feel remorse about excluding an alternative option, the paradox of choice manifests more frequently. In different domains, males and females exhibit different levels of self-efficacy. Boys have about 0.30 SD higher self-efficacy in mathematics than females.[8] On the other side, girls tend to be more confident in reading

[7] To answer a question appropriately, people should read it and think carefully about the answer. Therefore, answering questions without enough time to read or think about the answer can be considered a sign of inattention.

[8] In 2012, PISA surveyed students on their level of mathematics self-efficacy. The survey uses several items to assess mathematics self-efficacy, defined as a belief about own ability to complete a task. OECD (2014b) explains in detail how self-efficacy is measured.

and the humanities (Lundberg, 2020; Bordalo et al., 2019; McNally, 2020). Thus, a higher share of multiple-choice questions can result in different cognitive loads for males and females in mathematics and reading.

The higher cognitive load required by students who lack self-efficacy to answer multiple-choice questions not only includes cognitive efforts to rule out distractors, but also the negative impact of regrets and feedback on future performance. Indeed, it is not uncommon for test takers to experience regret when answering multiple-choice items, as several possible answers are available (Merry et al., 2021; Kruger et al., 2005). The feeling of regret occurs more often in students with low self-efficacy (Johnson et al., 2021). Secondly, multiple-choice items allow for *unintended corrective feedback*, namely students could realize that their computation is incorrect if the answer they come up with doesn't fit the alternative options (Bridgeman, 1992). This negative unintended feedback can create a sense of frustration, especially among students with lower self-efficacy, which could negatively impact efforts and future test performance (Dweck et al., 1978; Spencer et al., 1999; Good et al., 2003; Machina and Siniscalchi, 2014).

I investigate the role of self-efficacy in explaining the differential impact of multiple-choice questions in mathematics and reading, by using a survey-measure of self-efficacy and looking at the heterogeneous impact of question difficulty. First, I document that in tests with more multiple-choice (compared to closed-response) questions, the performance gap between students with high and low self-efficacy increases. Secondly, the impact of self-efficacy on multiple-choice questions should vary with the difficulty of the questions (Steele, 1997). For easy multiple-choice questions, students should have no problem ruling out incorrect alternatives. But, for female students who lack mathematics self-efficacy, ruling out distractors may be more challenging in hard questions. I show that the gender gap in performance is similar in easy multiple-choice and easy closed-response questions. But, when it comes to hard questions, the gender gap in performance is much wider in multiple-choice questions than in closed-response questions. I show that the relationship between format and gender is driven neither by questions' characteristics, such as their cognitive domain or context, nor by differences in writing skills.

Can these results be generalized to high-stake exams? While I cannot give a final answer, I show that my results remain consistent across countries with different test stakes. Overall, PISA is a low-stake exam for students. Yet different countries perceive and consider the test differently. In general, Asian or North European countries take the PISA test into high consideration, putting considerable effort to perform well. On the contrary, students in Middle Eastern countries tend to perform poorly due to their lower engagement in the test (Zamarro et al., 2019). I show that the relationship between format and gender differences in

performance does not depend on the stake of PISA. Male and female students' performance in mathematics is considerably affected by the format of the test, both in high-stakes and low-stakes settings.

I contribute to several stands of literature. First, I contribute to the literature on the effectiveness of standardized tests (Freedle, 2010; Borghans et al., 2016; Borgonovi et al., 2021). Recently, several papers have documented that differences in performance on these tests are driven not only by the difference in cognitive ability or knowledge but rather by differences in students' economic backgrounds (Goodman et al., 2020; Duquennois, 2022) and sociocultural status (Dobrescu et al., 2021). This paper provides new insights into this debate, by showing that even apparently fair and blind standardized tests that rely heavily on multiple-choice items can reinforce gender inequalities in education.

The second contribution of my paper is to shed light on a new mechanism through which large inequalities arise and persist in education. Females' under-performance in mathematics and males' under-performance in reading have been associated with culture and social norms (Guiso et al., 2008; Nollenberger et al., 2016), as well as with teacher stereotypes (Carlana, 2019). In this paper, I suggest a new mechanism behind large education inequalities: the testing process. Standardized assessments often consist primarily of multiple-choice questions. This is the first paper that documents how these questions, compared to other formats, provide an upper-bound for the gender gaps in performance. Third, I contribute to the literature on the role of confidence in explaining gender differences in educational and economic outcomes. Gender differences in competitiveness have been associated with the under-enrollment of women in STEM-related tracks (Goulas et al., 2022) and STEM-related occupations (Niederle and Vesterlund, 2007; Niederle et al., 2011), as well as gender differences in self-promotion and salary negotiation (Reuben et al., 2012; Exley and Kessler, 2022; Biasi and Sarsons, 2022). This paper suggests a new way for confidence and self-efficacy to influence gender differences in economic outcomes: by affecting students' performance on standardized tests. Women's underperformance in mathematics is most influenced by the share of multiple-choice questions in countries with higher mathematics self-efficacy among males. On the contrary, mathematics performance is stable across tests' formats when males and females have similar levels of mathematics self-efficacy. Fourth, my results contribute to the literature on female performance in multiple-choice assessments. Multiple-choice tests with a penalty for answering incorrectly discriminate against women (Baldiga, 2014; Riener and Wagner, 2017; Coffman and Klinowski, 2020).[9] I provide evidence that multiple-choice ques-

[9]Women tend to be more risk-averse and less confident in the correctness of their responses. Therefore, when negative marking is applied, they are more likely to skip questions than men, even conditional on underlying knowledge. Women's higher omission rates negatively impact their performance, thus increasing

tions harm girls' mathematics performance even in a context where penalties for answering incorrectly do not apply. As a consequence, women's under-performance in multiple-choice questions is not only driven by risk aversion, but also by confidence. Moreover, by comparing performance across different formats, I can document the spillover effect of multiple-choice questions on other formats and their impact on efforts and future performance.

2 Data: the Program for International Student Assessment

This paper uses data from the Program for International Student Assessment (PISA). PISA is an international standardized test administered by the Organization for Economic Co-operation and Development (OECD) to 15-year-old students in more than 60 countries (OECD, 2014b). The survey takes place every three years since 2000, and with over half a million students taking part, PISA is now the biggest international large-scale assessment. The test is designed to compare 15-years-old students' performance across countries and over time in three domains: mathematics, reading, and science.[10]

The results of PISA test have enormous impact on countries educational policies and national assessments. Indeed, PISA is considered a rather unique and reliable instruments that policy makers have to internationally benchmarking performance and changes over time (Breakspear, 2012). The so called *"PISA shock"*, which refers to the set of educational policies implemented by the German government after the surprisingly low PISA results, is probably the most notable example of how PISA has an impact on countries educational policies (Waldow, 2009).[11] Nevertheless, PISA can be consider a low-stakes exam for students, as students' performance on PISA test has no direct consequences on any educational outcomes. Yet, there are large variation across countries in the stake of PISA (Akyol et al., 2021).

The population sampling follows a two-stage stratified design. Firstly, schools of 15-years-old students are randomly selected with a probability proportional to the size of the school. Within each sampled school, students are randomly selected with equal probability. In total, approximately 150 schools and 5,250 students per country participate in PISA.

the gender gap in favor of men.

[10] PISA is performed every three years. Each year one domain is assessed in depth. In 2012 mathematics was the main domain, while in 2015, science was considered the main domain, while mathematics and reading were minor domains. This means that all students answer at least one section related to the main domain, and provide non-cognitive and attitudinal information regarding that particular domain.

[11] After the release of the PISA 2000 results, Germany received lower-than-expected results, which led to the introduction of important reforms in the educational system, such as the introduction of national standards and further support for disadvantage and immigrants students (Ertl, 2006; Niemann, 2010).

Up until wave 2012, the PISA test was paper-pencil. In 2015, computer-based exams were administered for the first time as the main mode of assessment.[12] In this paper, I employ data from waves 2012 and 2015 (for this last waves I focus only on students who complete the computer-based assessment)[13]

PISA contains information about students' demographics, home, and family background characteristics. Students' demographics information includes students' gender, SES status, parental education, and occupational level, language, immigration background, age in months, and grade level. In addition, PISA includes schools' background information, as well as their organizational and educational provision. My main sample consists of about 500,000 students surveyed either in 2012 or 2015, who answer at least one mathematics cluster and for whom information regarding the gender of the students, parental education, and occupation are available.

Figure 1a shows the timeline of the PISA test. The total assessment last approximately three and a half hours. The formal exam is designed as two-hour tests, both for the paper-based and computer-based assessment. The exam combines four 30-minutes sections, each one assessing a particular domain, mathematics, reading, or science.[14] At the end of the first two sections students are entitled to a short 5 minutes break. Students answer a 35-minute questionnaire at the end of the formal assessment. This questionnaire collects information about students attitudes, beliefs and non-cognitive skills.

Different groups of students receive different exam booklets, chosen among a pool of 409 different ones.[15] These booklets are different ordered combinations of mathematics, reading, and science sections. Booklets are assigned to students randomly, and contain different formats of questions.[16]

PISA employs three different formats of questions: multiple-choice items, where students need to select the correct answer among a set of possible ones; closed-response items, where students need to answer with a limited and concise response; and open-response items, where students can provide a full and extensive answer, with no constrain on the length of

[12] 58 countries complete PISA 2015 in computer-assessment mode. Only 15 countries use paper-based assessment, as they did not have the resource needed for computer-based testing (OECD, 2017b).

[13] I use the results from 2012, the last wave where mathematics was considered the main domain, and 2015, the first year in which students complete computer-based assessments. The advantage of computer-based assessment data is that it includes time undertaken by students in each task. In the 2009 wave and prior waves, the number of different booklets was really low. In 2018, the test was computer-adaptive. In computer-adaptive assessments, questions are not randomly assigned to students, but rather each receives questions that are tailored to his previous performance.

[14] Therefore, depending on the exam, some students can face more than one section of a specific domain.

[15] The set of questions assigned to students is called booklets, even if students answer a computer-based test.

[16] OECD (2014b, 2017b) explain in details the random assignment.

the response. Figures A1, A2, and A3 display examples of the three formats of questions. Depending on the exam booklets they receive, students could sit an exam with a small or large proportion of each format. Figure 2 show the variation in the proportion of different questions by booklet in mathematics (Figures A4.a and A4.b display the variation in reading and science respectively). As consequence, students could be randomly assigned to an exam with a proportion of multiple-choice questiosn as high as 72% or as low as 17%. PISA uses *number-right scoring*, namely, there is no penalty for answering incorrectly multiple-choice items. Even if this scoring rule, at least implicitly, encourages guessing, some students penalize themselves by failing to respond to every item. In the computer-based assessment students needs to answer the questions in the order they are provided, and they do not receive any feedback about their performance at any time during the test.[17] Questions of different formats can happen at any order within the section. Figure 1b provide an example of the sequences of questions in two different mathematics sections.

Since PISA re-administer some of their items in several waves, I do not observe the exact prompt for most of the items. Nevertheless, I have information regarding the item format (multiple-choice, closed-response, and open response), cognitive domain (mathematics, reading, and science), question difficulties, and domain-specific information for each of the questions (i.e. content, context, and cognitive process).

In my analysis, I focus mainly on mathematics performance, for several reasons. First, mathematics is the domain with a higher variation for all three formats of questions: multiple-choice, closed- and open-response questions. Figure 3 shows the proportions of formats in different domains. In Mathematics there is a similar proportion of questions for all three formats, and closed-response represent, on average, 31% of the questions. On the contrary, the proportion of close-response questions is below 10% in Reading and below 5% in Science. Second, mathematics is the domain where the gender gap in favor of boys is wider. Both in 2012 and 2015, in most countries, boys outperform girls in mathematics, especially among top-achieving students (Peña-López et al., 2016). In contrast, girls perform better than boys in reading, even if the gap has narrowed compared to previous waves. Boys and girls perform similarly in science, but boys show greater aspiration towards science-related careers.

PISA test is administered in each country by trained test administrators, who ensure the security and confidentiality of the assessment material, as well as a fairly, impartially, and uniform assessment of the test (OECD, 2014b, 2017b). The trained test administrators cannot be teachers of participating students. At the beginning of the exam, each student is allocated to a desk with the assigned booklet in the year 2012 and to a workspace with

[17]This changed in 2018 when students completed a computer-adaptive test, where the types of questions students receive depend on their performance on previous questions.

a computer and received a unique login form in 2015. During the exam, a staff member of the school monitors the students. In 2012, booklets were randomly assigned to coders.[18] In 2015, for the computer-based assessment, multiple-choice and closed-response questions are computer-coded. Open-response questions are marked by recruited and trained coders. Each coder receives a set of 100 randomly selected student responses.

3 The Impact of Exam Format on Performance

In this section, I document how the format of exam impact gender differences in mathematics performance. I take advantage of the random assignment of exams booklets with varying proportion of multiple-choice, closed and open-ended question. I find that exams that rely heavily on multiple-choice (closed-response) questions display greater (smaller) women underperformance in mathematics.

I estimate the following model:

$$\begin{aligned} Y_{isb} = \beta_0 + \beta_1 \text{Female}_{is} + \beta_2 \text{Female}_{is} \cdot \text{Prop. of MCQs}_b \\ + \beta_3 \text{Female}_{is} \cdot \text{Prop. of ORQs}_b + X'_{is}\gamma + b_b + s_s + \varepsilon_{isb} \end{aligned} \tag{1}$$

where Y_{isb} represents the proportion of questions answered correctly (standardized), by student i, attending school s, who receives booklet b.[19]

The main explanatory variables include a dummy for females and its interaction with the proportion of mathematics multiple-choice questions featured in the booklet b. The

[18] Coders were provided with detailed criteria for coding, as well as many examples of acceptable and unacceptable responses.

[19] In my analysis, I use the proportion of correct questions as an outcome, but PISA employs the Item Response Theory to estimate students' performance. In particular, PISA uses a combination of two-parameters Rasch Model and generalized partial credit model. Item Response Theory is particularly appropriate to scale students' responses when different groups of students receive a subset of questions from the total questions pool. It characterizes students' performance as the probability of answering correctly a question (among the entire pool of questions, not only the ones they answer) based on their proficiency. In other words, students' performance can be compared across all participating students, even if different subgroups answer different sets of questions. Performances are reported thought of ten plausible values, drawn from a distribution that combines Item Response Theory to latent regression using demographics students' information.The plausible values were randomly drawn from the distribution of ability estimates that could reasonably be assigned to a student, and the mean of the plausible values should be equal to the expected posterior (EAP) estimator (OECD, 2017b). As my identification strategy compares students who face the same text booklet, plausible values are not the appropriate outcomes for my analysis. Students' raw score, defined as the proportion of correct questions answered by each student, represents a cleaner measure of individual students' performances in the questions received. Indeed, this score depends only on the questions students face, and it is not affected by how similar students perform in other questions or booklets.

model includes the interaction between a dummy for female and the proportion of open-response questions featured in the booklet. Therefore β_2 documents how gender differences in performance varies for students sitting exams that relies more or less heavily on multiple-choice, rather than closed-response questions (holding the proportion of open-ended question unchanged). 5[20]

The model controls for several students' characteristics, such as students' age, grade, migration status, parental education level, and occupational status. The model includes booklet FE, b_b, and school FE, s_s.[21] By including booklet FE, I compared within groups of students gender variation in performance across booklets with similar average characteristics. Standard errors are clustered at the school level.[22] I estimate the model separately for the year 2012 and year 2015.

3.1 Validity of the Randomization

In order to interpret the estimate of β_2 in model (1) as the effect of multiple-choice salience on gender difference in performance, exam formats need to be orthogonal to students observable and unobservable characteristics. PISA explicitly state that booklets are randomly rotated among test-takers within each school (OECD, 2014a, 2017a).[23] I provide additional evidence of the validity of the randomization in Table 1. Overall, there is no correlation between students' observable characteristics and the percentage of mathematics multiple-choice, closed-response, and open-response questions that students receive in the test.[24]

3.2 Exam Format and Gender Difference in Performance

Table 2 shows the estimates for model (1) for the years 2012 (columns 1-3) and 2015 (columns 4-6). Columns 1 and 4 do not include booklet FE, and the proportion of mathematics

[20]Boys under-perform girls in mathematics. As open-ended questions requires students to write down their answer, the relation between open-ended questions and gender difference in mathematics performance could be driven by girls advantage in reading. Comparing multiple-choice and closed-response questions allows to distinguish between the impact of alternative option from the impact of boys under-performance in writing.

[21]Booklet FE accounts for average booklet's characteristics, such as questions sequences among others. The proportion of multiple-choice and open-response questions are perfectly collinear with booklet FEs, and therefore cannot be estimated when booklet FEs are included.

[22]The estimates remain significant when standard errors are clustered at the country level.

[23]Applying a rotated design to exam questionnaires allows for more material to be tested. Some questions were taken from previous waves, some others introduced as new in 2012 and 2015.

[24]Only the grade of the students is correlated with the proportions of questions of different formats in 2012. The correlation between grades and females is 0.05. In general, girls are more likely to be in a higher school year compared to boys. This result is consistent with the literature on grade repetition and early school entry. Nevertheless, model (1) include students' grade as control.

multiple-choice questions is included as an explanatory variable. Columns 2 and 5 report the results without including school FE, while columns 3 and 6 include a booklet, school, and year FEs. Columns 3 and 6 are the preferred specifications. The inclusion of school FE in columns 3 and 6 do not significantly impact the estimate for β_2 in model (1).

Estimates for β_2 in Table 2 imply that an increase (decrease) in the proportion of multiple-choice (closed-response) questions by 10 percentage points inflates (decreases) women underperformance by about 0.027 standard deviations. Because students can be randomly assigned to an exam with a share of multiple-choice questions that range from 17% to 72% (as reported in Figure 2), the magnitude of the effect format of exam is not small. This effect is comparable to a decrease in teacher quality of one-quarter of a standard deviation (Rivkin et al., 2005), or an increase in a class size of one student (Angrist and Lavy, 1999).

The estimate of baseline gender gap captured by the β_1 coefficient decreases to 0.025 (0.014) in 2012 (2015) compared to the raw gender gap of 0.100 (-0.140) and become statistically insignificant for exam that contains only closed-response questions.[25]

Interestingly, increasing the proportion of open-response question by reducing the proportion of closed-response questions has no significant impact on gender difference in performance, as indicated by the estimates of the interaction between the proportion of open-response questions and the dummy for female.

Figure A5 shows the distributions of predicted standardized scores in mathematics, for female and male students, for different shares of mathematics multiple-choice questions. Males' distribution shifts to the right as the share of multiple-choice questions increases from 17 to 70%. The more multiple-choice questions students face, the more likely males are to receive a top score in mathematics.

3.3 Spillover Effects on Other Questions

In the previous section, I document that the format of exam have important consequences for the gender differences in overall mathematics performance. In this section I show that the proportion of multiple-choice questions that students face in the test affect how they score on multiple-choice questions, but also how students perform on closed-response, and open-response questions. In particular, I estimate model (1) using as outcomes the score on multiple-choice, closed-response, and open-response questions separately. Table 3 shows the results.

Exams that relies heavily on multiple-choice and less on closed-response questions (con-

[25]Because the PISA data do not contains any exam with only closed-response questions, the estimate for β_1 represents an out-of-sample estimation.

trolling for the share of open-responded questions), differentially decrease female performance in closed- and open-response questions.[26] In particular, an increase by 10 percentage points in the proportion of multiple-choice (closed-response) questions increases (decreases) females' under-performance in closed-response questions by 0.014 of a standard deviation, and in open-response by 0.023 of a standard deviation. Figure 4 provides a graphical representation of women under-performance in different formats of questions by the format of exam. While students' score in multiple-choice questions is independent and stable across different means of testing, the scores in closed- and open-response questions varies significantly depending on the type of exam that students receive.

This is the first paper that document how a particular format, multiple-choice questions, can affect students performance in other items. In the following section I investigate the mechanisms behind these findings.

4 Mechanisms

The exam format can affect students' performance in several ways. Firstly, various formats of questions may capture different skills in which male and female students may be better at. Secondly, different questions require distinct answering strategies, which allow students to perform differently even with similar levels of knowledge or skills. The following paragraphs explore the two channels in more detail.

4.1 Format and Questions' Characteristics

Gender differences by formats could hide gender gaps in areas of competencies or knowledge. Indeed, multiple-choice and closed-response questions may capture different areas of competencies, in which males and females students perform differently (Willingham and Cole, 2013). For example, males could have an advantage in MCQs because these questions are more likely to test students' ability to recall definitions, skills in which males could be better than females. Table A1 shows that indeed MCQs assess different competencies. Multiple-choice questions are more likely to assess specific content, context, and cognitive process compared to closed- and open-response questions. Previous educational literature has documented that boys outperform girls in questions that require students to identify a reasonable response, while girls outperform boys in questions that requires interpretations, problem solving, and analysis (Reardon et al., 2018).

[26] Different formats of questions can occur at any point in the exam, and there are no specific section of particular formats.

To investigate whether differences in questions' areas of knowledge play a role in explaining my results, I exploit the rich set of information available about the content, context, and cognitive process of each question. In particular, I compared questions designed to assess the same areas of competencies, with the same difficulty, that only vary in their format: multiple-choice, closed- and open-response. I estimate the following model:

$$\begin{aligned} \mathrm{Y}_{isq} =& \gamma_1 + \gamma_2 \mathrm{Female}_{is} + \gamma_3 \mathrm{MCQ}_q + \gamma_4 \mathrm{Female}_{is} \cdot \mathrm{MCQ}_q + \\ & \gamma_5 \mathrm{ORQ}_q + \gamma_6 \mathrm{Female}_{is} \cdot \mathrm{ORQ}_q + X'_{is}\Gamma + Z'_q\Theta + s_s + \varepsilon_{isq} \end{aligned} \tag{2}$$

where Y_{isq} indicates one of the four outcomes of the mathematics question q faced by student i in school s; MCQ_q and ORQ_q are dummy variables indicating whether question q is a multiple-choice or open-response as opposed to a closed-response ones. The model includes students' controls X_{is}, such as student's age and grade attended, student's immigration status, parental education and occupation, and an index of home possession. Z_q represents a vector of question characteristics: content, context, cognitive process that students need to employ to answer the questions, and question difficulty.[27] The model includes school FE, s_s.

I am mainly interested in the coefficients γ_4 and γ_6. The first capture how gender differences vary in multiple-choice compared to closed-response questions, the latter, how gender differences vary in open-response compared to closed-response questions.

I consider four main outcomes: (1) whether a question was answered correctly by students, (2) whether a question was answered correctly by students, conditional on answering, (3) whether the question was skipped or omitted by the student, and (4) the time undertook to answer a question (only for 2015).

Table A2 reports the results of this investigation for the years 2012 and 2015. Overall, girls are about 2.5% less likely to answer correctly a closed-response question than boys, but this gap increases to 5% for multiple-choice questions (columns 1-2 and 4-5). Because the model controls for questions' characteristics, boys' advantage in multiple-choice questions is not driven by gender differences in competencies.

Males higher performance in multiple-choice could be driven by their higher probability of answering correctly, or by lower omission rate. I find no evidence of the latter. Columns 3 and 6 of Table A2 report the difference in probability of skipping questions by gender and format. There is no gender difference in skipping behavior by format in the year 2012. In

[27] Question difficulty is measured with the percentage of students that answer the question incorrectly in the field trial.

2015, girls are less (rather than more) likely to skip multiple-choice questions compared to boys, even if the effect is small in magnitude.

Figures 5a and 5b plot the predicted performance of male and female students, by formats, using estimation results from model 2. From the pictures is clear that in both waves, males performed better than females in all formats of mathematics questions. Yet, the gender difference in performance is significantly bigger in multiple-choice, than in other questions. Because the estimated model account for questions area of assessment, the gender differences by format plotted in these figures cannot be driven by gender differences in area of competencies.

4.2 Format and Test-taking Ability

Answering multiple-choice and closed-response questions involves different test taking ability While answering closed-response items require students to provide an answer, answering multiple-choice questions requires choosing among a set of alternative options.

In particular, answering multiple-choice questions can be modeled as a process where each student assigns to each possible answer a prior (defined as the subjective probability that a given answer is correct), and ranks answers based on these priors. The distribution of these priors, and therefore students' ability to rule out alternative answers, depends on the features of the question (e.g. difficulty, number of alternatives, etc.), but also on student's cognitive and non-cognitive traits (e.g. level of knowledge, self-efficacy, uncertainty and ambiguity aversion) (Fok et al., 2012; Machina and Siniscalchi, 2014).

4.2.1 Modeling Students' Answering Strategies

This section introduces a simple model to describe students' answering strategies. Consider two formats of questions: multiple-choice (MCQ) and closed-response (CRQ).[28] Consider also two groups of students, males, M, and females, F.

Performance on a question q, of student i is defined as P_{iq}. To answer closed-response questions, students just need to come up with an answer, therefore students' performance on a CRQ is going to depend on student i's level of knowledge, K_i, and question characteristics, such as question difficulty or content of assessment, C_q. Therefore, students performance in any given CRQs by males and females students can be written as:

$$P_{M,CRQ} = p(K_M, C_{CRQ}) \qquad P_{F,CRQ} = p(K_F, C_{CRQ})$$

[28] Here I do not consider open-response questions, where students writing could play a role.

Because PISA random assigned exams to students, males and females students face the same set of questions C_{CRQ}. Therefore the only gender difference in performance in CRQs need to be driven by gender differences in level of knowledge.

$$P_{M,CRQ} \neq P_{F,CRQ} \quad \leftrightarrow \quad K_M \neq K_F$$

Students performance in MCQs can be written as a function of students level of knowledge, K_i, question characteristics (e.g. question difficulty, and number of alternative) C_q, and students ability to rule out incorrect answers (e.g. which depend on students self-efficacy, uncertainty, ambiguity and feedback aversions) B_i. These non-cognitive traits are likely to affect the distribution of students' priors, as well as students' ability to rank distinctively and precisely alternative answers.

Therefore, students' performances in a given MCQ by males and females students can be written as:

$$P_{M,MCQ} = m(K_M, C_{MCQ}, B_M) \qquad P_{F,MCQ} = m(K_F, C_{MCQ}, B_F)$$

Also in this case, males and females are assigned to the same set of MCQs questions, C_{MCQ}, therefore the gender differences in performance in MCQs could be driven both by gender differences in knowledge, K_i, but also gender differences in students' test-taking abilities B_i.

$$P_{M,MCQ} \neq P_{F,MCQ} \quad \leftrightarrow \quad K_M \neq K_F \quad \text{and, or} \quad B_M \neq B_F$$

In the PISA tests, students are randomly assigned to different exams with a different share of multiple-choice and closed-response questions. Therefore the gender differences in students' knowledge is constant across MCQs and CRQs.

This means that the only driver of gender differences in performance by format is students test-taking abilities B_i. In other words, the only reason why women under-perform boys in mathematics are higher in multiple-choice questions than in closed-response ones need to be driven by gender differences in test-taking abilities.

$$[P_{M,CRQ} - P_{F,CRQ}] \neq [P_{M,MCQ} - P_{F,MCQ}] \quad \leftrightarrow \quad B_M \neq B_F$$

The ideal experiment to investigate gender differences in answering strategy, and the role of students non-cognitive traits on the and mis-estimation of priors, would observe students answers, as well as the distribution of priors for each students and for each question.[29] This is not possible in PISA dataset. While the advantage of PISA data relies in the real life assessments with students from more than 60 countries around the world, it does not to elicit students priors. Moreover, I do not have measures of students ambiguity or uncertainty aversion.

Yet, I provide evidence of the role of test-taking ability, and in particular student-self efficacy in two ways. First, I analyze gender differences in performance by format in easy and hard questions. In easy multiple-choice questions, ruling out incorrect answers should be straightforward. Therefore, as B_i play less of a role in easy MCQs, the distribution of performance in these questions should be similar to the the distribution of performance in closed-response one. On the contrary, for hard multiple-choice questions, ruling out incorrect answer can be more challenging, and distractors could affect students cognitive load. Figure 6 displays the gender gaps by question difficulty. The gender gap is the same for easy multiple-choice and closed-response questions. On the contrary, for hard questions, females' under-performance are bigger in closed-response compared to multiple-choice questions.

4.3 The Role of Self-efficacy in explaining Test-taking Ability

Using a survey measure of self-efficacy, I examine whether multiple-choice questions increase cognitive load among students with lower self-efficacy.[30] First, I include self-efficacy in the model, and second, I investigate how the share of multiple-choice questions in the exam differentially affects the performance of students with low or high mathematics self-efficacy. Column 2 of Table 4 reports the estimates of model (1) controlling students' mathematics self-efficacy (column 1 reports the estimates without controlling for it for reference). Once math self-efficacy is controlled for, the effect of format on gender differences in performance decreases from 0.312 to 0.201 and becomes less precise. Table 5 report the estimates for model 1, when the dummy female is replaced by survey measured of self-efficacy. The gap on performance between students with high versus low mathematics self-efficacy is double in

[29] This measurement technique is known in the educational literature as *confidence weighting*, and aims to measure what respondents believe is a correct answer, and the degree of certainty toward the correctness of these beliefs (Ebel, 1965).

[30] In 2012, students were asked several questions in the PISA questionnaire, aiming to assess their mathematics self-efficacy. In particular, students were asked how confident they were on performing several mathematics tasks, such as"Solving an equation like 3x+5= 17", or "Finding the actual distance between two places on a map with a 1:10 000 scale". OECD (2014b) provides information on how self-efficacy is measured and how is the index calculated.

in multiple-choice compared to closed response questions.

In addition, I investigate whether the effect of multiple-choice questions is stronger in countries where males have greater mathematics self-efficacy compared to females. Figure 7 plots the correlation between gender differences in self-efficacy, controlling for mathematics performance, and the impact of the share of multiple-choice questions on female performance. In countries where boys have significantly greater self-efficacy than girls, a larger proportion of multiple-choice questions in the exam displays greater women-underperformance. In other words, women's underperformance in mathematics varies widely depending on the testing technologies in countries where the gender gap in self-efficacy is bigger.

In the next sections, I provide evidence that sections of exam with higher proportion of multiple-choice questions have a negative effect on subsequent performance and students' engagement level.

4.4 The Impact of Multiple-choice and Following Sections

In the previous sections, I document that an exam with higher proportion of mathematics multiple-choice questions has an effect on mathematics performance in closed- and open-response questions. In this part, I analyze whether the format of a mathematics section of the exam has an effect on how male and female students perform in subsequent mathematics sections. In particular, I exploit the random assignment of exam booklets with different mathematics sections, and measure students' performance in a specific section, after they receive an higher or lower proportion of multiple-choice questions in the previous section. Figure 9a provides an intuition of the identification. I compare gender differences in performance for two groups of students receiving the same mathematics section A: the first group of students has already answered a mathematics section with a large proportion of multiple-choice questions; the second group of students has already faced a mathematics section with a small proportion of multiple-choice questions in the previous section of the exam. If the format has not spillover effect on following parts of the same exam, we should expect the gender differences in performance in section A to be similar across the two groups of students. Columns 1 and 2 of Table 6 shows the results. A 10% higher proportion of multiple-choice questions increases females' under-performance in subsequent mathematics sections by 0.013 of a standard deviation (column 1). Similarly, a section with 10% higher proportion of multiple-choice questions increase females' likelihood to omit a question by 0.2% compared to males, about 2% of omission rate (column 2).

Students need to answer each section in a specific order. Once they move forward to a section, they cannot go back to the previous ones. Therefore, students performance in

a particular mathematics sections should not be affected by the format of the subsequent ones. To investigate whether this is the case I run a placebo analysis as display in Figure 9b. In particular, I compare the gender differences in performance in section X among students who are randomly assigned to a subsequent section with more or less multiple-choice versus closed-response questions. Columns 3 and 4 of Table 6 report the results. The format of a subsequent section is not correlated with gender differences in performance and omission rate on a previous section.

5 Exam Format and Students' Level of Engagement

In the previous section, I show that the format of an exam section impacts males and females performance in subsequent sectionss. Females perform differentially worse (less worse) than males after they faced a mathematics sections with higher proportion of multiple-choice (closed-response) questions. In this section, I analyze the effect of format of exam on students engagement during the test. [31]

First, I identify *disengaged students*, namely those students who exert low level of effort in the exam. Secondly, I investigate whether the format of exam students receive has an effect on their probability of becoming *disengaged students*.

5.1 Identify Disengaged Students

There are two paths to identify disengaged students: looking at omission rate and employing time response data to analyze students' rapid response (Akyol et al., 2021; Zamarro et al., 2019).

Omission rate is crucial to identify disengaged students. PISA does not employ negative marking for incorrect multiple-choice questions. Therefore, students should always have an incentive to guess multiple-choice questions when they do not know the answer, and skipping could be considered a sign of students' low level of attention. While for multiple-choice questions answering (including guessing) is always a weakly dominant strategy, for close- and open-response questions time constraints could lead students to omit some answers. Although students have a 30 minutes time limit in each section, time is not a binding constraint for most students.[32] Consequently, omitting any questions could be interpreted as a sign of

[31] For this section, I only use 2015 data. Indeed, in 2015 students completed the computer-based exam. This allows to track the time they took to answer each question.

[32] On average, students take around 18 minutes to answer each mathematics section and 90% of students finish the mathematics section in 27 minutes.

students' low engagement.[33]

In addition, time response data can be employed to investigate students' level of engagement in the test. Indeed, in order to answer a question, a minimum amount of time is needed to read and understand the question. Therefore, too little time spent on a question could be considered a sign of low effort. Figure A6 plot an example of the distribution of time spent on a specific questions by student of a given country. The bimodal distribution seems to be composed by two separate populations: one of students answering in normal time and centered around 1.18 minutes, and another one of students answering too-rapidly. It is therefore possible to compute a threshold that divides the two different normally distributed sub-populations. I employ a Gaussian mixture model to compute country-question specific thresholds to identify questions answered too rapidly. Therefore, I define the time spent r_{qij} on an item q by student i, in country j, as too rapid if the time is less than a threshold τ_{qj}.[34] This country-question specific threshold allows me to distinguish between questions answered rapidly or in normal time. It does not provide yet a measure of disengagement or guessing behavior. At this point I have two question-specific criteria: omission, and rapid response questions.

I consider a student disengaged in the test if he/she 1] does not answer 3 or more questions, even if there is enough time remaining in the section (i.e. at least 5 minutes); 2] answer 3 or more questions rapidly, and he/she scores lower in rapidly-answered questions than in questions answered in normal time.[35] These criteria allow me to consider about 10% of students as non-serious, a rate similar to the ones found by Akyol et al. (2021).

Table A3 shows the summary statistics of disengaged boys and girls. Consistent with previous literature, boys are significantly more likely than girls to be identified as disengaged students (the proportion of disengaged boys is 9.39%, while the proportion of disengaged girls is 8.51%, and the difference is statistically different from zero).

Someone may argue that omitting an unknown answer should not be interpreted as a sign of a low level of attention during the tests, but rather a sign of seriousness and reliability. Table A3 provides the summary statistics for the proportion of students who belong to the first criterion (omission rate) and the second criterion (rapid response) and used these two

[33]Akyol et al. (2021) document that skipping behavior increases with question order within the exam sections. They argue that, as there is no correlation between questions' difficulty and questions' position within the section, this pattern is consistent with students skipping questions as a sign of reducing exam effort.

[34]The average threshold is similar to the one in Akyol et al. (2021). Using this threshold, approximately 5 percent of questions are defined as too-rapid response.

[35]Students who are really smart can answer questions quickly because they can read, think, and respond faster than their peers. This would not imply they are exerting low effort. However, if this were the case, the proportion of correct answers would not be lower than when questions are answered normally.

as additional measures of students level of engagement. There is no statistical difference between the number of boys and girls that omit 3 or more questions. Yet, boys are more likely to be identified as disengaged because they answer too rapidly 3 or more questions.

To study whether the proportion of multiple-choice questions has a differential effect on students' engagement in the exam by gender, I estimate model 1 using a dummy variable for disengaged students as an outcome. Table 7 shows the results. The first three columns show the results for disengaged students using the above-mentioned definition. Columns 1 and 2 estimate the specification 1 using OLS, without and with school FE, while column 3 uses Logit regression, and reports the marginal effect. The estimate for the female coefficient in column 3 implies that girls are overall 1.2% less likely to be identified as disengaged than boys. Nevertheless, when the proportion of multiple-choice question features in the exam increases, girls become differentially more disengaged than boys. The estimate for the interaction between females and the proportion of mathematics multiple-choice questions is bigger in magnitude than the estimate for females. This means that a 10 percentage point increase in the proportion of multiple-choice questions can reverse the gender gap in students' engagement level.

5.2 Possible Confounding Factors

In the above sections, I document a relationship between the share of multiple-choice questions and the gender difference in students' performance in mathematics. Nevertheless, booklet characteristics that are correlated with the share of multiple-choice questions and differentially affect boys' and girls' performance could bias my results. In this section, I provide evidence that my results remain robust to possible omitted variables.

First, sections with a larger share of multiple-choice questions could have a larger proportion of questions of a particular context, content, or cognitive process. The PISA test tries to make each booklet as comparable as possible in terms of questions' context, content, or cognitive process. Yet, I further investigate whether the proportion of multiple-choice in each section is correlated with a larger proportion of particular question characteristics. Tables A4a, A4b, and A4c show that indeed the proportion of multiple-choice questions is correlated with the number of questions in the booklet assessing occupational and societal context, as well as questions testing uncertainty and space and shape.

Table A5 displays the results from model 1 once the interaction term between the dummy female and the proportion of questions assessing different contents (columns 2 and 6), different contexts (columns 3-4 and 7-8) are included. The estimates for the interaction term between the dummy female and the proportion of multiple-choice questions in the booklet remain

negative and significantly different from zero in the majority of the specifications.[36]

Second, the number of questions featured in the mathematics section could affect students' performance and be correlated with the proportion of multiple-choice questions. The higher proportion of multiple-choice questions could be correlated with the total number of overall mathematics questions. This could bias my results, as the number of questions can affect students' performance and the time to respond differentially for boys and girls. Column 7 of Table A2 indicates that women take overall more time to answer MCQs and ORQs than boys. Table A6 shows that on average each mathematics section of the test has 14 questions. The sections with a higher proportion of multiple-choice questions do not have a significantly larger number of questions overall. In addition, the proportion of multiple-choice questions in the section is not correlated with average questions difficulty, as well as the number of easy, medium-hard, and hard questions in the cluster.

Third, multiple-choice questions may appear in a specific position within the cluster. The position of a question affects students' performance (Schweizer et al., 2009). Moreover, the item position effect varies across males and females. In particular, girls are better than boys to sustain their performance through an exam (Balart and Oosterveen, 2019). Therefore, my estimated effect could be biased if multiple-choice questions systematically appear in a specific position of the mathematics section (e.g. at the beginning of the end of the section). This could be an issue in 2015 PISA tests, where students are required to answer the questions in the order they are provided. Table A7 shows that there is no relationship between multiple-choice questions and the question position within the mathematics section for 2015 data. Hence, my results are unlikely to be driven by the correlation between the format of the item and its position within the cluster.

5.3 Heterogeneity by the Stake of the Test

For students, PISA is not a high-stakes exam. As a consequence, students' incentive to perform well might be minimal, and varying across gender and countries.[37]

[36]Including the proportion of questions assessing particular context, content, and cognitive process creates a problem of imperfect multicollinearity. For this reason, the preferred specification only includes the interaction term between female and proportion of multiple-choice questions.

[37]Several studies document gender differences in the level of effort exerts in low-stakes examination (Buser et al., 2014; Azmat et al., 2016; Gneezy et al., 2019). On one side, the performance of men increases more than women when the stakes of the test increase (Buser et al., 2014; Balart and Oosterveen, 2019). On the other side, women seem to perform better when they sit for low-stakes examinations (Ors et al., 2013; Azmat et al., 2016; Cai et al., 2019). In particular, while girls exert a similar level of effort in low- and high-stakes tests, boys exert much less effort than girls in low-stakes examinations. A different stand of the literature has documented that variation in students' level of effort explains much of the variation in the performance across gender and countries. In particular, using PISA data Zamarro et al. (2019) and Akyol et al. (2021)

While the stake of PISA does not represent a threat for my analysis, it undermines the external validity of my findings. Indeed, it is not possible to generalize the finding of this paper to high-stake examinations. Yet, different countries consider the PISA test and its stake differently. Generally speaking, students of Asian countries take PISA test really seriously, while students in middle-east countries do not perform as well as expected in the test (Borgonovi and Biecek, 2016; Zamarro et al., 2019). I provide suggestive evidence of the validity of my results, and show the heterogeneity of my analysis by the stake that each country assign to the PISA test.

First, I rank countries in terms of the average proportion of disengaged students (defined as in section 5). The proportion of disengaged students could be considered a proxy for the perceived level of the stake of PISA in the country (Gneezy et al., 2019). Figure 10 plot the proportion of disengaged students and the estimates for the interaction between the dummy female and the proportion of mathematics multiple-choice questions by countries. While the effect differ by country, the relationship between format of exam and gender differences in performance remain stable across countries that face different stakes for PISA.

5.4 Exam Format and Immigration and Socio-economic Background

Someone may argue that gender is only one of the dimensions in which multiple-choice questions can affect the performance of the different groups of students.[38] I find no evidence that multiple-choice questions differentially affect the performance of students by immigration status or socio-economic background. Table A8 shows that the share of multiple-choice questions does not differentially affect the performance of native and immigrant students (columns 1 and 3), and of student of different socio-economic background (columns 2 and 4).

6 Heterogeneous Effects

6.1 Frequency of Assessments

The main findings of this paper highlight that boys have an advantage in answering multiple-choice questions and the gender gap in mathematics performance may be inflated in favor of boys for exams that use a larger proportion of multiple-choice questions.

provide evidence that accounting for student effort explains between about 30 percent of the differences in performance across countries. Similarly, Balart and Oosterveen (2019) reveals that gender differences in students' effort could increase the gender gap in mathematics performance by 6 times in favor of boys.

[38] Immigrants have slightly higher self-efficacy than native students (the difference is about 0.05 of a standard deviation), while students with high socioeconomic status are on average more self-efficacy than those of low socioeconomic status.

One possible affirmative action to increase girls' ability to perform well in mathematics multiple-choice questions is training students to answer multiple-choice questions. I investigate the possible efficacy of this intervention by analyzing the heterogeneous effect of the share of multiple-choice questions by the frequency with which students are assessed using standardized tests in their school. I use the information provided by PISA in 2015 regarding how often each school assesses students using mandatory standardized assessments.

Table A9 shows the results. The estimate for the interaction between females and the proportion of multiple-choice questions is negative for students who are assessed using standardized assessments less than five times per year. On the contrary, the negative effect of multiple-choice question on girls' performance halved and becomes insignificant among students who are assessed using standardized assessments monthly.[39] This provides evidence that boys' advantage in mathematics multiple-choice assessments could potentially disappear when students are used to standardized assessments.

6.1.1 Mathematics Performance

In this subsection, I analyze the heterogeneous effect of the proportion of multiple-choice questions on students' performance. Figures A7a and A7b display the estimates for female and its interaction with the proportion of multiple-choice questions students receive, as in model 1, by decile of mathematics performance.

The format of questions has a differential impact on the gender gap in performance with respect to the performance distribution. In 2012, the proportion of multiple-choice questions that students receive has no differential effect for outcomes of males and females in the bottom 10 percentile, and only marginal effect for male and female students in the top 10 percentile. The proportion of multiple-choice questions has the greatest effect on the gender difference in performance among students in the middle of the performance distribution. The findings are similar in 2015, where the main effect of multiple-choice questions of gender difference in performance is driven by students in the top 70 and 80 percentile of the performance distribution.

6.2 Reading and Science

The main analysis of this paper focuses on mathematics. As shown in Figure 3, mathematics is the domain with a higher variation for all three formats of questions: multiple-choice, closed- and open-response questions. On the contrary, reading and science domains do not

[39] Figure 11 displays the marginal effect of females on performance when the proportion of multiple-choice questions varies from 0 to 1, by frequency of the standardized assessment.

have sufficient variation to provide valid estimate to the relationship between format and gender differences in performance.

Nevertheless, in this section, I extend the analysis to reading and science, and extend my results to reading and science domains. According to previous literature, girls have greater self-efficacy in reading, while boys have greater self-efficacy in science (Mostafa, 2019). Therefore, we can expect exams with a higher percentage of multiple-choice reading questions to inflate the gender gap in reading in favor of females. On the contrary, the science gender gap should increase in favor of males if more multiple-choice questions are included in the exam. I document that multiple-choice heavy exams can reinforce existing inequality in educational performance. Table 8 shows the results. Females outperform boys in reading by 0.23 SD. Yet, this difference highly depends on the format of exams that students faced. A 10 percentage point higher (lower) proportion of reading multiple-choice (closed-response) questions increases males under-performance in reading by 0.035 SD, about 15 % of the raw gap (column 1). Figure A8 shows the distributions of predicted standardized scores in reading, for female and male students, for different shares of reading multiple-choice questions. Females' distribution shifts on the right of the males' one as the share of multiple-choice questions increases from 29 to 69%. In science, where males perform better than females, a 10 percentage points greater (lower) proportion of science multiple-choice (closed-response) questions increase women under-performance in science by 0.022 SD, about one quarter of the raw gender gap in Science (column 2). Figure A9 shows the distributions of predicted standardized scores in science, for female and male students, for different shares of science multiple-choice questions. Males' distribution shifts on the right of the females' one as the share of multiple-choice questions increases from 52 to 75%.

7 Discussion and Conclusion

Standardized tests are widely used for measuring human capital, determining university admissions, providing licenses, and certifying students, as well as assessing the effectiveness of policies and educational interventions. In this paper, I document that the gender differences in performance measured in these tests are not technology-independent, but rather educational inequalities vary with the format of the test. The most common form of question employed in standardized assessments, multiple-choice questions, is associated with greater gender disparities in educational performance than closed-ended questions.

Importantly, this analysis compared multiple-choice and closed-response questions, formats that are both marked by computer. Due to this, increasing the number of closed-response questions in a test does not entail an increase in costs.

Students' performances are both directly and indirectly affected by multiple-choice questions. On one hand, on exams that feature a large number of multiple-choice questions, females under-perform in mathematics and males under-perform in reading, compared to similar exams with a larger share of closed-response questions. Females under-perform in mathematics and males under-perform in reading, on exams that feature a large number of multiple-choice questions compared to similar exams with a large share of closed-response questions. Additionally, multiple-choice questions affect students' level of effort and how they perform in subsequent questions. For example, the level of effort a woman puts into a test decreases as the proportion of multiple-choice questions increases. Furthermore, female students who have previously faced an exam section with a high proportion of multiple-choice questions perform worse in the following mathematics section.

An investigation of the mechanisms reveal that multiple-choice questions capture not only the knowledge of a student but also other individual skills, such as their level of confidence and self-assessment abilities. When faced with a set of alternative answers, students with low self-assessment abilities have difficulty choosing the correct answer. In turn, this negatively impacts students' performances on subsequent sections of the exam.

While students' levels of confidence about their ability in a given topic could be worthwhile measuring, as it may be an important predictor of future earnings, it is important to highlight that this trait is influenced by the environment in which individuals are raised, and the level of stereotypes they are exposed to (Guiso et al., 2008; Nollenberger et al., 2016).

It is unlikely that we will ever have a single undisputed method to measure a student's cognitive ability, which is probably a combination of several factors. Yet, it is important to understand whether there are constant gender gaps in performance among different testing technologies, and what dimension of human capital each exam measures. This is the main scope of this paper. This study is the first to demonstrate that gender differences in performance are unstable across exam formats, and rather that the most common types of questions, multiple-choice questions, affect student performance and effort in subsequent sections. To reduce inequality in educational and economic opportunities, it is essential to have a clear understanding of how different test technologies affect human capital measures and the reasons for these discrepancies.

References

Akyol, P., K. Krishna, and J. Wang (2021). Taking pisa seriously: How accurate are low-stakes exams? *Journal of Labor Research 42*(2), 184–243.

Angrist, J. D. and V. Lavy (1999). Using maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly journal of economics 114*(2), 533–575.

Azmat, G., C. Calsamiglia, and N. Iriberri (2016). Gender differences in response to big stakes. *Journal of the European Economic Association 14*(6), 1372–1400.

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

Baldiga, K. (2014). Gender differences in willingness to guess. *Management Science 60*(2), 434–448.

Biasi, B. and H. Sarsons (2022). Flexible wages, bargaining, and the gender gap. *The Quarterly Journal of Economics 137*(1), 215–266.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2019). Beliefs about gender. *American Economic Review 109*(3), 739–73.

Borghans, L., B. H. Golsteyn, J. J. Heckman, and J. E. Humphries (2016). What grades and achievement tests measure. *Proceedings of the National Academy of Sciences 113*(47), 13354–13359.

Borgonovi, F. and P. Biecek (2016). An international comparison of students' ability to endure fatigue and maintain motivation during a low-stakes test. *Learning and Individual Differences 49*, 128–137.

Borgonovi, F., A. Choi, and M. Paccagnella (2021). The evolution of gender gaps in numeracy and literacy between childhood and young adulthood. *Economics of Education Review 82*, 102119.

Breakspear, S. (2012). The policy impact of pisa: An exploration of the normative effects of international benchmarking in school system performance.

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement 29*(3), 253–271.

Bucher-Koenen, T., R. Alessie, A. Lusardi, and M. Van Rooij (2016). Women, confidence, and financial literacy. *European Investment Bank*.

Buser, T., M. Niederle, and H. Oosterbeek (2014). Gender, competitiveness, and career choices. *The Quarterly Journal of Economics 129*(3), 1409–1447.

Cai, X., Y. Lu, J. Pan, and S. Zhong (2019). Gender gap under pressure: Evidence from china's national college entrance examination. *Review of Economics and Statistics 101*(2), 249–263.

Carlana, M. (2019). Implicit stereotypes: Evidence from teachers' gender bias. *The Quarterly Journal of Economics 134*(3), 1163–1224.

Coffman, K. B. and D. Klinowski (2020). The impact of penalties for wrong answers on the gender gap in test scores. *Proceedings of the National Academy of Sciences*.

Dickerson, A., S. McIntosh, and C. Valente (2015). Do the maths: An analysis of the gender gap in mathematics in africa. *Economics of Education Review 46*, 1–22.

Dobrescu, L., R. Holden, A. Motta, A. Piccoli, P. Roberts, and S. Walker (2021). Cultural context in standardized tests.

Doty, E., T. J. Kane, T. Patterson, and D. O. Staiger (2022, December). What do changes in state test scores imply for later life outcomes? Working Paper 30701, National Bureau of Economic Research.

Duquennois, C. (2022). Fictional money, real costs: Impacts of financial salience on disadvantaged students. *American Economic Review 112*(3), 798–826.

Dweck, C. S., W. Davidson, S. Nelson, and B. Enna (1978). Sex differences in learned helplessness: Ii. the contingencies of evaluative feedback in the classroom and iii. an experimental analysis. *Developmental psychology 14*(3), 268.

Ebel, R. L. (1965). Confidence weighting and test reliability. *Journal of Educational Measurement 2*(1), 49–57.

Ertl, H. (2006). Educational standards and the changing discourse on education: The reception and consequences of the pisa study in germany. *Oxford Review of Education 32*(5), 619–634.

Exley, C. L. and J. B. Kessler (2022). The gender gap in self-promotion. *The Quarterly Journal of Economics 137*(3), 1345–1381.

Fok, D., R. Paap, and B. Van Dijk (2012). A rank-ordered logit model with unobserved heterogeneity in ranking capabilities. *Journal of applied econometrics 27*(5), 831–846.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American psychologist 39*(3), 193.

Freedle, R. (2010). On replicating ethnic test bias effects: The santelices and wilson study. *Harvard Educational Review 80*(3), 394–404.

Gierl, M. J., O. Bulut, Q. Guo, and X. Zhang (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research 87*(6), 1082–1116.

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

Good, C., J. Aronson, and M. Inzlicht (2003). Improving adolescents' standardized test performance: An intervention to reduce the effects of stereotype threat. *Journal of Applied Developmental Psychology 24*(6), 645–662.

Goodman, J., O. Gurantz, and J. Smith (2020). Take two! sat retaking and college enrollment gaps. *American Economic Journal: Economic Policy 12*(2), 115–58.

Goulas, S., S. Griselda, and R. Megalokonomou (2022). Comparative advantage and gender gap in stem. *Journal of Human Resources*, 0320–10781R2.

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

Hampf, F., S. Wiederhold, and L. Woessmann (2017). Skills, earnings, and employment: exploring causality in the estimation of returns to skills. *Large-scale Assessments in Education 5*(1), 1–30.

Hanushek, E. A. and D. D. Kimko (2000). Schooling, labor-force quality, and the growth of nations. *American economic review 90*(5), 1184–1208.

Hong, L. and S. E. Page (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences 101*(46), 16385–16389.

Johnson, B. N., W. J. Kim, J. Blue, A. Summerville, and B. P. Kirkmeyer (2021). Gender differences in the functionality of regret on academic performance. In *2021 CoNECD*.

Kahn, S. and D. Ginther (2017). Women and stem. Technical report, National Bureau of Economic Research.

Kida, T., K. K. Moreno, and J. F. Smith (2010). Investment decision making: do experienced decision makers fall prey to the paradox of choice? *The journal of behavioral finance 11*(1), 21–30.

Kruger, J., D. Wirtz, and D. T. Miller (2005). Counterfactual thinking and the first instinct fallacy. *Journal of personality and social psychology 88*(5), 725.

Lindner, M. A., A. Eitel, G.-B. Thoma, I. M. Dalehefte, J. M. Ihme, and O. Köller (2014). Tracking the decision-making process in multiple-choice assessment: Evidence from eye movements. *Applied Cognitive Psychology 28*(5), 738–752.

Lundberg, S. (2020). Educational gender gaps. *Southern economic journal 87*(2), 416–439.

Machina, M. J. and M. Siniscalchi (2014). Ambiguity and ambiguity aversion. In *Handbook of the Economics of Risk and Uncertainty*, Volume 1, pp. 729–807. Elsevier.

McNally, S. (2020). Gender differences in tertiary education: what explains stem participation? Technical report, IZA Policy Paper.

Merry, J. W., M. K. Elenchin, and R. N. Surma (2021). Should students change their answers on multiple choice questions? *Advances in Physiology Education 45*(1), 182–190.

Miller, C., K. Stassun, et al. (2014). A test that fails. *Nature 510*(7504), 303–304.

Mostafa, T. (2019). Why don't more girls choose to pursue a science career?

Niederle, M. and L. Vesterlund (2007). Do women shy away from competition? do men compete too much? *The Quarterly Journal of Economics 122*(3), 1067–1101.

Niederle, M., L. Vesterlund, et al. (2011). Gender and competition. *Annual Review of Economics 3*(1), 601–630.

Niemann, D. (2010). Turn of the tide—new horizons in german education policymaking through io influence. In *Transformation of education policy*, pp. 77–104. Springer.

Nollenberger, N., N. Rodríguez-Planas, and A. Sevilla (2016). The math gender gap: The role of culture. *The American Economic Review 106*(5), 257–261.

OECD (2014a). *PISA 2012 Assessment and Analytical Framework*.

OECD (2014b). Pisa 2012 technical report.

OECD (2017a). *PISA 2015 Assessment and Analytical Framework*.

OECD (2017b). Pisa 2015 technical report.

Ors, E., F. Palomino, and E. Peyrache (2013). Performance gender gap: does competition matter? *Journal of Labor Economics 31*(3), 443–499.

Peña-López, I. et al. (2016). *PISA 2015 results (volume I): Excellence and equity in education*. Organisation for Economic Co-operation and Development, OECD Publishing.

Reardon, S. F., D. Kalogrides, E. M. Fahle, A. Podolsky, and R. C. Zárate (2018). The relationship between test item format and gender achievement gaps on math and ela tests in fourth and eighth grades. *Educational Researcher 47*(5), 284–294.

Reuben, E., P. Rey-Biel, P. Sapienza, and L. Zingales (2012). The emergence of male leadership in competitive environments. *Journal of Economic Behavior & Organization 83*(1), 111–117.

Riener, G. and V. Wagner (2017). Shying away from demanding tasks? experimental evidence on gender differences in answering multiple-choice questions. *Economics of Education Review 59*, 43–62.

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

Schoellman, T. (2012). Education quality and development accounting. *The Review of Economic Studies 79*(1), 388–417.

Schwartz, B. and A. Ward (2004). Doing better but feeling worse: The paradox of choice. *Positive psychology in practice*, 86–104.

Schweizer, K., M. Schreiner, and A. Gold (2009). The confirmatory investigation of apm items with loadings as a function of the position and easiness of items: A two-dimensional model of apm. *Psychology Science Quarterly 51*(1), 47.

Spencer, S. J., C. M. Steele, and D. M. Quinn (1999). Stereotype threat and women's math performance. *Journal of experimental social psychology 35*(1), 4–28.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American psychologist 52*(6), 613.

Waldow, F. (2009). What pisa did and did not do: Germany after the 'pisa-shock'. *European Educational Research Journal 8*(3), 476–483.

Willingham, W. W. and N. S. Cole (2013). *Gender and fair assessment*. Routledge.

Zamarro, G., C. Hitt, and I. Mendez (2019). When students don't care: Reexamining international differences in achievement and student effort. *Journal of Human Capital 13*(4), 519–552.

Table 1: Formats of Tests and Students' Characteristics

	Mathematics in 2012			Mathematics in 2015		
	(1) % of Multiple-Choice Questions b/se/se_sc	(2) % of Closed-Response Questions b/se/se_sc	(3) % of Open-Response Questions b/se/se_sc	(4) % of Multiple-Choice Questions b/se/se_sc	(5) % of Closed-Response Questions b/se/se_sc	(6) % of Open-Response Questions b/se/se_sc
Female	-0.060	0.046	0.014	0.049	-0.048	-0.002
	(0.043)	(0.026)*	(0.025)	(0.065)	(0.046)	(0.022)
	[0.053]	[0.034]	[0.028]	[0.075]	[0.053]	[0.025]
Age in Months	-0.028	0.005	0.023	0.084	-0.064	-0.020
	(0.073)	(0.047)	(0.040)	(0.114)	(0.081)	(0.040)
	[0.070]	[0.046]	[0.038]	[0.121]	[0.083]	[0.044]
Immigration Status:	-0.192	0.146	0.046	-0.044	0.061	-0.017
First-Generation	(0.143)	(0.093)	(0.078)	(0.205)	(0.146)	(0.071)
	[0.195]	[0.120]	[0.107]	[0.190]	[0.138]	[0.063]
Immigration Status:	-0.025	0.042	-0.017	0.073	-0.065	-0.007
Second-Generation	(0.124)	(0.076)	(0.069)	(0.174)	(0.124)	(0.061)
	[0.112]	[0.069]	[0.062]	[0.188]	[0.131]	[0.067]
Mother's Highest	-0.002	0.004	-0.001	-0.004	0.003	0.000
Education	(0.019)	(0.013)	(0.010)	(0.026)	(0.019)	(0.009)
	[0.020]	[0.012]	[0.010]	[0.030]	[0.021]	[0.010]
Father's Highest	0.006	-0.001	-0.005	-0.002	0.002	0.001
Education	(0.018)	(0.012)	(0.010)	(0.025)	(0.018)	(0.009)
	[0.016]	[0.009]	[0.010]	[0.027]	[0.018]	[0.010]
Highest Parental	-0.002	0.001	0.001	0.003	-0.001	-0.001
Occupational Status	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)**
	[0.001]	[0.001]	[0.001]	[0.002]*	[0.001]	[0.001]***
Home Possession	0.005	-0.010	0.006	0.010	-0.010	0.000
Index	(0.027)	(0.018)	(0.014)	(0.039)	(0.028)	(0.014)
	[0.029]	[0.020]	[0.015]	[0.031]	[0.023]	[0.011]
Student's Country of	0.000	-0.000	-0.000	-0.000	0.000	0.000
Birth	(0.000)	(0.000)	(0.000)**	(0.000)	(0.000)	(0.000)
	[0.000]	[0.000]	[0.000]*	[0.000]	[0.000]	[0.000]
Mother's Country of	-0.000	0.000	0.000	-0.000	0.000	0.000
Birth	(0.000)*	(0.000)*	(0.000)	(0.000)	(0.000)	(0.000)
	[0.000]**	[0.000]*	[0.000]	[0.000]	[0.000]	[0.000]
Father's Country of	0.000	-0.000	0.000	0.000	-0.000	-0.000
Birth	(0.000)	(0.000)*	(0.000)	(0.000)	(0.000)	(0.000)
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Obs	344,040	344,040	344,040	154,299	154,299	154,299
Mean Y	40.49	31.29	28.22	40.49	31.29	28.22
St Dev Y	13.43	8.86	6.94	13.43	8.86	6.94
Country FE	Yes	Yes	Yes	Yes	Yes	Yes
Booklet FE	No	No	No	No	No	No
Students' Controls	Yes	Yes	Yes	Yes	Yes	Yes
F Statistics	0.94	1.12	0.67	0.55	0.50	0.75
P-Value for Model	0.501	0.340	0.769	0.868	0.907	0.696

Notes: This table shows no significant relationship between the percentage of different formats of questions and students' observable characteristics. The outcome variables are the percentage (ranging from 0 to 100) of multiple-choice, closed-response, and open-response questions students receive in their exams. The explanatory variables include student's demographic characteristics, such as a dummy for females, the age of the student in months, the grade that the student is attending, and whether the student is a first or second-generation immigrant, as opposed to native. The specifications also include information about the parents, including their highest education level (measured by ISCED), their highest occupational status, and a summary index for their home possessions (which includes information such as the number of books in the house, the number of desks the students have to study, the internet connection, etc.). Columns 1-3 include data from the 2012 assessment, while columns 4-5 include data from the 2015 ones. Standard errors are clustered at the school level (in parentheses) or country level (in square brackets).

Table 2: Format of Test and Gender Gaps in Performance

	Std. Scores in Mathematics in 2012			Std. Scores in Mathematics in 2015			Tot Time (Mins)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	b/se/se_sc	b/se/se_sc	b/se/se_sc	b/se/se_sc	b/se/se_sc	b/se/se_sc	b/se/se_sc
Female	-0.044	-0.045	-0.025	0.099	0.100	-0.014	1.165
	(0.079)	(0.078)	(0.075)	(0.194)	(0.192)	(0.169)	(3.090)
	[0.098]	[0.095]	[0.095]	[0.134]	[0.135]	[0.121]	[1.821]
Prop. Math MCQs	-0.087			-0.221			
	(0.057)			(0.121)*			
	[0.159]			[0.139]			
Prop. Math ORQs	-0.355			-1.020			
	(0.109)***			(0.347)***			
	[0.445]			[0.540]*			
Female ×	-0.258	-0.257	-0.275	-0.340	-0.345	-0.266	0.450
Prop. Math MCQs	(0.085)***	(0.085)***	(0.083)***	(0.170)**	(0.168)**	(0.146)*	(2.690)
	[0.135]*	[0.133]*	[0.128]**	[0.122]***	[0.128]***	[0.116]**	[1.413]
Female ×	0.117	0.120	0.083	-0.461	-0.467	-0.195	3.967
Prop. Math ORQs	(0.169)	(0.169)	(0.161)	(0.486)	(0.482)	(0.433)	(7.838)
	[0.157]	[0.149]	[0.158]	[0.323]	[0.321]	[0.289]	[4.145]
Obs	349,951	349,951	349,951	159,539	159,539	159,539	159,468
Mean Y	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	65.94
St Dev Y	1.00	1.00	1.00	1.00	1.00	1.00	23.07
School FE	No	No	Yes	No	No	Yes	Yes
Booklet FE	No	Yes	Yes	No	Yes	Yes	Yes
Students' Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R-sq	0.25	0.25	0.28	0.23	0.23	0.52	0.43
Raw Gender Gap	-0.10	-0.10	-0.10	-0.14	-0.14	-0.14	2.98

Notes: This table presents the estimations of model (1), for the standardized score in mathematics in year 2012 (columns 1-3), the standardized score in mathematics (columns 4-6), and the total time undertook to complete the exam in 2015 (column 7) . The outcomes in columns 1-6 are the standardized proportions of correct questions in mathematics. Each specification controls for student's age in months, grade (compared to modal grade in the country), immigration status, parent highest level of education and occupation, and home possession index. The main explanatory variables are the proportions of multiple-choice and open-response questions (ranging from 0 to 1) and their interaction with the dummy for female students. Columns 1 and 4 do not include Booklet FE, while columns 2, 3, 5, 6, and 7 do (so that the proportion of MCQ and ORQ is captured by these FEs). Standard errors are clustered at school level in parentheses, and at country level in square brackets. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table 3: Format of Test and Performance in Other Questions

	(1) Multiple-choice Questions b/se/se_sc	(2) Closed-response Questions b/se/se_sc	(3) Open-Response Questions b/se/se_sc
Female	-0.221	-0.099	-0.008
	(0.033)***	(0.033)***	(0.032)
	[0.048]***	[0.049]**	[0.038]
Female ×	-0.002	-0.141	-0.226
Prop. Math MCQs	(0.035)	(0.042)***	(0.037)***
	[0.052]	[0.056]**	[0.043]***
Female ×	0.243	0.281	-0.034
Prop. Math ORQs	(0.076)***	(0.072)***	(0.071)
	[0.109]**	[0.101]***	[0.080]
Obs	509,490	509,490	509,490
Mean Y	-0.00	-0.00	0.00
St Dev Y	1.00	1.00	1.00
School FE	Yes	Yes	Yes
Booklet FE	Yes	Yes	Yes
Students' Controls	Yes	Yes	Yes
R-sq	0.23	0.23	0.24

Notes: This table presents the estimations of model (1) for the year 2012 and 2015, using as an outcome the standardized score only in multiple-choice (column 1), closed-response (column 2), and open-response questions (column 3). Each specification controls for student's age in months, grade (compared to modal grade in the country), immigration status, parent highest education, and occupational levels, home possession index, school and booklet FE. The proportions of multiple-choice and open-response questions range from 0 to 1. Standard errors are clustered at school level in parentheses, and at country level in square brackets. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table 4: Format and Gender Gap: Controlling for Mathematics Self-efficacy

	Std. Scores in Mathematics in 2012	
	(1)	(2)
	b/se/se_sc	b/se/se_sc
Female	-0.025	-0.108
	(0.075)	(0.094)
	[0.095]	[0.079]
Female × Prop. Math MCQs	-0.275	-0.114
	(0.083)***	(0.106)
	[0.128]**	[0.096]
Mathematics Self-Efficacy		0.118
		(0.052)**
		[0.056]**
Mathematics Self-Efficacy × Prop. Math MCQs		0.139
		(0.057)**
		[0.061]**
Obs	349,951	231,702
Mean Y	-0.00	-0.00
St Dev Y	1.00	1.00
School FE	Yes	Yes
Booklet FE	Yes	Yes
Students' Controls	Yes	Yes
R-sq	0.28	0.37
Raw Gender Gap	-0.10	-0.10

Notes: This table presents the estimations of model (1), using data from 2012, where survey measures of mathematics self-efficacy are available. Column 1 reports the estimation results as in column 3 of Table 2 for reference. Column 2 includes mathematics self-efficacy and the interaction between mathematics self-efficacy and the proportion of multiple-choice questions and open-response questions. Standard errors are clustered at the school level in parentheses, and country level in square brackets. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table 5: Differences in Performance by Format and Students' Levels of Self-efficacy

	Std. Scores in Mathematics in 2012	
	(1)	(2)
	Males	Females
	b/se/se_sc	b/se/se_sc
Mathematics Self-efficacy	0.163	0.151
	(0.031)***	(0.030)***
	[0.046]***	[0.054]***
Mathematics Self-efficacy × Prop.	0.176	0.208
Math MCQs	(0.035)***	(0.036)***
	[0.040]***	[0.038]***
Obs	113,928	117,715
Mean Y	0.05	-0.05
St Dev Y	1.03	0.97
School FE	Yes	Yes
Booklet FE	Yes	Yes
Students' Controls	Yes	Yes
R-sq	0.31	0.31

Notes: This table presents the estimations of model (1), where the dummy female is replaced by a measure of mathematics self-efficacy. The model includes data from 2012, where survey measures of mathematics self-efficacy are available. Column 1 reports the estimation including only male students, while Column 2 includes only females. Standard errors are clustered at the school level in parentheses, and country level in square brackets. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table 6: Format of Test and Subsequent and Previous Performance

	Subsequent Math Section		Previous Math Section	
	(1)	(2)	(3)	(4)
	Std. Score	Prop. of Skipped Questions	Std. Score	Prop. of Skipped Questions
	b/se/se_sc	b/se/se_sc	b/se/se_sc	b/se/se_sc
Female	-0.112	-0.003	-0.184	0.012
	(0.035)***	(0.006)	(0.031)***	(0.004)***
	[0.039]***	[0.006]	[0.036]***	[0.005]**
Female × Prop. Math	-0.132	0.020		
MCQs Previous Section	(0.052)**	(0.010)**		
	[0.055]**	[0.011]*		
Female × Prop. Math			0.017	-0.011
MCQs Following Section			(0.060)	(0.009)
			[0.065]	[0.010]
Obs	102,257	102,257	102,257	102,257
Mean Y	0.00	0.10	-0.00	0.08
St Dev Y	1.00	0.17	1.00	0.15
School FE	Yes	Yes	Yes	Yes
Booklet FE	Yes	Yes	Yes	Yes
Students' Controls	Yes	Yes	Yes	Yes
R-sq	0.40	0.27	0.39	0.28

Notes: This table presents the estimations of model (1), for the years 2012 and 2015. In columns 1 and 2 the outcome variables are the standardized score and the proportion of omitted questions in subsequent mathematics sections. The main explanatory variable is the interaction between the dummy female and the proportion of multiple-choice questions in the previous mathematics section. Figure 9a provides an intuition of the estimation strategy. In columns 3 and 4 the outcome variables are the standardized score and the proportion of questions that were omitted in previous mathematics sections. In this case, the main explanatory variable is the interaction between the dummy for female students and the proportion of multiple-choice questions in the following mathematics section. Figure 9b provides an intuition for the estimation strategy. Each specification controls for the student's age in months, grade (compared to modal grade in the country), immigration status, parents' highest level of education and occupation, home possession index, school, and booklet FE. Standard errors are clustered at the school level in parentheses, and country level in square brackets. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table 7: Format of Tests and Level of Efforts

	Inattentive Students			Students Omitting 3 or more Questions			Students Answering too rapidly 3 or more Questions		
	(1) OLS b/se/se_sc	(2) OLS b/se/se_sc	(3) Logit (dy/dx) b/se/se_sc	(4) OLS b/se/se_sc	(5) OLS b/se/se_sc	(6) Logit (dy/dx) b/se/se_sc	(7) OLS b/se/se_sc	(8) OLS b/se/se_sc	(9) Logit (dy/dx) b/se/se_sc
Female	-0.013	-0.011	-0.012	-0.004	-0.002	-0.005	-0.018	-0.016	-0.016
	(0.006)**	(0.006)*	(0.005)**	(0.005)	(0.005)	(0.004)	(0.005)***	(0.005)***	(0.005)***
	[0.006]**	[0.005]**	[0.005]**	[0.005]	[0.005]	[0.004]	[0.006]***	[0.005]***	[0.005]***
Female x Prop. of	0.027	0.024	0.023	0.018	0.016	0.020	0.026	0.024	0.021
Multiple-choice Math	(0.013)**	(0.013)*	(0.011)**	(0.011)*	(0.010)	(0.009)**	(0.011)**	(0.011)**	(0.011)*
Questions	[0.012]**	[0.012]**	[0.011]**	[0.010]*	[0.010]	[0.009]**	[0.012]**	[0.010]**	[0.011]*
Obs	120,307	120,307	120,307	120,307	120,307	120,307	120,307	120,307	120,307
Mean Y	0.09	0.09	0.09	0.05	0.05	0.05	0.06	0.06	0.06
St Dev Y	0.29	0.29	0.29	0.23	0.23	0.23	0.24	0.24	0.24
School FE	No	Yes	No	No	Yes	No	No	Yes	No
Booklet FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Students' Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table reports the estimates for model (1) using as outcomes: 1] whether a student is identified as disengaged as he did not answer 3 or more questions, even if there was enough time remaining in the cluster (i.e. at least 5 minutes), and/or he answered too rapidly 3 or more questions, and the proportion of correct questions answered too rapidly is lower than the proportion of correct questions answered in normal time. (columns 1-3); 2] whether a student did not answer 3 or more questions, even if there was enough time remaining in the cluster (columns 4-6); 3] whether a student answered too rapidly 3 or more questions, and the proportion of correct questions answered too rapidly is lower than the proportion of correct questions answered in normal time (columns 7-9). Columns 1-2-4-5-7-8 use OLS estimation, while columns 3-6-9 estimate the specification 1 using Logit model and report the marginal effects. Each specification includes student controls (student's age in months, grade (compared to modal grade in the country), immigration status, parent highest education, occupational levels, and home possession index), and school and booklet FEs. Standard errors are clustered at the school level in parentheses, and country level in square brackets. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

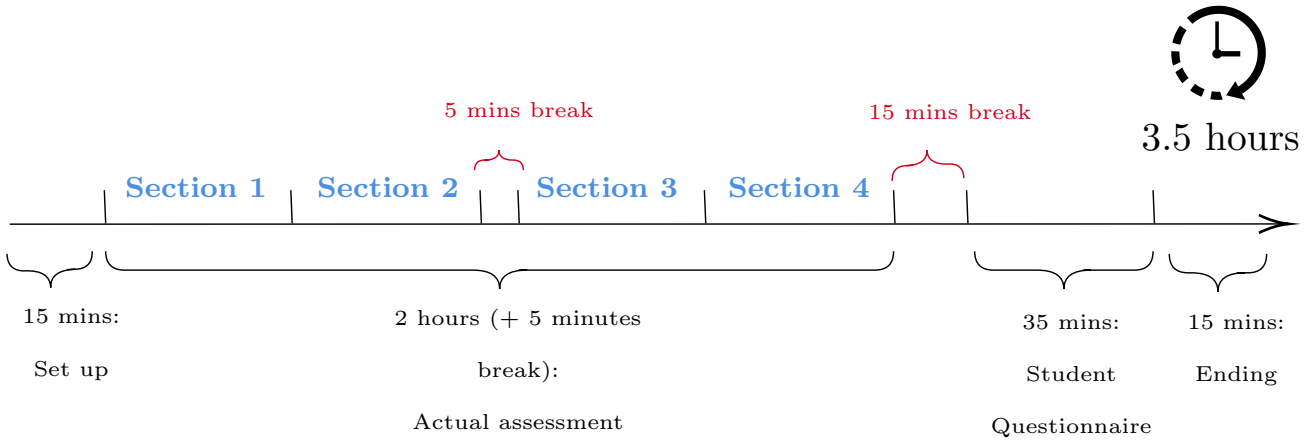
Table 8: Format of Test and Gender Gap in Reading and Science

	(1) Std. Scores in Reading b/se/se_sc	(2) Std. Scores in Science b/se/se_sc
Female	0.322	0.279
	(0.378)	(0.046)***
	[0.430]	[0.063]***
Female ×	0.345	
Prop. Reading MCQs	(0.149)**	
	[0.176]*	
Female ×		-0.219
Prop. Science MCQs		(0.052)***
		[0.065]***
Female ×	-0.604	
Prop. Reading ORQs	(0.768)	
	[0.873]	
Female ×		-0.871
Prop. Science ORQs		(0.045)***
		[0.084]***
Obs	249,327	381,826
Mean Y	-0.00	0.00
St Dev Y	1.00	1.00
Gender Gap	0.23	-0.07
School FE	Yes	Yes
Booklet FE	Yes	Yes
Students' Controls	Yes	Yes
R-sq	0.22	0.45

Notes: This table presents the estimations of model (1), for the reading and science domains. The outcome variables are the standardized score in reading sections, column 1, and the standardized score in science sections, column 2. The main explanatory variables are the proportion of multiple-choice questions in reading and science, which range from 0 to 1. The specifications include data from 2012 and 2015. Each specification controls for student's age in months, grade (compared to modal grade in the country), immigration status, parent highest education, and occupational levels, and home possession index. Standard errors are clustered at school level, in parenthesis, and at country level, in squared brackets. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Figure 1: Structure of PISA Test

(a) Timeline of the Test



(b) Example of Mathematics Sections

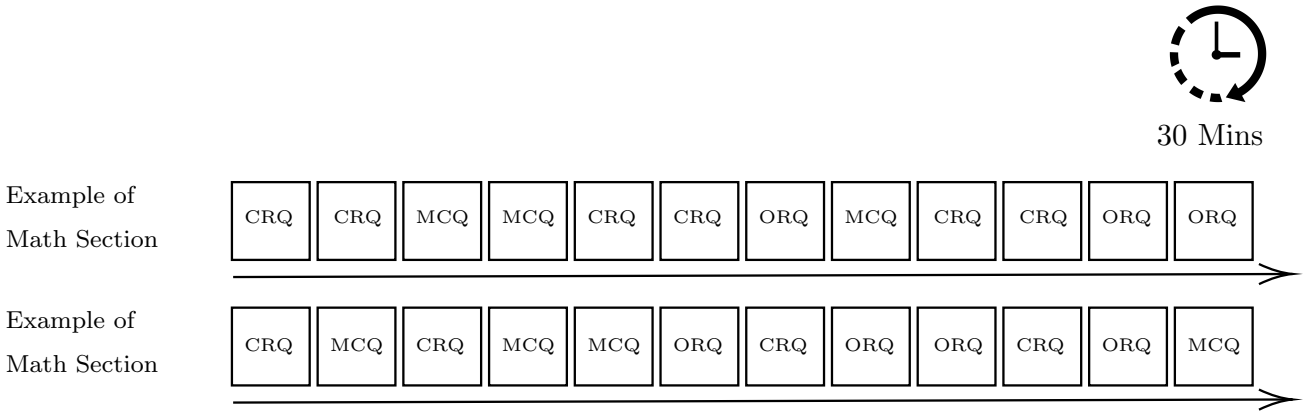
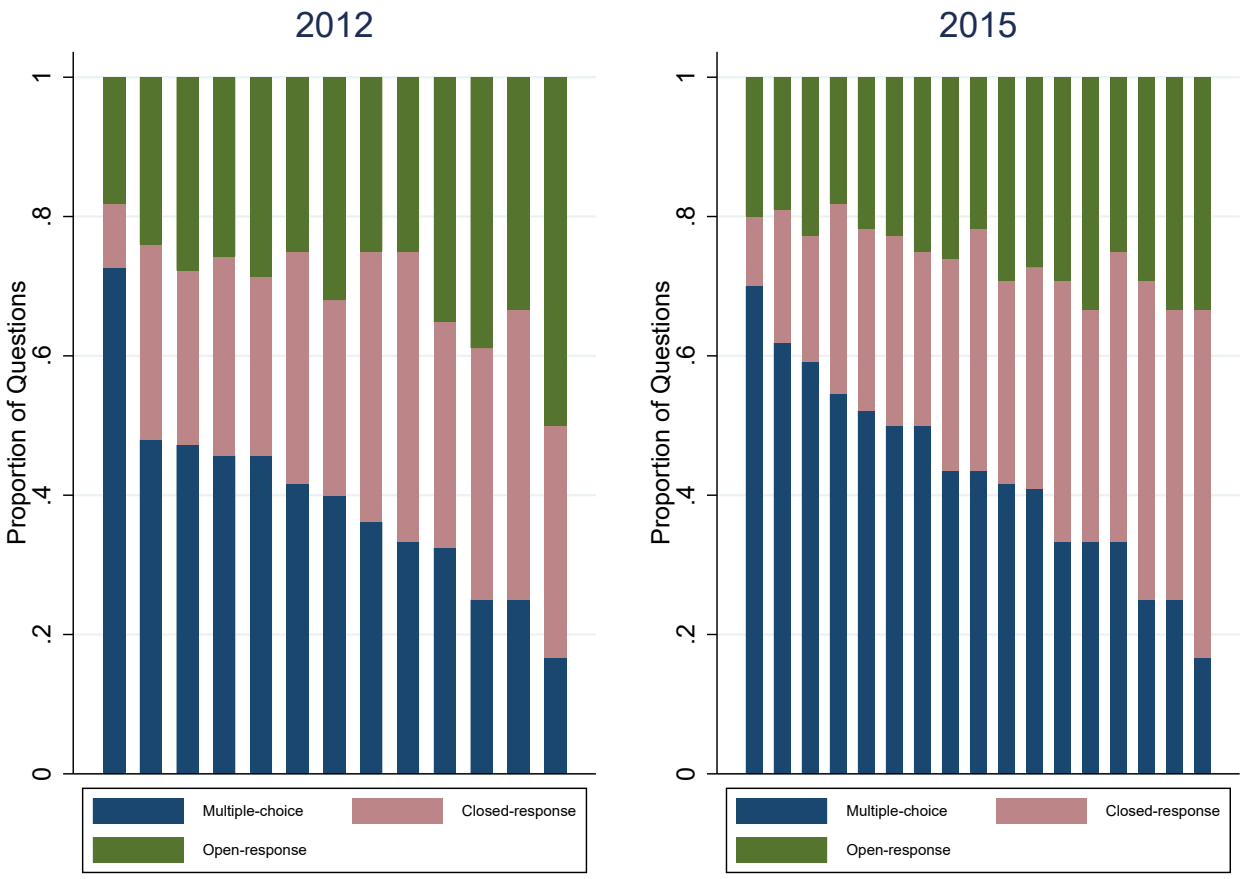
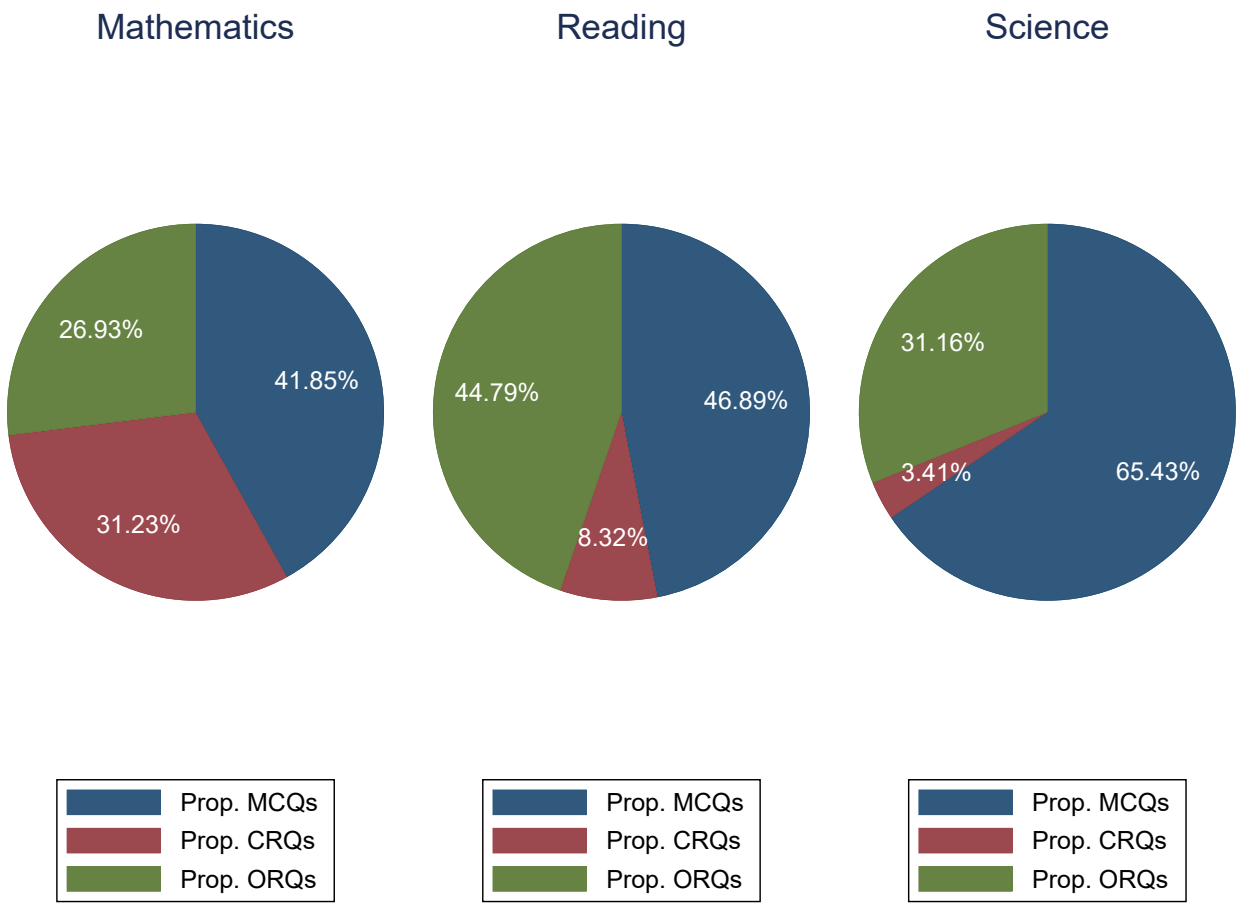


Figure 2: Variation in the Format of Mathematics Tests



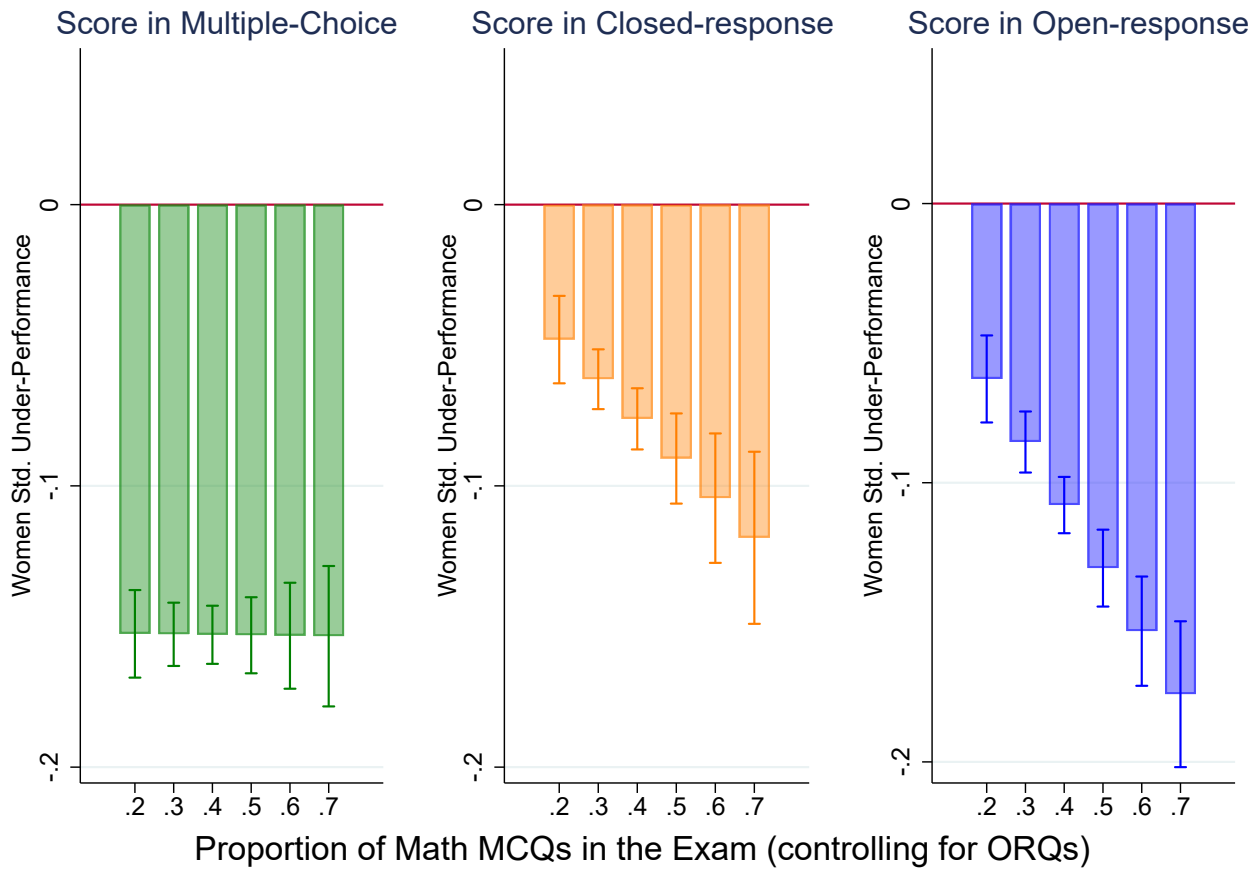
This figure shows the natural variation in the proportion of multiple-choice, closed-response, and open-response questions that arise from different exam booklets. The proportion of multiple-choice questions varies from 0.17 to 0.73 in 2012, and from 0.17 to 0.70 in 2015.

Figure 3: Proportions of Questions by Formats and Domains



Notes: This figure shows the variation in the proportion of multiple-choice, closed-response, and open-response questions in the different domains (using data from the 2012 and 2015 waves). In Mathematics there is a similar proportion of questions for all three formats. On the contrary, the proportion of close-response questions is below 10% (5%) in Reading (Science).

Figure 4: Format of Exam and Performance in Different Questions



Notes: This figure shows the estimates for females' under-performance in multiple-choice, closed- and open-response questions for exams with different proportions of multiple-choice questions, controlling for open-response questions. The results are obtained from the estimates of model 1.

Figure 5a: Gender Differences in Performance by Formats of Questions (2012)

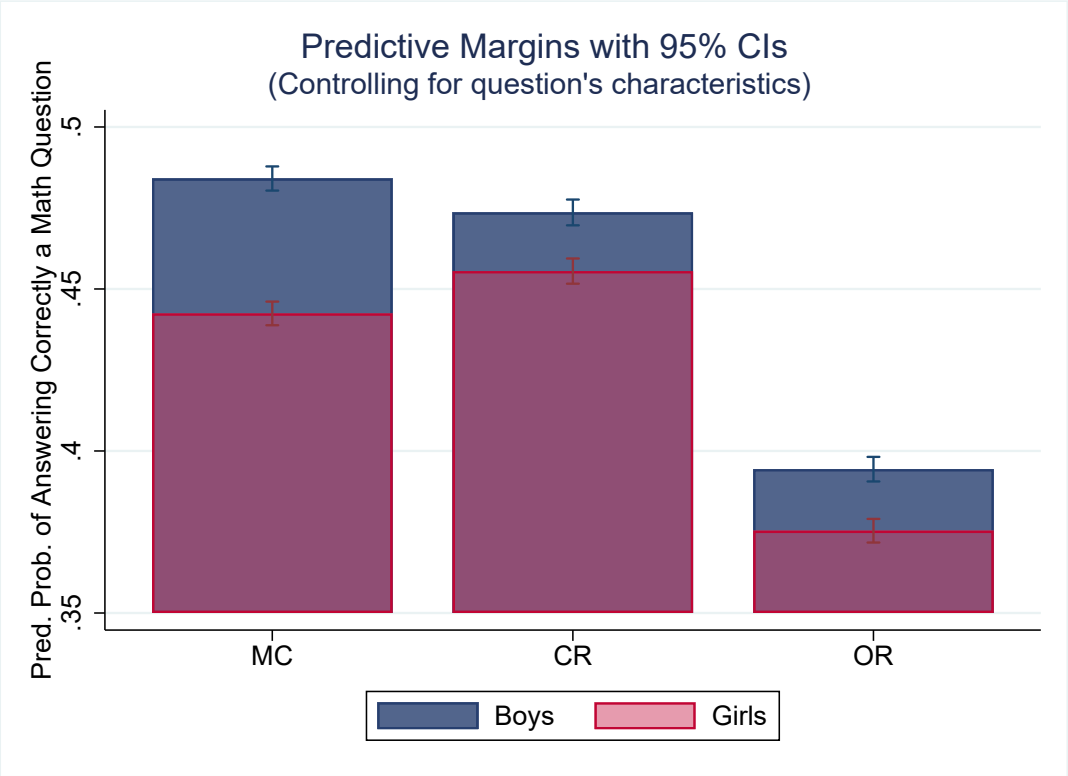
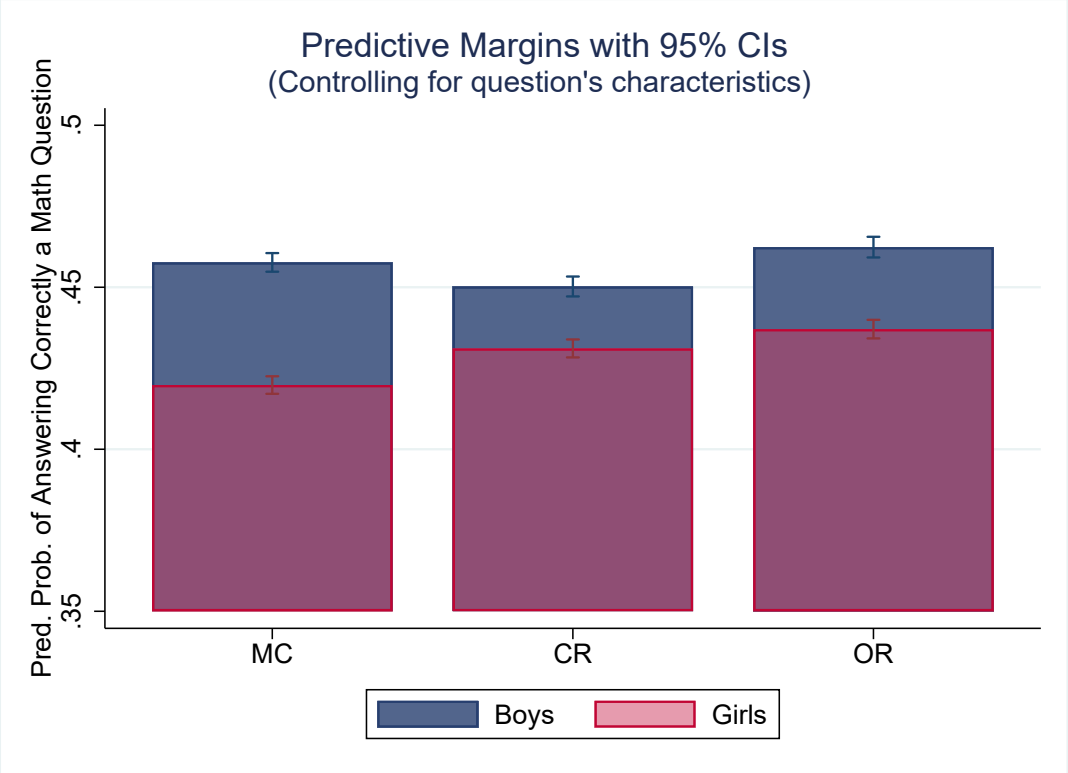
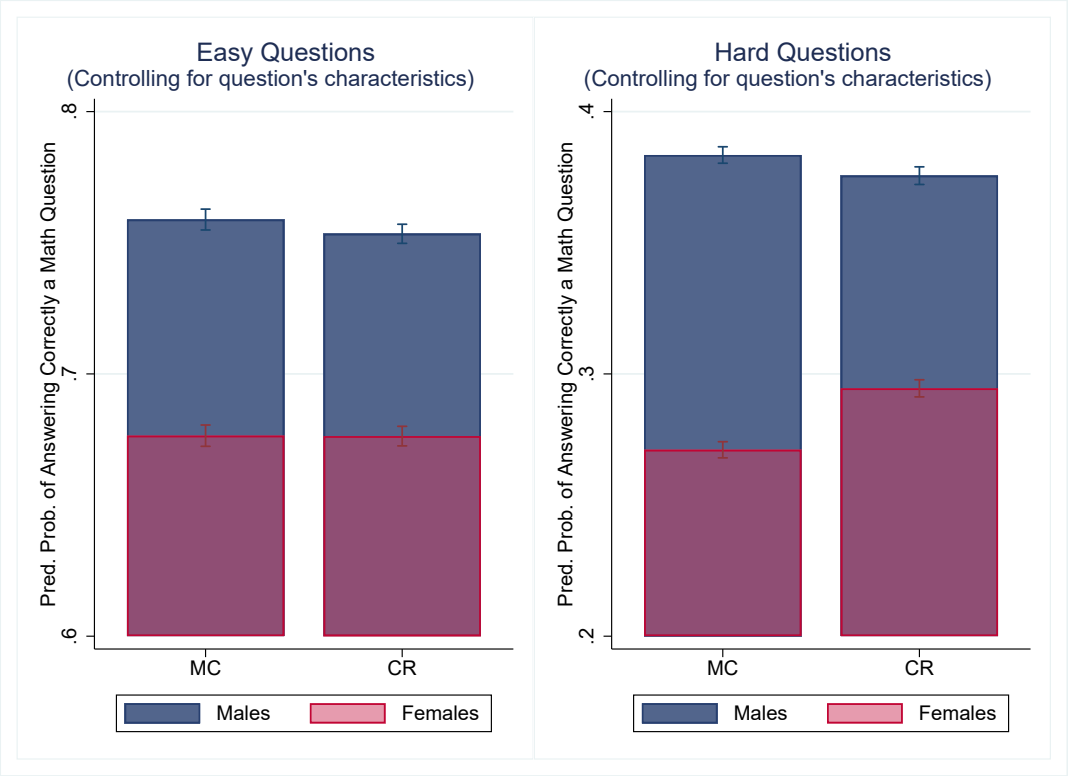


Figure 5b: Gender Differences in Performance by Formats of Questions (2015)



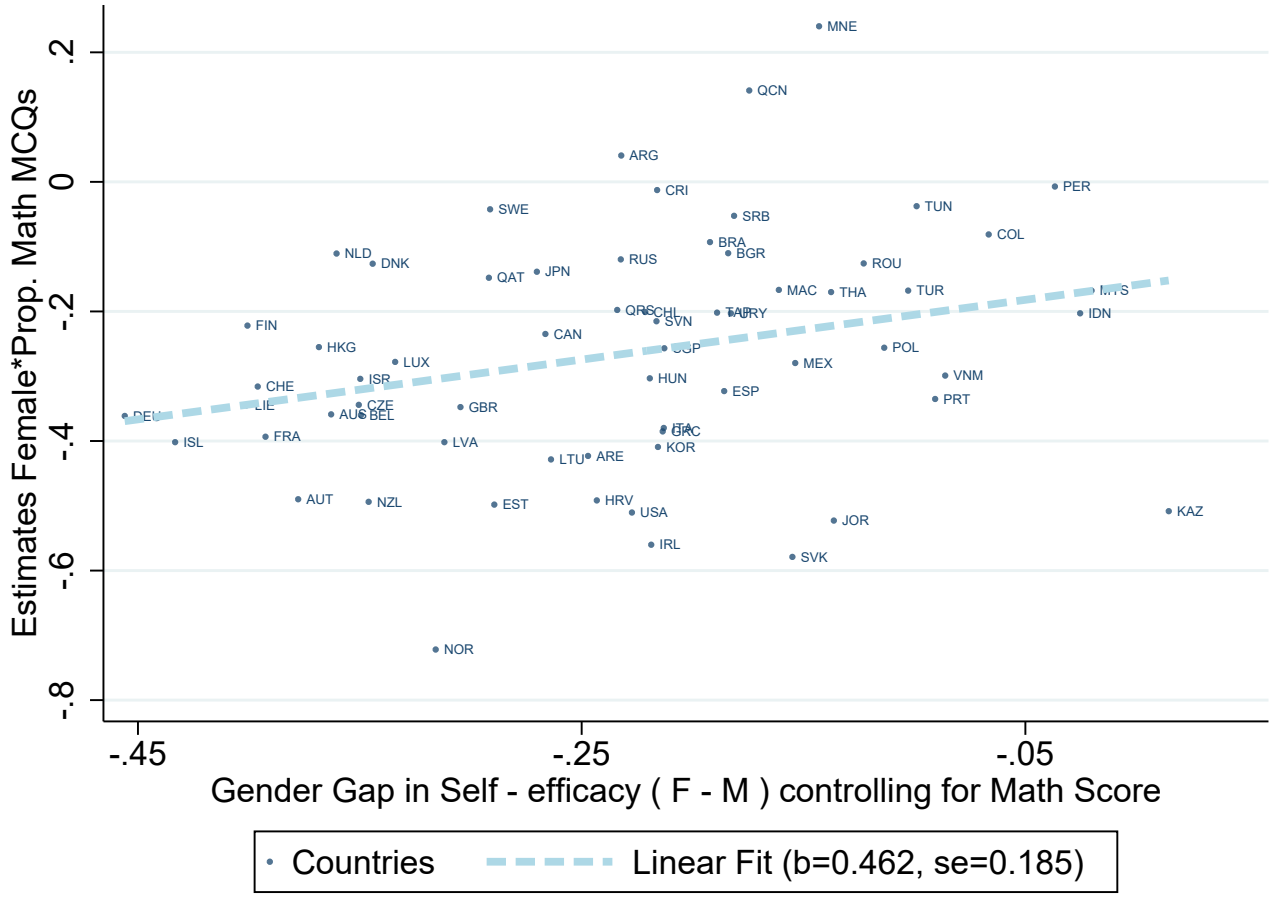
Notes: This figure shows the mathematics performance of male and female students in different formats of questions, as in model 2, using data from 2012 and 2015. Overall, males score better than females across all formats. The gender gap is, however, much greater for multiple-choice questions than for closed- and open-ended questions.

Figure 6: Gender Differences in Performance by Formats and Difficulty



Notes: Figure 6 shows the mathematics performance of male and female students in different formats of questions, as in model 2, by question difficulty. Questions are defined as easy if they required a level of proficiency not greater than 2, while questions are defined as hard if their required a level of proficiency of at least 5. See https://www.oecd.org/pisa/pisaproducts/PISA%202012%20Technical%20Report_Chapter%2015.pdf for a detailed description of the levels of proficiency. Data from the years 2012 and 2015 are included.

Figure 7: Impact of Exam Format and Gender Differences in Self-Efficacy



Notes: This graph plots for each country, the gender differences in students' self-efficacy in mathematics in the horizontal axis and the estimates for the interaction between the dummy female and the proportion of mathematics multiple-choice questions as in model 1, in the vertical axes. The gender differences in students' self-efficacy in mathematics represent the estimates β_1 for the dummy female on the following regression run separately for each country: Self Efficacy$_i = \beta_0 + \beta_1 Female_i + \gamma$Score in Math$_i + \varepsilon_i$. There is a positive and significant correlation between women's lower mathematics self-efficacy and the impact of the share of multiple-choice questions on female performance in mathematics (the fitted line has a slope coefficient equal to 0.462, with standard error equal to 0.185).

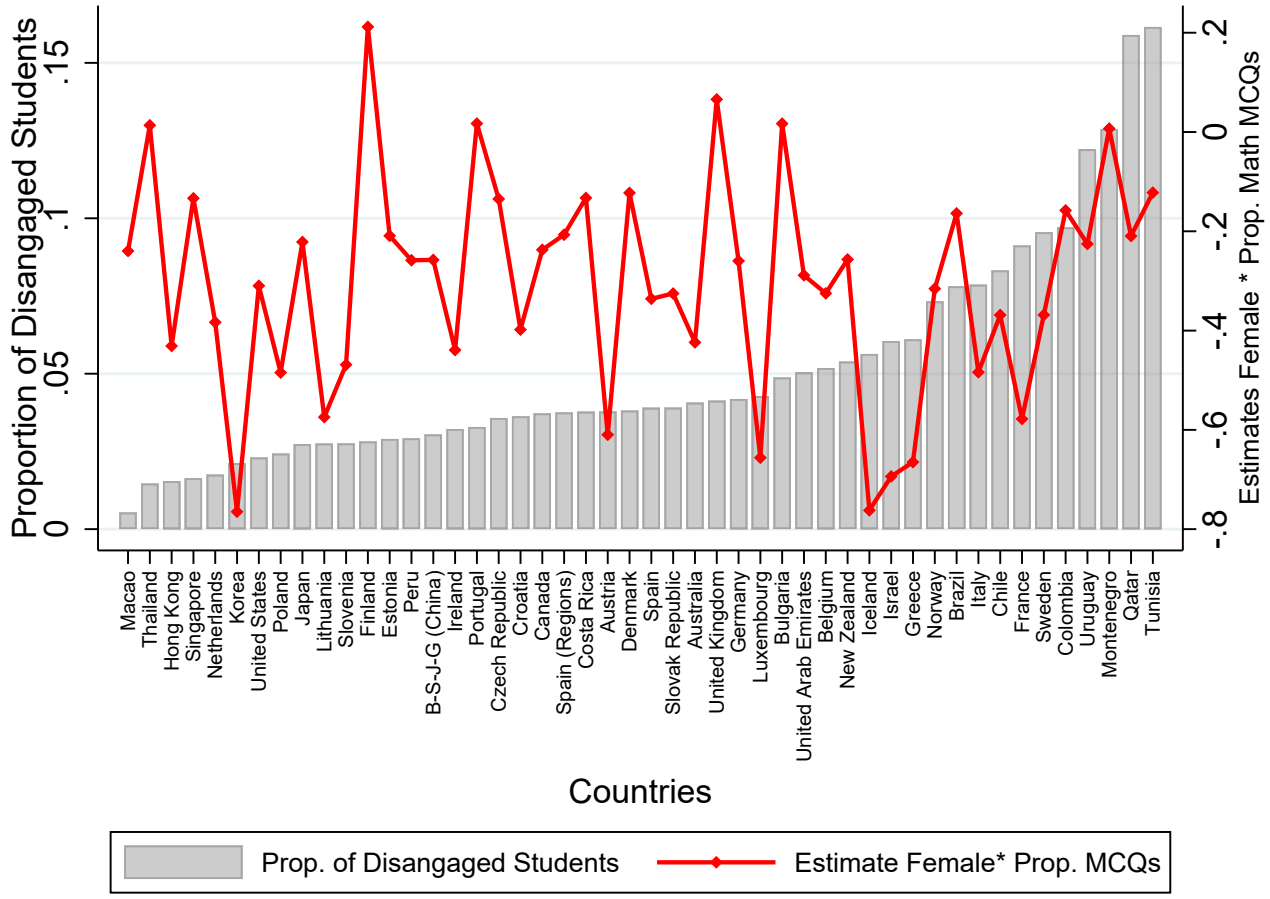
Figure 8: The Impact of Format on Different Sections

(a) The Impact of Format on Subsequent Sections

(b) The Impact of Format on Previous Sections: Placebo Investigation

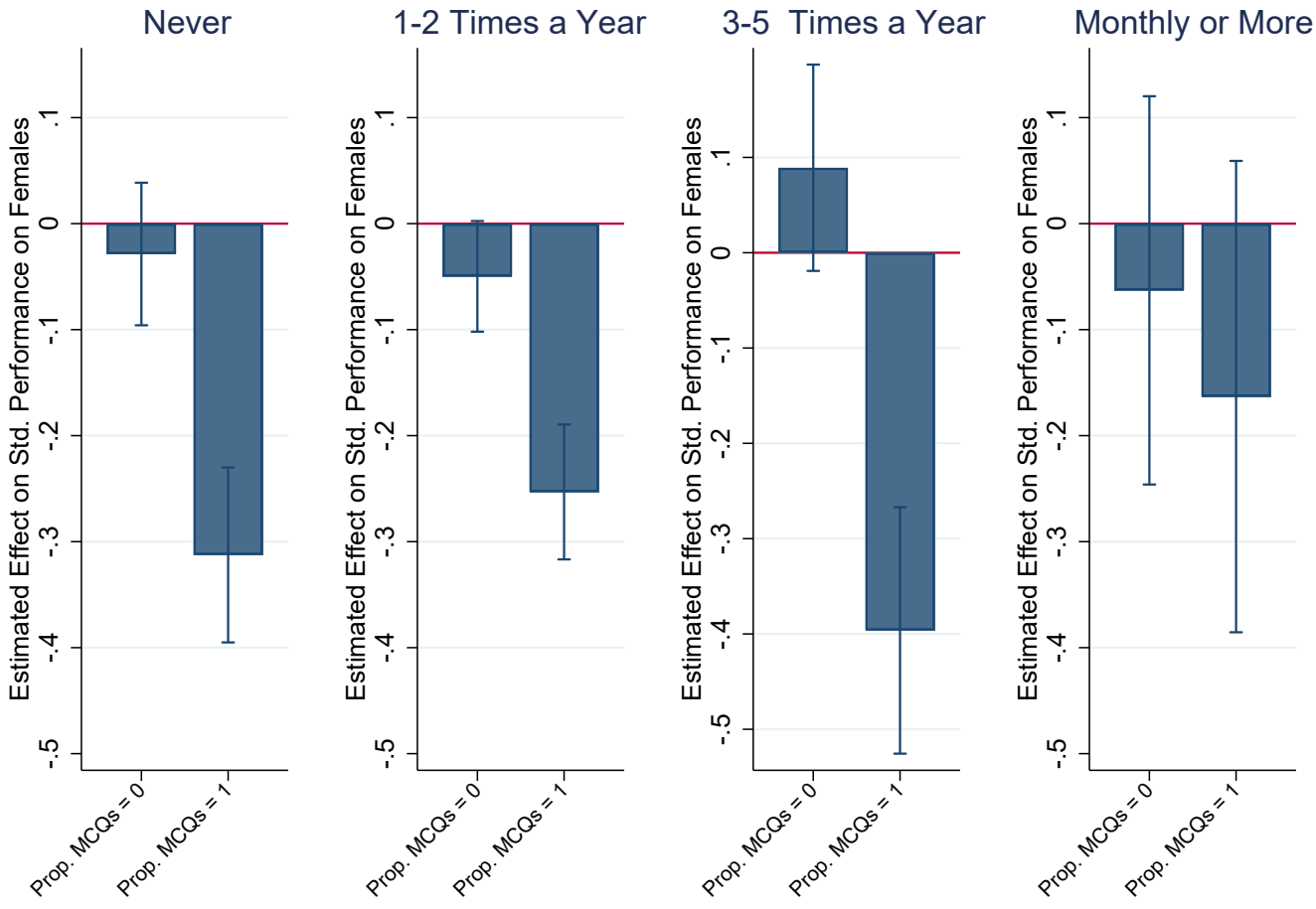
Notes: These figures provide an intuition of the identification used to estimate the spillover effect of formats on other exam sections. Figure (a) provides an example of the identification of the effect of the format of sections on subsequent ones. The analysis compares performance in the same mathematics section, section A, by students who previously faced a math section with more (in red) or less (in green) multiple-choice questions. Figure (b) provides an example of the identification of the effect of the format of sections on previous ones. This analysis compares performance in the same mathematics section, for example, section X, by students who face a subsequent section with more (in red) or less (in green) multiple-choice questions.

Figure 10: Format of Exam and Gender Differences in Performance by Stake of the Test



Notes: This graph plots 1] the proportion of disengaged students by countries (gray bars), 2] the estimates for the interaction between the dummy female and the proportion of mathematics multiple-choice questions as in model 1 (red line). The proportion of disengaged students by country represents a measure of the stake of the PISA test in the country (Akyol et al., 2021). A student is defined as disengaged if she omits 3 or more questions, even if there is enough time remaining in the cluster (i.e. at least 5 minutes), or if she answers too rapidly 3 or more questions, and the score in these questions is lower than the score of questions answered in normal time. The figure considers only countries where students complete the computer-based PISA assessment in 2015, where time data were available.

Figure 11: Format of Exam and Gender Differences in Performance by Frequency of Assessments



This figure displays females' under-performance in mathematics for exams with a higher share of multiple-choice questions, as in model 1 by frequency of the assessment. Information regarding the frequency of standardized assessment is obtained from the PISA school questionnaire. The analysis contains data from the 2015 assessment, where the information about frequency is available. Each specification controls for the student's age in months, grade (compared to modal grade in the country), immigration status, parents' highest education and occupational levels and home possession index, country, and booklet FE. Within countries, different schools assess students at different frequencies: this allows the estimation including country FE.

Appendix Tables and Figures

Table A1: Characteristics of Mathematics Questions

Questions' Characteristics:	**Format of the Question**			
Content	**Multiple-Choice**	**Closed-Response**	**Open-Response**	**Total**
Change and Relationship	14	13	22	49
Quantity	24	20	5	49
Space and Shape	17	15	14	46
Uncertainty and Data	24	11	11	46
Context				
Occupational	13	15	16	44
Personal	17	13	4	34
Scientific	15	11	22	48
Societal	34	20	10	64
Process				
Employ	32	33	20	85
Formulate	16	20	19	55
Interpret	31	6	13	50
Total	79	59	52	190
Question Difficulty (% of international incorrect)	44.67	52.20	74.54	55.44

Table A2: Performance and Time Response by Gender and Format of Questions

	Math Questions in 2012			Math Questions in 2015			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Correct Answer	Correct Answer Conditional on Answering	Skipped Question	Correct Answer	Correct Answer Conditional on Answering	Skipped Question	Time (Mins)
Female	-0.022***	-0.032***	-0.002	-0.026***	-0.029***	0.001	0.014
	(0.004)	(0.004)	(0.001)	(0.003)	(0.003)	(0.002)	(0.009)
Multiple-choice	0.011***	0.010***	-0.002***	0.011***	0.003**	-0.017***	-0.170***
	(0.001)	(0.001)	(0.000)	(0.001)	(0.001)	(0.001)	(0.003)
Female ×	-0.025***	-0.025***	0.000	-0.024***	-0.025***	-0.003***	-0.010**
Multiple-choice	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.001)	(0.004)
Open-response	-0.082***	0.017***	0.001*	0.017***	0.038***	0.092***	0.387***
	(0.001)	(0.001)	(0.000)	(0.001)	(0.001)	(0.001)	(0.005)
Female ×	0.003*	-0.007***	0.000	-0.012***	-0.014***	0.005***	0.065***
Open-response	(0.002)	(0.002)	(0.000)	(0.002)	(0.002)	(0.002)	(0.007)
Obs	1,756,093	1,728,324	1,756,093	1,579,733	1,471,571	1,579,733	1,579,733
Mean Y	0.43	0.46	0.05	0.43	0.46	0.05	1.59
St Dev Y	0.49	0.50	0.21	0.49	0.50	0.21	1.26
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Booklet FE	No	No	No	No	No	No	No
Students' Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Question's Difficulty	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Question's Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Observations are at the student-question level. Each specification is estimated using linear-model and controls for questions' characteristics (difficulty, content, context, and process), and their interaction with a female dummy, students characteristics (student's age in months, grade compared to modal grade in the country, immigration status, parent highest education, and occupational levels, and home possession index). Each specification includes school FE. The omitted category is the closed-response question. Columns 1-3 contain observations from wave 2012, while columns 4-7 contain data from wave 2015. The time information is available only in the 2015 wave. Standard errors, in parenthesis, are clustered at school level. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table A3: **Descriptive Statistics: Disengaged Students**

	Male	Female	Difference	*p-value*
	(1)	(2)	(3)	(4)
Inattentive Students	0.0939	0.0851	-0.0088	0.0000
Students who omit 3 or more questions	0.0544	0.0551	0.0007	0.3713
Students who answer too fast 3 or more questions and score lower in rapidly-answered questions	0.0654	0.0536	-0.0118	0.0000

Notes: This table reports the summary statistics for 1] disengaged students, namely students who omit 3 or more questions, even if there is enough time remaining in the cluster (i.e. at least 5 minutes), or who answer too rapidly 3 or more questions, and the proportion of correct questions answered too rapidly is lower than the proportion of correct questions answered in normal time; 2] students who omit 3 or more questions even if there enough time left to answer; and 3] students who answer 3 or more questions too rapidly, and the proportion of correct questions answered too rapidly is lower than the proportion of correct questions answered in normal time, by gender (columns 1 and 2, respectively). The table displays the gender difference between column (2) and (1) (column 3); and p-values for the t-test on the gender difference (column 4).

Table A4a: Proportion of Multiple-choice Questions and N. of Questions Assessing Different Context

	N. of Question By Context			
	(1) Occupational	(2) Personal	(3) Scientific	(4) Societal
Prop. of Multiple-choice Questions	-6.901*	-0.218	-2.985	8.114***
	(0.986)	(1.195)	(0.728)	(0.110)
Obs	16	16	16	16
Mean Y	2.75	2.13	3.00	4.00
St Dev Y	2.05	1.20	2.28	2.28

Notes: Observations are at the booklet level and include data from 2012 and 2015. The proportion of multiple-choice questions ranges from 0 to 1. Each specification includes survey year FE. Standard errors, in parenthesis, are clustered at the survey year level. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table A4b: Proportion of Multiple-choice Questions and N. of Questions Assessing Different Content

	N. of Question By Content			
	(1) Change and Relation	(2) Quantity	(3) Space and Shape	(4) Uncertanty
Prop. of Multiple-choice Questions	-3.182	1.366	2.533**	-2.706*
	(1.010)	(0.231)	(0.077)	(0.246)
Obs	16	16	16	16
Mean Y	3.06	3.06	2.88	2.88
St Dev Y	1.00	0.77	0.96	1.02

Notes: Observations are at the cluster level and include data from 2012 and 2015. The proportion of multiple-choice questions ranges from 0 to 1. Each specification includes survey year FE. Standard errors, in parenthesis, are clustered at the survey year level. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table A4c: Proportion of Multiple-choice Questions and N. of Questions Assessing Different Cognitive Process

	N. of Question By Cognitive Process Employed		
	(1) Employ Mathematical Concepts	(2) Formulating Situations	(3) Interpreting, Applying, Evaluating
Prop. of	0.608	-0.245	-2.353
Multiple-choice Questions	(1.158)	(1.856)	(1.450)
Obs	16	16	16
Mean Y	5.31	3.44	3.13
St Dev Y	1.08	1.36	1.41

Notes: Observations are at the cluster level and include data from 2012 and 2015. The proportion of multiple-choice questions ranges from 0 to 1. Each specification includes survey year FE. Standard errors, in parenthesis, are clustered at the survey year level. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table A5: Test Format and Gender Gap: controlling for Questions' Context and Content

	Std. Score in Mathematics in 2012				Std. Score in Mathematics in 2015			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female	0.015	-0.059***	-0.017	-0.176***	-0.090***	-0.025	-0.055***	-0.090
	(0.022)	(0.020)	(0.012)	(0.038)	(0.031)	(0.028)	(0.014)	(0.165)
Female ×	-0.313***	-0.248***	-0.174***	-0.050	-0.206***	-0.271***	-0.108***	-0.218
Prop. MCQs	(0.051)	(0.025)	(0.036)	(0.053)	(0.069)	(0.034)	(0.039)	(0.232)
Female ×		0.131***				-0.128*		
Prop. of Uncertainty Q.		(0.049)				(0.070)		
Female ×			-0.102***				-0.191***	
Prop. of Societal Q.			(0.032)				(0.033)	
Female ×				0.411***				0.048
Prop. of Occupational Q.				(0.094)				(0.361)
Obs	349,951	349,951	349,951	349,951	159,539	159,539	159,539	159,539
Mean Y	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
St Dev Y	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Booklet FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Students' Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R-sq	0.28	0.17	0.17	0.17	0.52	0.50	0.50	0.50

Notes: Observations are at the student level. Columns 1-4 include data on the year 2012 assessment; columns 5-8 include data on the year 2015 ones. Each specification controls for student's age in months, grade (compared to modal grade in the country), immigration status, parent highest education, and occupational levels, and home possession index. Column 1 and 5 report the specification in columns 3 and 6 of table 2 for reference. Columns 2 and 6 include the interaction terms with the dummy female and the proportion of questions assessing uncertainty, as opposed to quantity, change and relationship, and space and shape. Columns 3 and 7 include the interaction terms with the dummy female and the proportion of questions of societal rather than occupational, scientific, and personal context in the booklet. Columns 4 and 8 include the interaction terms with the dummy female and the proportion of questions that of occupational context, rather than societal, scientific and personal. Standard errors, in parenthesis, are clustered at school level. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table A6: Proportion of Multiple-choice Questions and Clusters Characteristics

	(1) N. of Question per Cluster	(2) Av. Quest. Difficulty	(3) N. of Easy Questions	(4) N. of Medium Questions	(5) N. of Hard Questions
Prop. of	0.811	3.390	-0.875	1.348	0.339
Multiple-choice Questions	(2.254)	(6.273)	(1.890)	(2.565)	(1.946)
Obs	41	41	41	41	41
Mean Y	14.00	51.83	3.51	6.98	3.51
St Dev Y	2.51	7.73	2.20	2.63	1.86
Year FE	Yes	Yes	Yes	Yes	Yes
Domain FE	Yes	Yes	Yes	Yes	Yes

Notes: Observations are at the cluster level and include data from 2012 and 2015. The proportion of multiple-choice questions ranges from 0 to 1. The outcome in column 2, the questions' difficulty, is defined as the percentage of questions answered incorrectly by all students in the participating countries. It ranges from 0 to 100%. The definition of easy, medium, and hard questions are computed from proficiency level provided by PISA (OECD, 2017b). In particular, a question is defined as of easy difficulty if means to assess students proficiency level 1, and 2. A question is defined as of medium difficulty if aims to assess students proficiency level 3, and 4, while as of hard difficulty if aims to assess students proficiency level 5 and 6. For a detailed analysis of the proficiency levels read OECD (2017b). Standard errors, are in parenthesis. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table A7: Questions' Format and Position

	Multiple-choice Question	
Sequence in cluster	-0.018	
	(0.016)	
Question Order= 1st to 3th		0.124
(Omitted: Question Order= 10th or above)		(0.157)
Question Order= 4th to 6th		-0.050
(Omitted: Question Order= 10th or above)		(0.159)
Question Order= 7th to 9th		0.000
(Omitted: Question Order= 10th or above)		(0.159)
Obs	81	81

Notes: Observations are at the question level and contain data only from the year 2015. In 2012, students completed a paper-based exam, so they were entitled to answer their questions in their preferred order. Only in 2015, when students completed a computer-based exam, the order of the questions would be relevant. The outcome variable is a dummy variable indicating whether the question is a multiple-choice question as opposed to a closed- or open-response one. The first specification shows the relationship between multiple-choices questions and the order of the questions within the cluster. The second specification includes three dummies variables indicating whether the question is included among the first 3 questions in the cluster, the second three questions, or the question's order range between the 7th and the 9th questions. The baseline category is a dummy indicating whether the question order is above 9th. Standard errors, are in parenthesis. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table A8: Test Format and Performance by Migration and Socioeconomic Status

	Std. Scores in Mathematics in 2012		Std. Scores in Mathematics in 2015	
	(1)	(2)	(3)	(4)
	b/se/se_sc	b/se/se_sc	b/se/se_sc	b/se/se_sc
1st or 2nd Gen. Immigrant × Prop.	-0.194		-0.071	
Math MCQs	(0.125)		(0.162)	
	[0.055]		[0.119]	
Socio-economic Status × Prop.		-0.009		0.009
Math MCQs		(0.011)		(0.014)
		[0.026]		[0.027]
1st or 2nd Gen. Immigrant	-0.002	-0.062	-0.023	-0.034
	(0.050)	(0.016)***	(0.078)	(0.009)***
	[0.106]	[0.101]	[0.078]	[0.037]
Socio-economic Status	0.177	0.068	0.060	0.043
	(0.006)***	(0.008)***	(0.006)***	(0.007)***
	[0.024]***	[0.044]	[0.011]***	[0.014]***
Obs	340,750	340,750	156,516	156,516
Mean Y	-0.00	-0.00	-0.00	-0.00
St Dev Y	1.00	1.00	1.00	1.00
School FE	Yes	Yes	Yes	Yes
Booklet FE	Yes	Yes	Yes	Yes
Students' Controls	Yes	Yes	Yes	Yes
R-sq	0.27	0.17	0.52	0.49
Raw Gender Gap	-0.10	-0.10	-0.14	-0.14

Notes: This table presents the estimations of model (1) where the main explanatory term become the interaction between the proportion of mathematics multiple-choice questions and migration status (columns 1 and 3), or socio-economics status (columns 2 and 4). Migration status is measured as dummy indicating whether a student is a first or second-generation immigrant, as opposed to a native student. The socio-economic status is a continues standardized measure of economic and cultural possession in the household. The proportion of mathematics multiple-choice questions ranges from 0 to 1. Standard errors, are clustered at school level, in parenthesis, and at country level, in square brackets. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table A9: Heterogenous Effect by Frequency of Standardized Assessments

	Std. Scores in Mathematics in 2015			
	(1)	(2)	(3)	(4)
	Never	1-2 per Year	3-5 per Year	Montly or More
	b/se/se_sc	b/se/se_sc	b/se/se_sc	b/se/se_sc
Female	-0.029	-0.050	0.089	-0.063
	(0.034)	(0.027)*	(0.055)	(0.093)
	[0.038]	[0.030]*	[0.053]	[0.087]
Female × Prop. Math MCQs	-0.284	-0.203	-0.485	-0.100
	(0.074)***	(0.057)***	(0.117)***	(0.199)
	[0.070]***	[0.057]***	[0.102]***	[0.180]
Obs	39,511	72,061	17,016	5,614
Mean Y	-0.00	-0.00	-0.00	-0.00
St Dev Y	1.00	1.00	1.00	1.00
School FE	No	No	No	No
Country FE	Yes	Yes	Yes	Yes
Booklet FE	Yes	Yes	Yes	Yes
Students' Controls	Yes	Yes	Yes	Yes
R-sq	0.34	0.30	0.36	0.33

Notes: This table shows the heterogeneous effect of the proportion of multiple-choice questions on the gender gap in mathematics performance as in equation (1), by the frequency in which schools assess students using standardized assessments. Information regarding the frequency in which the school assesses students using standardized assessment is obtained from the PISA school questionnaire. The analysis contains data from the 2015 assessment, where the information about frequency is available. Each specification controls for student's age in months, grade (compared to modal grade in the country), immigration status, parent highest education, and occupational levels and home possession index, year, and booklet FE. Within countries, different schools assess students at different frequencies: this allows the estimation including country FE. Standard errors, are in parenthesis. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

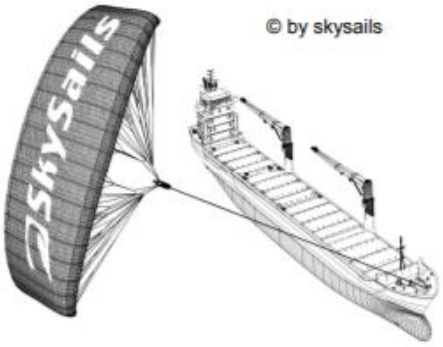
Figure A1: Example of a Multiple-Choice Question

SAILING SHIPS

Ninety-five percent of world trade is moved by sea, by roughly 50 000 tankers, bulk carriers and container ships. Most of these ships use diesel fuel.

Engineers are planning to develop wind power support for ships. Their proposal is to attach kite sails to ships and use the wind's power to help reduce diesel consumption and the fuel's impact on the environment.

Translation Note: "© by skysails": Do not adapt skysails as this is a registered label.



Question 1: SAILING SHIPS

PM923Q01

One advantage of using a kite sail is that it flies at a height of 150 m. There, the wind speed is approximately 25% higher than down on the deck of the ship.

At what approximate speed does the wind blow into a kite sail when a wind speed of 24 km/h is measured on the deck of the ship?

A 6 km/h
B 18 km/h
C 25 km/h
D 30 km/h
E 49 km/h

This figure shows an example of a multiple-choice question in the PISA 2012 mathematics exam. Students are required to pick the correct answer among a set of possible 5 answers. Students do not receive any penalty for choosing the wrong response. *Source:* https://www.oecd.org/pisa/pisaproducts/pisa2012-2006-rel-items-maths-ENG.pdf

Figure A2: Example of a Closed-Response Question

Question 2: SAUCE PM924Q02 – 0 1 9

You are making your own dressing for a salad.

Here is a recipe for 100 millilitres (mL) of dressing.

Salad oil:	60 mL
Vinegar:	30 mL
Soy sauce:	10 mL

How many millilitres (mL) of salad oil do you need to make 150 mL of this dressing?

Answer: …….. mL

This figure shows an example of a closed-response question in the PISA 2012 mathematics exam. Students are required to provide a short and concise answer. *Source:* https://www.oecd.org/pisa/pisaproducts/pisa2012-2006-rel-items-maths-ENG.pdf

Figure A3: Example of a Open-Response Question

Question 4: SAILING SHIPS PM923Q04 – 0 1 9

Due to high diesel fuel costs of 0.42 zeds per litre, the owners of the ship *NewWave* are thinking about equipping their ship with a kite sail.

It is estimated that a kite sail like this has the potential to reduce the diesel consumption by about 20% overall.

Name: *NewWave*

Type: freighter

Length: 117 metres

Breadth: 18 metres

Load capacity: 12 000 tons

Maximum speed: 19 knots

Diesel consumption per year without a kite sail: approximately 3 500 000 litres

The cost of equipping the *NewWave* with a kite sail is 2 500 000 zeds.

After about how many years would the diesel fuel savings cover the cost of the kite sail? Give calculations to support your answer.

.....

.....

.....

.....

.....

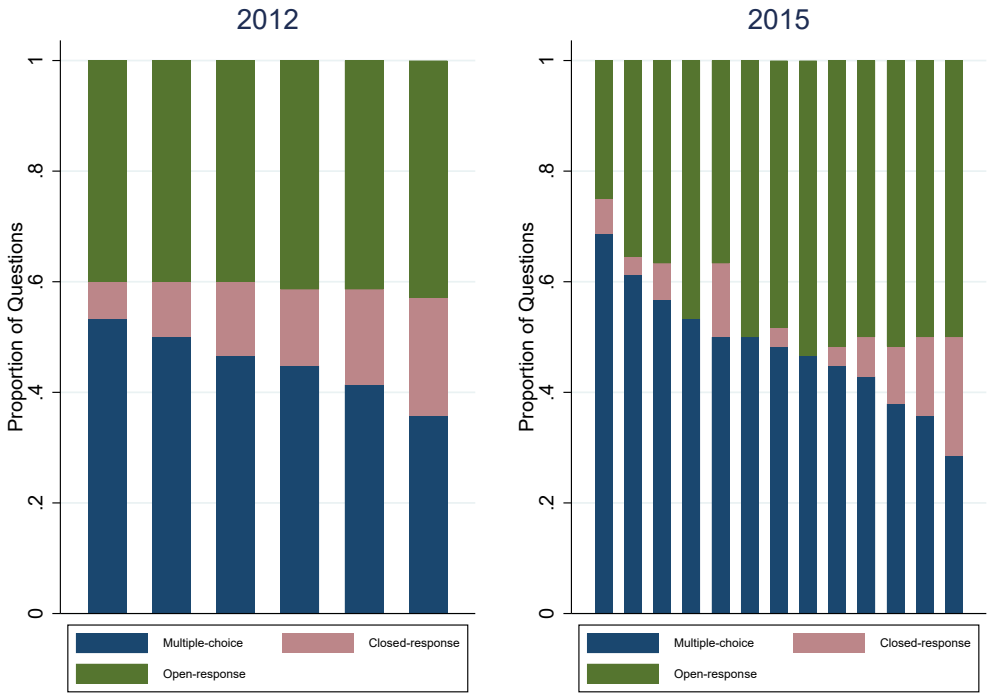
.....

.....

Number of years:.....

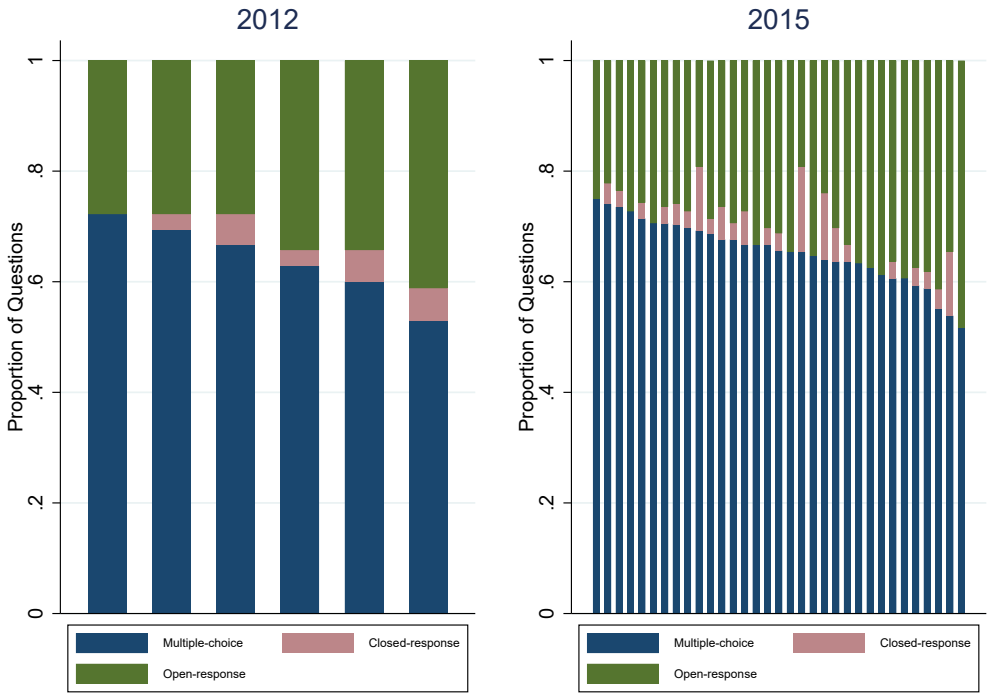
This figure shows an example of an open-response question in the PISA 2012 mathematics exam. Students are required to provide a short and concise answer alongside a detailed explanation to support the answer. *Source:* https://www.oecd.org/pisa/pisaproducts/pisa2012-2006-rel-items-maths-ENG.pdf

Figure A4.a: Variation in the Proportion of Questions by Formats in Reading



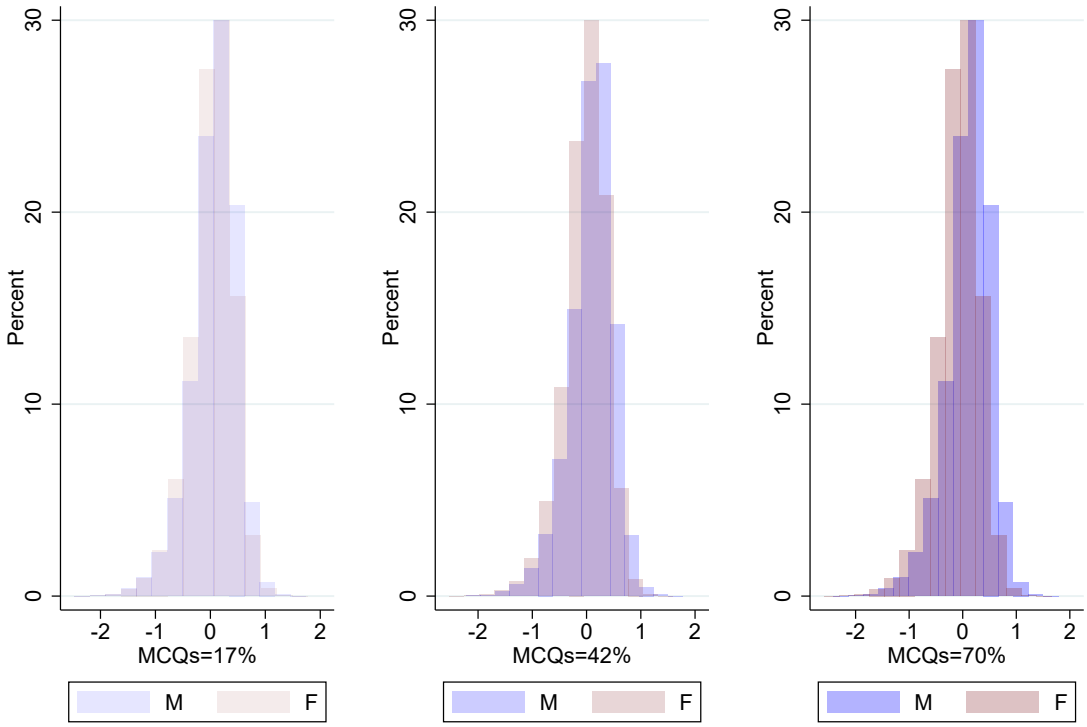
Notes: This figure shows the variation in the proportion of multiple-choice, closed-response, and open-response questions in the different combinations of reading clusters in 2012 and 2015. The proportion of multiple-choice questions varies from 0.36 to 0.53 in 2012, and from 0.29 to 0.68 in 2015.

Figure A4.b: Variation in the Proportion of Questions by Formats in Science



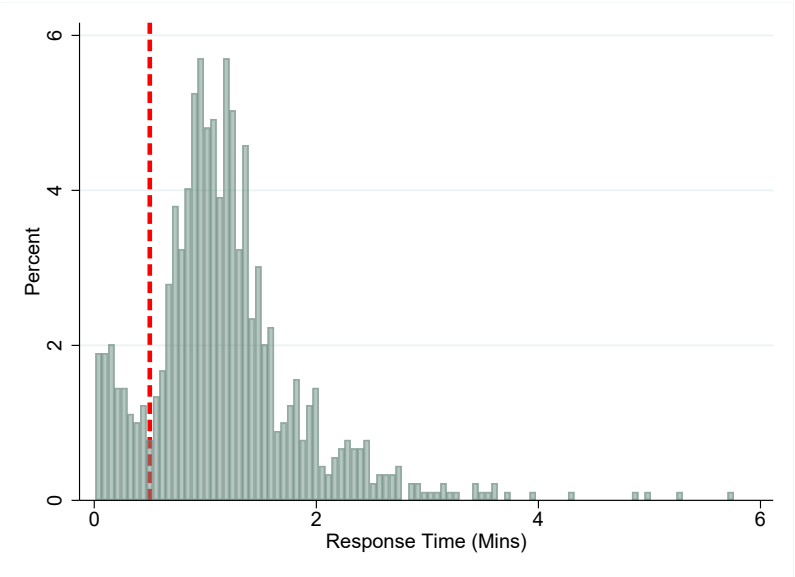
Notes: This figure shows the variation in the proportion of multiple-choice, closed-response, and open-response questions in the different combinations of science clusters in 2012 and 2015. The proportion of multiple-choice questions varies from 0.53 to 0.73 in 2012, and from 0.52 to 0.75 in 2015.

Figure A5: Predicted Distributions of Mathematics Scores and Format of Test



Notes: These figures show the predicted distributions of mathematics scores for male and female students, as the share of mathematics multiple-choice questions increases from 17% (the lower proportion of multiple-choice questions that students received in the sample) to 42% (the average proportion of multiple-choice questions in the sample) to 70% (the highest proportion of multiple-choice questions that students received in the sample). The distributions are obtained using linear prediction from model 1, when the proportion of multiple-choice questions increases from 17 to 42 to 70%.

Figure A6: Example of Response Time Distribution



Notes: This Figure plots the time spent in answering a specific question, in a given country. The red-dashed line represents the Gaussian mixture model threshold.

Figure A7a: Heterogeneous Effects by Mathematics Performance (2012)

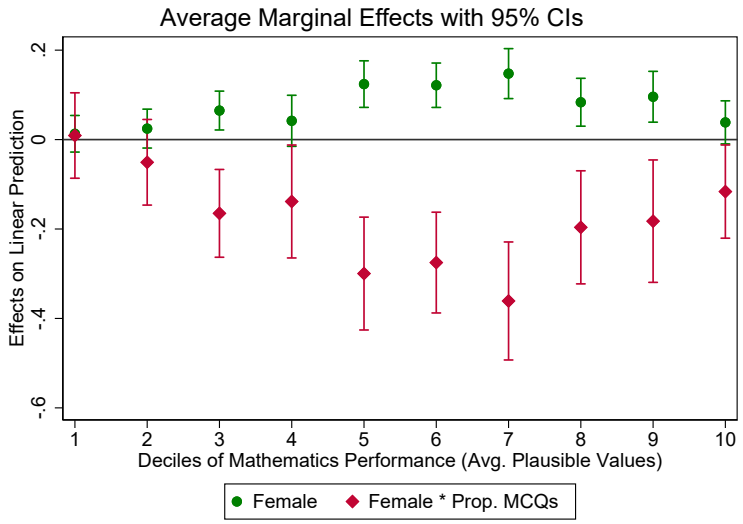
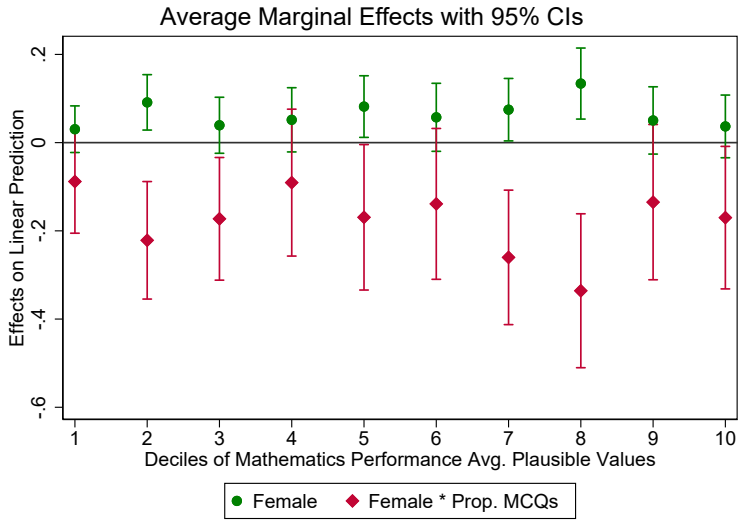
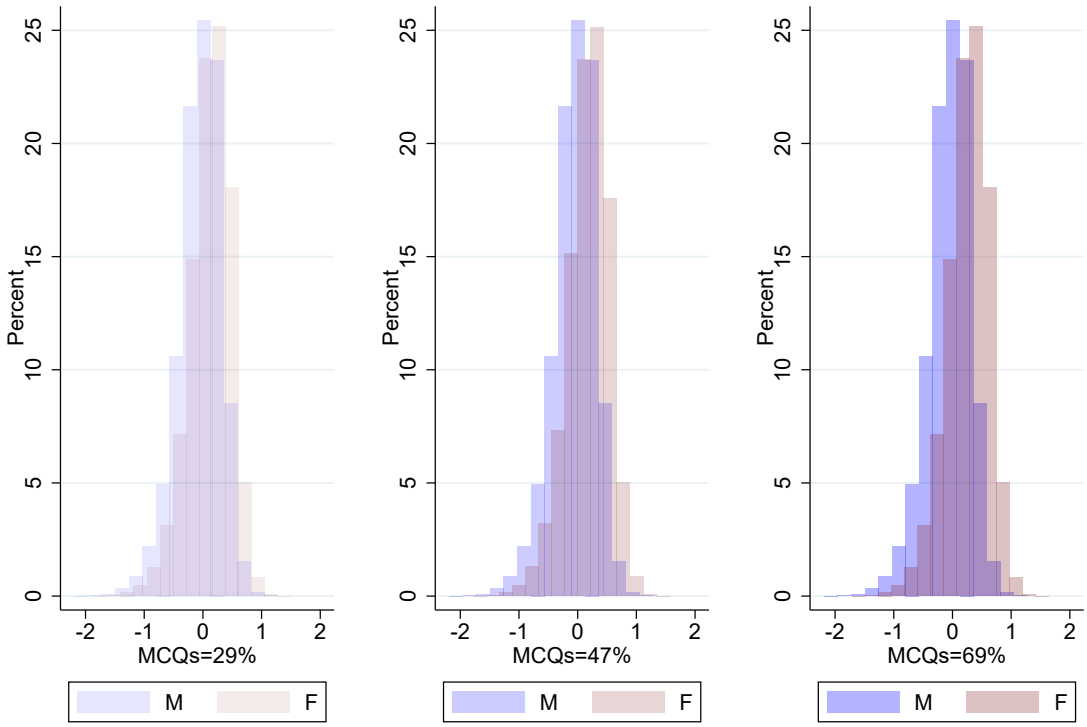


Figure A7b: Heterogeneous Effects by Mathematics Performance (2015)



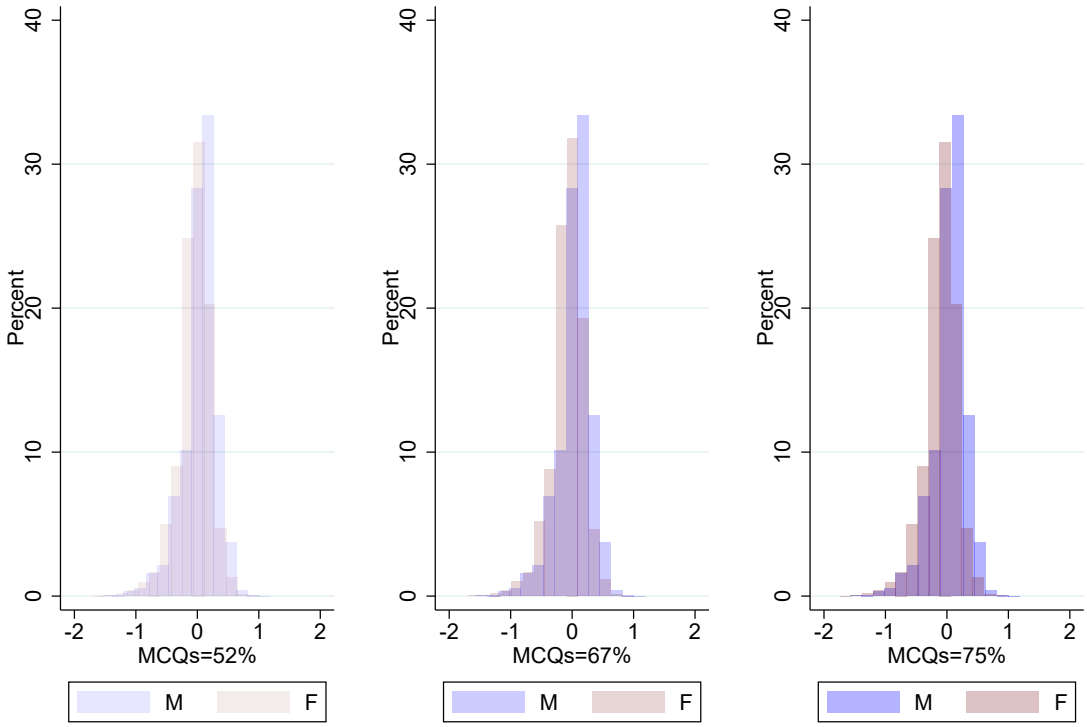
Notes: These figures show the estimate and the 95% confidence intervals for female and its interaction with the proportion of mathematics' question in models 1 by decile of Mathematics Performance in 2012 (Figure A7a) and 2015 (Figure A7b). Each specification controls for student's age in months, grade (compared to modal grade in the country), immigration status, parental highest level of education, and occupational levels and home possession index, schools, and booklet FE. The horizontal axes represent each decile of mathematics performance computed as the average of the Plausible Values. The lowest decile, 1, indicates students among the bottom 10% of performance, while the highest decile, 10, represents students among the top 10% of performance.

Figure A8: Predicted Distributions of Reading Scores and Format of Test



Notes: These figures show the predicted distributions of reading scores for male and female students, as the share of reading multiple-choice questions increases from 29% (the lower proportion of multiple-choice questions that students received in the sample) to 47% (the average proportion of multiple-choice questions in the sample) to 69% (the highest proportion of multiple-choice questions in reading that students received in the sample). The distributions are obtained using linear prediction from model 1, when the proportion of multiple-choice questions increases from 29 to 47 to 69%.

Figure A9: Predicted Distributions of Science Scores and Format of Test



Notes: These figures show the predicted distributions of science scores for male and female students, as the share of science multiple-choice questions increases from 52% (the lower proportion of multiple-choice questions that students received in science) to 67% (the average proportion of multiple-choice questions in the sample) to 75% (the highest proportion of multiple-choice questions in reading that students received in the sample). The distributions are obtained using linear prediction from model 1, when the proportion of multiple-choice questions increases from 52 to 67 to 75%.