Different Questions, Different Gender Gap:
Can the Format of Questions Explain the Gender Gap in Mathematics?

Silvia Griselda*

November 13, 2020

Job Market Paper
Latest version available here

Abstract

Standardized assessments are widely used to determine educational and economic opportunities. These standardized assessments exclusively, or in large part, use multiple-choice questions. But multiple-choice exams may not be adequate for comparing students' competencies across genders. In this paper, I show that female students receive lower marks when randomly assigned to exams with a larger proportion of multiple-choice questions. Specifically, a 10 percentage point increase in the proportion of multiple-choice questions widens the gender difference in mathematics performance by 0.026 standard deviations in favor of men, an effect that represents about 50% of the overall gender gap. Moreover, a higher proportion of multiple-choice questions has negative spillovers to other open-ended questions on the same exam. Female students exert less effort than males on tests that contain a larger proportion of multiple-choice questions. I provide suggestive evidence that these results are driven by women's lower confidence and by the stereotypes that women face in traditionally male domains.

Keywords: achievement gap, gender gap, mathematics, multiple-choice
JEL Classification: I21, I24, J24

*PhD Candidate at Department of Economics, University of Melbourne. Email: sgriselda@student.unimelb.edu.au. I am very grateful for the great guidance and support from Victoria Baranov, Lisa Cameron, and Michael Bernard Coelli. I am also thankful for the invaluable comments from David Byrne, Edwin Chan, Sarah Dahmann, Shashi Karunanethy, Leslie Martin, Cain Polidano, Maria Recalde, Nicolas Salamanca, Anna Sanz-de-Galdeano, Sofia Karina Trommlerová, Haikun Zhan, and the seminar participants at the Australian Applied Young Economist Webinar, the 35th Annual Conference of the Italian Association of Labour Economists, the 5th IZA Workshop: The Economics of Education, the Graduate Students in Economics of Education Zoom (GEEZ) seminar, the Prometeo Workshop on Applied Microeconomics and Gender, and the economics seminar at Cardiff University. I acknowledge financial support from the University of Melbourne's FBE Doctoral Program Scholarship for this research. All errors are my own.

1 Introduction

Despite catching up in most educational outcomes, females continue to under-perform and be under-represented in mathematics-intensive fields. This has important implications for both women and society overall, as mathematics skills play a crucial role in determining future earnings (Joensen and Nielsen, 2009) and entrance into highly paid STEM occupations (Beede et al., 2011; Card and Payne, 2017). Previous research has extensively analyzed how environmental factors such as stereotypes, culture, and role models can affect gender differences in mathematics performance, choices and preferences of individuals (Guiso et al., 2008; Kahn and Ginther, 2017; Carlana, 2019).

This paper builds new knowledge by showing how screening technologies used in education and job applications — standardized tests — can themselves perpetuate and reinforce systematic gender differences in academic and labor market outcomes. Standardized tests are widely used around the world to determine university admissions, provision of licenses and certifications, as well as to determine the effectiveness of different educational inputs (Coffman and Klinowski, 2020). Many of these standardized tests employ, totally or in large part, multiple-choice questions to assess students' competences.[1] These questions are considered objective, low cost and easy to implement on a large scale (Frederiksen, 1984; Duquennois, 2019).[2] Yet, there is limited understanding of whether these tests capture individuals' underlying knowledge, or whether their results are also affected by other factors (Freedle, 2003; Riener and Wagner, 2017).

In this paper, I explore whether the wide use of multiple-choice questions with no negative marking on mathematics tests contributes to widen the gender difference in mathematics performance. I uncover three key findings: first, I document that boys have an advantage in multiple-choice compared to constructed-response questions. Second, I show that girls obtain lower marks when they randomly received an exam with a larger proportion of multiple-choice questions. Interestingly, the differing proportion of multiple-choice questions has spillover effects on females' performance on constructed-response questions. Third, I establish that a larger proportion of multiple-choice questions induces girls to exert less effort during the

[1] US and other universities 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.

[2] Multiple-choice questions require students to choose among a set of possible alternatives. This format is different from closed-response items, where students need to come up with a short answer, and open-response items, where students need to provide an extensive explanation of the answer, or provide an analytical and complete solution.

test: they omit more questions, even when they have enough time left, and answer carelessly.

The educational literature widely documents that boys perform better in multiple-choice tests and girls perform better in constructed-response ones (Bolger and Kellaghan, 1990; DeMars, 2000; Willingham and Cole, 2013). Yet, it is not clear whether these gender differences by format are simply driven by gender gaps in different areas of knowledge (i.e. *construct-relevant skills*) or by other factors that tests do not intend to assess, such as willingness to guess and omission strategies (i.e. *construct-irrelevant skills*).[3]

In this paper, I am able to isolate the role of question's characteristics from gender differences in answering strategies in explaining boys' advantage in multiple-choice questions. I employ data from the Programme for International Student Assessment (PISA). PISA is the largest international assessment, involving more than 500,000 15-year-old students from more than 60 countries. PISA specifies the exact construct-relevant skills that each question aims to assess. I document that girls' under-performance is greater in multiple-choice than in other formats of questions even after controlling for construct-relevant skills and question difficulty. Further, I set aside the role of writing skills by documenting boys' advantage in multiple-choice compared to closed-response questions, a format that requires students to provide only a short and concise answer.[4]

After documenting boys' advantage in answering multiple-choice questions, I analyze the effect of differing proportions of multiple-choice questions on performance and its spillover effect on other formats. In particular, I exploit the random assignment of exam booklets with different proportions of multiple-choice questions to students. My results show that a 10 percentage point increase in the proportion of mathematics multiple-choice questions differentially decreases girls' performance in mathematics by 0.026 of a standard deviation

[3]Construct-relevant skills are abilities that are meant to be assessed in the tests, while construct-irrelevant skills are nuisance factors that are not meant to be assessed in the tests, but could affect test performance. Liu and Wilson (2009) and Taylor and Lee (2012) reveal that multiple-choice questions require students to identify a reasonable response, a competence in which boys outperform girls, while constructed-response questions require students to provide their interpretation and analysis, a skill where in general girls outperform boys. In addition, Reardon et al. (2018) highlight boys' better performance in geometry, probability, and algebra (contents usually more likely assessed using multiple-choice questions), and girls' advantage in problem-solving and statistical interpretation (contents that are often assessed in constructed-response questions). Interpreting and identifying reasonable responses are both considered construct-relevant skills. (Bridgeman, 1992) documents that different genders have different guessing and omission strategies, and they react differently to multiple answers, therefore performance on multiple-choice assessments may be impaired by these gender differences in construct-irrelevant skills. Ben-Shakhar and Sinai (1991) document boys' greater tendency to guess and girls' higher omission rate at any level of knowledge. Yet von Schrader and Ansley (2006) find that boys' greater tendency to guess did not explain their higher performance in multiple-choice tests with no penalty for wrong responses.

[4]PISA assesses students using another construct-response format of question: open-response. Open-response questions ask students to provide their answers alongside a detailed explanation of their reasoning.

compared to that of boys, an effect that represents about 50% of the baseline gender gap in mathematics. A decrease in girls' performance compared to men by 0.026 of a standard deviation is comparable to a decrease in teacher quality of one-quarter of a standard deviation (Rivkin et al., 2005), or an increase in class size of one student (Angrist and Lavy, 1999). Furthermore, the higher proportion of multiple-choice questions has negative spillover effects on females' likelihood of correctly answering closed- and open-response questions. In particular, a 10 percentage point increase in the proportion of mathematics multiple-choice questions decreases girls' performance in closed- and open-response questions by 0.036 and 0.024 of a standard deviation, respectively.

I show that receiving a higher proportion of multiple-choice questions has a negative effect on females' level of effort during the test. To measure student effort, I follow Akyol et al. (2018) and Anaya and Zamarro (2020) and I use information on omission rates and time spent for each question. First, I identify students who omit questions even if they have enough time left to answer. Second, I identify students who answer questions too rapidly. To answer a question appropriately, individuals 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 low effort. Consistent with the literature, I find that boys exert lower effort in tests than girls.[5] Yet, a higher proportion of multiple-choice questions has gender differential effects on the engagement level of students. Girls decrease their level of effort when they face an exam with more multiple-choice items: they skip more questions, and/or they answer them too rapidly. I show that the number of questions, the order of questions in the booklet, and gender differences in writing and motor skills do not drive the gender differential effect of the proportion of multiple-choice questions on students' performance.

I provide suggestive evidence of the role of students' confidence and self-stereotypes in explaining the boys advantage in multiple-choice questions. The concept of self-stereotypes refers to females' greater sensitivity to negative performance feedback in male-dominated fields. Specifically, in quantitative domains girls are more likely than boys to attribute negative feedback to their own lower ability (Dweck et al., 1978), and less likely to identify themselves within those fields (Steele, 1997; Spencer et al., 1999). Previous literature has shown that daughters whose mothers work in STEM-related occupations have greater confidence in mathematics and are less likely to believe boys are better than girls in mathematics (Oguzoglu and Ozbeklik, 2016; Bowden et al., 2018; van der Vleuten et al., 2018; Bertrand,

[5]PISA is a low-stakes exam for students, as their performance has no direct impact on their future educational outcomes. As a consequence, students' incentives to perform well in the test can be minimal and vary across students. The previous literature documents that boys exert less effort than girls in low-stakes exams (Attali et al., 2011; Buser et al., 2014; Azmat et al., 2016).

2019). I show that indeed the negative effect of the proportion of multiple-choice questions on girls' performance disappears for students whose mothers work in STEM-related occupations. In addition, I show that the effect of multiple-choice questions disappears in reading, a domain where girls are known to outperform boys. There are several explanations for why the gender gap in mathematics varies with the proportion of multiple-choice questions featured in an exam. First, multiple-choice items allow for guessing, either randomly or based on partial knowledge. Since there is no penalty for incorrect responses in the PISA test, students' tendency to guess depends on their beliefs about their knowledge of mathematics (Ben-Shakhar and Sinai, 1991).[6] Coffman et al. (2019) document that students' beliefs about their own ability are related to gender-stereotypes: conditional on underlying competencies, boys are more confident in male-type domains, while girls are more confident about their ability in female-type domains.[7] Second, multiple-choice questions allow for unintended corrective feedback: students are able to realize their computation is incorrect once their solution does not appear among the set of possible choices (Bridgeman, 1992). Boys and girls may have different responses to negative unintended corrective feedback in mathematics, due to the self-stereotypes that women face in traditionally male-type domains.

My results contribute to the literature on female performance in multiple-choice assessment. Several papers document that multiple-choice tests with a penalty for answering incorrectly discriminate against women (Baldiga, 2014; Riener and Wagner, 2017; Conde-Ruiz et al., 2020; Coffman and Klinowski, 2020). Indeed, 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 the gender gap in favor of men. I provide evidence that multiple-choice questions have a negative effect on girls' mathematics performance, even in a context where penalties for answering incorrectly do not apply. In addition, this paper provides evidence of the spillover effects of multiple-choice questions on other questions in the exam.

The rest of the paper is organized as follows. Section 2 describes the PISA exams and the data and provides a descriptive analysis of the gender difference in performance by format. Section 3 presents the identification strategy and reports the main results. Section 4 employs question-response data to investigate the mechanisms behind the results. Section 5 concludes.

[6] According to Ben-Shakhar and Sinai (1991), males are more likely to answer when uncertain, but individual guessing tendencies depend on other situational factors, such as the test instructions, time pressure, the content, and difficulty of the items.

[7] Cho (2017) shows that boys are more confident than girls about their knowledge in both real and fictional mathematics concepts, using data from the PISA 2012 assessment.

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 65 countries (OECD, 2014). 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.

PISA 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 assessments (Schleicher, 2019). The test is designed to compare 15-years-old students' performance across countries and over time. Nevertheless, PISA is a low-stakes exam for students, as their performance on the PISA exam has no direct consequences on any educational outcomes.

PISA test assesses students' competencies in three domains: mathematics, reading, and science.[8] In 2015, computer-based exams were administered for the first time as the main mode of assessment.[9] In this paper, I focus on mathematics performance in 2015 for students completing the computer-based assessment.[10]

There are several reasons why I decide to focus on mathematics performance. First, mathematics is the domain with a higher variation for all three formats of questions: multiple-choice, closed- and open-response questions.[11] Second, mathematics is the domain where the gender gap in favor of boys is wider. In 2015, in most countries, boys outperform girls in mathematics, especially among top-achieving students (Peña-López et al., 2016). In contrast,

[8]PISA is performed every three years. Each year one domain is assessed in depth. 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.

[9]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, 2017).

[10]I use the results from 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 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.

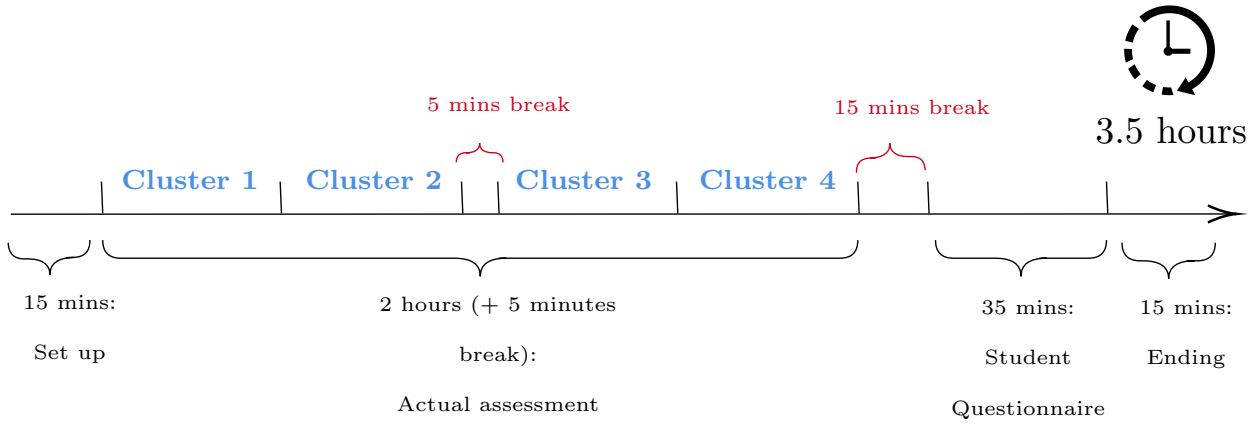
[11]Figure A1 shows the proportion of questions in the three formats by booklets both in reading and science. There are on average only 7% of closed-response reading questions, with 3 out of 13 combinations of clusters with no closed-response questions. In science there are on average only 4% of closed-response items, with several combinations of clusters with no closed-response questions.

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.

The computer-based assessment has the advantage of containing information on students' response time in each question. In addition, 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 159,211 students who answer at least one mathematics cluster and for whom information regarding the gender and parental education and occupation are available.

Figure I shows the timeline of the PISA test. The total assessment last approximately three and a half hours. The formal computer-based exam is designed as two-hour tests. The exam combines four 30-minutes sections, called clusters, each one assessing a particular domain. After the first two clusters students are entitled to a short 5 minutes break. Students answer a 35 minutes questionnaire at the end of the formal assessment.

Figure I: Time-line of PISA Computer-Based Assessment in 2015



Different groups of students receive different exam booklets, chosen among a pool of 396 different ones.[12] These booklets are different ordered combinations of 7 mathematics, 7 reading, and 12 science clusters. Booklets are assigned to students randomly. In particular, each student receives two random numbers. The first number, CC, assigns one of the possible 66

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

standard booklet forms. These booklet forms determine the clusters' order and the exact non-science clusters.[13] The second number, S, determines the exact science clusters combination. Therefore, the combination of the first and second number determines, for each student, the combination of the exact cluster.[14]

PISA test includes 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 not constrain on the length of the response. Figures A5, A6, and A7 display example of the three formats of questions. 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 (von Schrader and Ansley, 2006). Each student needs to answer the questions in the order they are provided, and students do not receive any feedback about their performance in the test.

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).

PISA employs Item Response Theory to estimate students' performance. In particular, PISA 2015 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. Item Response Theory 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.[15]

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, 2017). The trained test administrators cannot be

[13] Figure A2 shows the different booklet forms for 2015 assessment.

[14] Figure A3 and A4 show the clusters for each possible combination of the two numbers.

[15] 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.

teachers of participating students. At the beginning of the exam, each student is allocated to a workspace with a computer and received a unique logon form. During the exam, a staff member of the school monitors the students. 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.

2.1 Descriptive Statistics

As above mentioned, different students receive a different booklet among a set of 396 possible ones. These booklets can contain up to two mathematics clusters. Therefore, some students receive an exam booklet with two of the 7 possible mathematics clusters, some students receive an exam booklet with only one mathematics cluster, while others receive a booklet with no mathematics clusters, but only reading and science ones. For my main analysis, I use data on students that receive at least one mathematics cluster. This assignment determines 18 different combinations of mathematics clusters. Figure 1 shows the proportion of questions for each format on these different booklet combinations. On average, booklets have 44.7% multiple-choice questions, but this percentage varies from 29 to 70%.

Each mathematics booklet has on average 12 mathematics questions. Table 1 shows the summary statistics for all mathematics questions in PISA 2015 exams. Out of 81 mathematics questions, 34 are multiple-choice, 26 closed-response, and 21 open-response. On average, multiple-choice questions are easier than other formats. 50.82% of students answer incorrectly multiple-choice questions, while 60.06 and 70.26% of students answer incorrectly closed-response and open-response questions respectively.

Questions of different formats assess different content, context, as well as different cognitive processes. Figure 2 displays the proportion of questions by format and questions characteristics. In general, multiple-choice questions are more likely to refer to quantity related content. Open-response questions are much less likely to refer to quantity related content and more likely to involve a change and relationships content. With respect to the context, multiple-choice questions tend to refer to societal context, while open-response questions are more likely to involve scientific ones. Different formats require students to employ different cognitive processes. In particular, multiple-choice questions are more likely to assess students' ability to employ mathematical concepts, or interpreting, applying and evaluating ideas. Yet, only a small number of multiple-choice questions require students to formulate situations mathematically.

2.2 Descriptive Investigation: Students' Performance by Format of Questions

The educational literature traditionally recognizes males greater performance in multiple-choice items, and females greater performance in close- or open-response questions (Bolger and Kellaghan, 1990; DeMars, 2000; Willingham and Cole, 2013; Lindberg et al., 2010). These differences may be driven by the different areas of assessment of multiple-choice and close- or open-response questions. In this section, I employ question-level performance and response data, to investigate whether gender difference in performance by format persists once several questions characteristics are controlled for. In particular, I estimate the following model:

$$\mathrm{Y}_{isq} = \gamma_1 + \gamma_2 \mathrm{Female}_{is} + \gamma_3 \mathrm{MC}_q + \gamma_4 \mathrm{Female}_{is} \cdot \mathrm{MC}_q + X'_{is}\Gamma + Z'_q\Theta + s_s + \varepsilon_{isq} \qquad (1)$$

where Y_{isq} indicates whether the mathematics question q is answered correctly by student i in school s, and the time undertaken to answer; MC_q is a dummy variable indicating whether question q has multiple-choice format as opposed to close- or open-response ones. The model includes students' controls, 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, measured as the percentage of students that answer the question incorrectly in all countries. In addition, each specification includes school FE, s_s.

Table 3 reports the results of this descriptive investigation. Overall, girls are 3% less likely to answer correctly a question than boys, and this gap increases to 4.6% for multiple-choice questions. Girls spend more time than boys to answer questions, but for multiple-choice, this difference narrows.

3 Empirical Analysis: the Effect of Proportion of Multiple-choice Questions on Gender Gap in Mathematics

In this section, I analyze whether the proportion of multiple-choice questions featured on an exam affects students' performance differently by gender. I exploit the random assignments of test booklets with different proportions on multiple-choice questions to students. Table 2 provides evidence of the validity of the randomization. There is no correlation between students' demographics characteristics and the percentage of mathematics multiple-choice,

closed-response, and open-response questions they receive in the test.

To analyze whether the proportion of mathematics multiple-choice questions affects the gender gap in performance, I estimate the following model:

$$Y_{isb} = \beta_0 + \beta_1 \text{Female}_{is} + \beta_2 \text{Female}_{is} \cdot \text{Prop. of MC questions}_b + X'_{is}\gamma + b_b + s_s + \varepsilon_{isb} \quad (2)$$

where Y_{isb} represents either the standardize raw exam score or the average of plausible values in mathematics, for student i, attending school s, who receives booklet b. I calculate the exam raw score as the proportion of correct questions that each student answers. 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 model controls for several students' characteristics, such as students' age, grade, migration status, parental education level, and occupational status. Importantly, the model accounts for average booklet characteristics by including booklet FE, b_b. The model also includes school FE, s_s. Standard errors are cluster at the school level.[16]

There are several reasons to employ both the exam raw score (i.e. the proportion of correct questions) and the average of plausible values as outcomes. PISA calculates students' performance using Item Response Theory (IRT). IRT estimates the probability of answering a question correctly as a function of students' skills. It allows the comparison of students' performance among students who did not necessarily answer the same questions, by characterizing items, as well as students' characteristics. PISA provides 10 plausible values. These plausible values are drawn from a posterior distribution obtained by combining IRT and students' information (OECD, 2017). In other words, plausible values come from a distribution of the potential performance of all students with similar characteristics and identical responses to each question. This means that students' plausible values depend not only on the score on the questions a student answers but also on how similar students have performed in different questions (Allen et al., 2005). As my identification strategy compares students who face the same text booklet, plausible values may not be 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' performance 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.[17]

Table 4 shows the results. The first two columns report the results with the standardized raw score as the outcome, while the second two columns report the results for the average of

[16] The model remains robust when standard errors are clustered at the country level.

[17] In main tables, I report both the standardized raw score and the average of plausible values.

plausible values as an outcome. Columns 1 and 3 report the results without including school FE, while columns 2 and 4 include school FE. Column 2 is the preferred specification. The inclusion of school FE does not significantly impact the estimate for β_2. This estimate implies that an increase in the proportion of multiple-choice questions by 10 percentage points differentially reduces girls' scores by 0.026 standard deviations compared to boys. Table A1 estimates the effect of proportions of other formats. The proportions of close- and open-response questions have both differential positive effects on girls' performance compared to boys' ones.

The effect of multiple-choice questions on the gender gap in performance is not small. It is useful to compare this estimate to the baseline gender gap captured by β_1 coefficient. The 0.026 standard deviation differential decrease in girls' performance resulting from a 10 percentage point increase in the proportion of multiple-choice questions is almost half the magnitude of the female coefficient. In addition, the maximum variation in the proportion of multiple-choice questions is 40 percentage points, as it varies from 0.29 to 0.70.

My results are unlikely to be driven by gender differences in motor skills required to type close- and open-ended responses. Table A2 reports the results for PISA 2012 paper-based assessment. Employing data from the paper-based assessment, the table shows the negative effect of the proportion of multiple-choice questions on women's performance in mathematics and science.

3.1 The Effect of Proportion of Multiple-choice Questions on Performance by Formats

In this section, I study the effect of the proportion of multiple-choice questions on the performance in different formats. Table 5 estimates model 2 using the proportion of correct multiple-choice, closed-response and open-response questions as outcomes (columns 1, 2 and 3 respectively). Column 1 reports the standardized score in all questions (as in column 2 of Table 4) for comparison.

The proportion of multiple-choice questions featured in the exam affect the proportion of correct closed-response and open-response questions. An increase in the proportion of multiple-choice questions by 10 percentage points differentially depresses girls' performance compared to boys in closed-response questions by 0.036 standard deviations, and in open-response questions by 0.024 standard deviations (columns 3 and 4 respectively). The estimates for β_2 is not significant in column 2. Importantly, this does not mean that the gender gap in performance is insignificant for multiple-choice questions. Girls' lower score in multiple-choice is shown in table 3. On the contrary, the not significant estimate for β_2 in column 2

indicates that the gender gap in multiple-choice questions in favor of boys remains similar across exams that feature different proportions of multiple-choice questions.

The estimates of the female coefficient provide useful insight. Girls under-performance in mathematics are driven only by lower performance in multiple-choice questions. The β_1 estimate is negative and significant in column 2, where the proportion of correct multiple-choice questions is used as the outcome but is positive or not significant in columns 3 and 4, where the proportion of correct close- and open-response questions are used as outcome.

3.2 Possible Confounding Factors

In the above section, I document a relationship between the proportion of multiple-choice questions and the gender difference in students' performance in mathematics. Several confounding factors could drive my results.

First, the number of questions featured in the cluster could affect students' performance and be correlated with the proportion of multiple-choice questions. Table 3, indicates that women take overall more time to answer each question than boys. If the number of questions in a cluster affects student performance and the time to respond, my main results could be driven by exam booklets that feature a larger proportion of multiple-choice questions having more questions in general. Table 6 shows that clusters with a higher proportion of multiple-choice questions do not have a higher number of questions. Each cluster has 12 or 13 questions. In addition, the proportion of multiple-choice questions in the cluster is not correlated with average questions difficulty, as well as the number of easy, medium-hard, and hard questions in the cluster.

Second, multiple-choice questions may appear in a specific position within the cluster. The education literature document that the position of the question in the exam affects students' performance (Schweizer et al., 2009; Debeer et al., 2014). Moreover, the item position effect varies across men and women. In particular, girls are better than boys to sustain their performance through an exam Battaglia and Hidalgo-Hidalgo (2018); Wu et al. (2019); Balart and Oosterveen (2019). Therefore, the position of multiple-choice questions may represent an omitted variable for my analysis. Table 7 shows that there is no relationship between multiple-choice questions and the question position within the cluster. Hence, my results are unlikely to be driven by the correlation between the format of the item and its position within the cluster.

4 Mechanisms: Students' Engagement Level

For students, PISA is a low-stakes exam. While the PISA results are employed by several stakeholders to evaluate different educational systems or compare the performance of students over time and across countries, they have no direct impact on students' education outcomes (OECD, 2017).

As a consequence, students' incentive to perform well might be minimal, and varying across gender and countries. In this framework, variation in students' performance may be the result of differences in students' knowledge, but also variation in the level of effort exerted. Several studies document gender differences in the level of effort exerts in low-stakes examination (Attali et al., 2011; Buser et al., 2014; Azmat et al., 2016). On one side, the performance of men increases more than women when the stakes of the test increase (Attali et al., 2011; Buser et al., 2014). On the other side, women seem to perform better when they sit 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-stake tests, boys exert much less effort than girls in low-stakes examinations.[18]

In the following part of this section, I show that the proportion of multiple-choice questions has a gender differential effect on students' level of engagement during the examination. I follow the literature to identify low engaged students, namely students who exert low effort in the exam (Akyol et al., 2018; Anaya and Zamarro, 2020). There are two paths to identify low engaged students: looking at omission rate and employing time response data to analyze students' rapid response.

4.1 Students' Omission Behavior

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 effort.[19] For close-

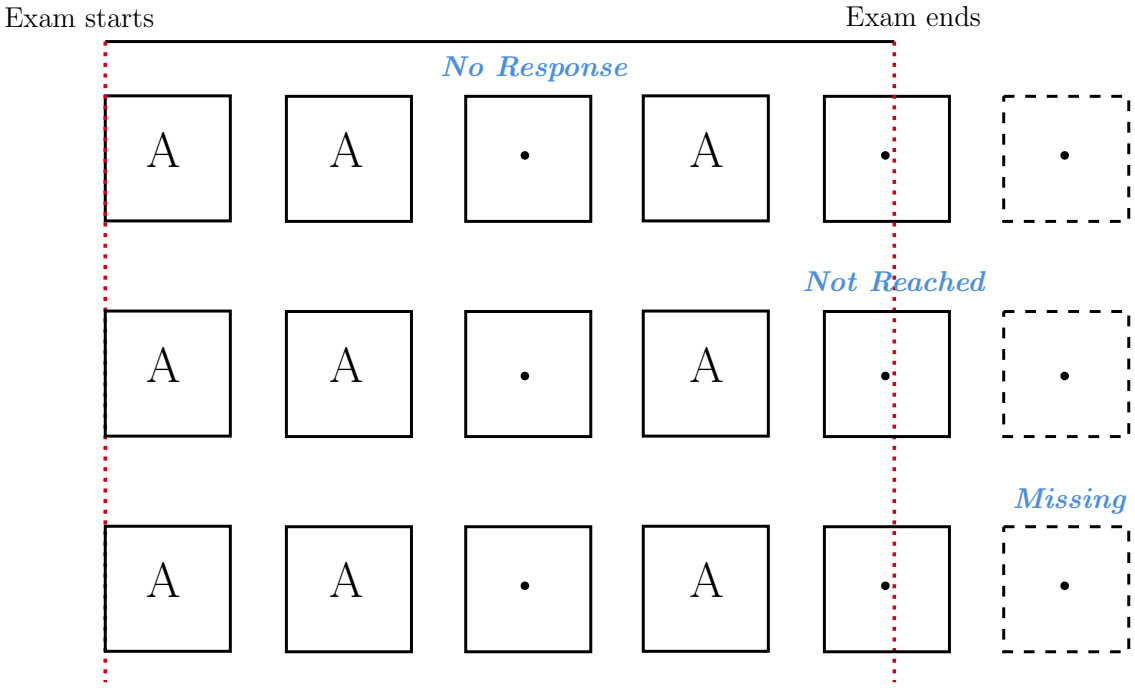
[18]Previous literature documents that variation in students' level of effort explains much of the variation in the performance across gender and countries. In particular, using PISA data Zamarro et al. (2016) and Akyol et al. (2018) provide evidence that accounting for student effort explains between about 30 percent of the differences in performance across countries. Similarly, (Anaya and Zamarro, 2020) reveals that gender differences in students' effort could increase the gender gap in mathematics performance by 6 times in favor of boys.

[19]For closed-response and open-response questions skipping may not be the result of students' low seriousness. Indeed, as the PISA test has a time limit, students may decide to skip these formats of questions to save time to concentrate on other questions. Nevertheless, students usually finish the exam much earlier than the permitted time, so skipping behavior may be considered as a sign of low seriousness even in the context of closed-response and open-response questions.

and open-response questions time contains could lead students to omit some answers. Nevertheless, even if students have 30 minutes to answer each cluster, time is not a binding constraint for most students.[20] Consequently, omitting any question could be interpreted as a sign of students' low effort. Indeed, Akyol et al. (2018) document that skipping behavior increases with question order within the exam clusters. They argue that, as there is no correlation between questions' difficulty and questions' position within the cluster, this pattern is consistent with students skipping questions as a sign of reducing exam effort.

PISA adopts specific terminologies for questions that have not been answered. Figure II helps understanding this terminology. A question is defined as no response if a student spends some time on it but decides to move on to the next question without answering it. A question is marked as not reached if a student spends some time on it, does not answer, and does not move to the next question as the time is up. A question is defined as missing if a student does not spend any time on it. Therefore, by definition, missing questions can only be at the end of the test, following a non-reached question. Importantly, in computing the plausible value, PISA considered both non-reached and missing questions as not administered to students.

Figure II: Notation for Question without Answer in PISA 2015



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

4.2 Students' Rapid Response

Students need to read and understand the question in order to answer it appropriately. As a consequence, too little time spent on the question could be considered another sign of students' low effort. I follow Akyol et al. (2018) and Wise and Ma (2012) and compute a threshold that determines whether students are answering the questions too rapidly. This threshold is question and country-specific.[21]

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 0.10 the average time spent to answer the question q in country j, $mean_{qj}$. For example, suppose that in country A the average time spent to answer a particular question was 2 minutes (120 seconds). The answer of student I, in country A, for question Q is flag as too-rapid if the response time is less than 12 second. I choose the 0.10 threshold to have approximately 5 percent of questions defined as too-rapid response.

4.3 Low Effort Students

To identify students who exert reducing effort during the exam I employ two criteria, considering both the number of omitted questions (no response, not reached or missing) and the number of questions answered too rapidly. In particular, I consider as low effort a student who does not answer 3 or more questions, even if there is enough time remaining in the cluster (i.e. at least 5 minutes). In addition, I consider as low effort a student who answers 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. These criteria allow me to consider about 10% of students as non-serious, a rate similar to the ones found by Akyol et al. (2018). Table A4 shows the summary statistics of non-serious boys and girls. Consistent with previous literature, boys are significantly more likely than girls to be identified as low effort students. Indeed, the proportion of low effort boys is 9.38%, while the proportion of low effort girls is 8.51%. The gender difference is statistically different from zero.

To study whether the proportion of multiple-choice questions has a differential effect on students' engagement in the exam by gender, I estimate model 2 using a dummy variable that identifies low effort students as an outcome. Table 8 shows the results. Columns 1 and 2 estimate the specification 2 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 1.3% less likely to be identified as low effort than boys.

[21] Note that while Akyol et al. (2018) apply the methodology only on science-related questions, I only consider mathematics questions.

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 student engagement level.

4.4 Heterogeneous Effects with Respect to Maternal Occupation

In the previous sections, I document that women's performance is differentially affected by the proportion of multiple-choice questions they receive in the test. As mentioned in previous sections, students' level of confidence about their mathematics abilities, as well as students' beliefs about their ability to perform well in mathematics are likely to play a role in women under-performance in multiple-choice assessments.

I provide suggestive evidence of the role of confidence and students' beliefs by analyzing the effect of the proportion of multiple-choice questions for two groups of students: students whose mother does and does not work in STEM related occupations. Previous literature has shown that daughters whose mum works in STEM-related occupations have greater confidence in mathematics and are less likely to believe boys are better than girls in math. (Oguzoglu and Ozbeklik, 2016; Bowden et al., 2018; van der Vleuten et al., 2018; Bertrand, 2019). Figure 3 displays the estimate for females and its interaction with the proportion of mathematics questions in model 2. The proportion of multiple-choice questions has a negative effect on female performance only among students whose mother is not employed in STEM-related occupations. The effect of the proportion of multiple-choice is almost zero and not statistically significant among students whose mothers work in STEM-related occupations.

5 Conclusion

Performance in standardized assessment plays a key role in determining future educational and economic opportunities. Standardized tests are used around the world to determine university admission, provision of license and certifications, as well as used to determine the effectiveness of intervention policies and educational inputs. These tests use exclusively or in large part multiple-choice items. This format of question is perceived as efficient and cost-effective, especially when assessments are implemented on a large scale. Nevertheless, multiple-choice items may result in unfair assessments if they favor particular individuals.

In this paper, I provide evidence of a greater gender gap in mathematics performance

on exams that present a higher proportion of multiple-choice items. In particular, a 10 percentage points increase in the proportion of multiple-choice questions widens the gender gap by 50% in favor of boys. The effect of the proportion of multiple-choice questions has spillover effects on the gender gap in the performance in close- and open-response questions. An investigation of the mechanisms reveals that girls become differentially less engaged than boys when they receive an exam with a higher proportion of multiple-choice questions.

The results of my analysis have important policy implications. Multiple-choice questions are widely used in standardized assessments as cost-efficient, and more objective. The use of computer-based assessments makes the scoring of multiple-choice questions significantly less costly than open-response questions. Moreover, as the marking does not require human-coding, the results are seen as more objective and less subjected to score manipulation.

References

Akyol, Ş. P., K. Krishna, and J. Wang (2018). Taking pisa seriously: How accurate are low stakes exams? Technical report, National Bureau of Economic Research.

Allen, N. L., C. A. McClellan, and J. J. Stoeckel (2005). Naep 1999 long-term trend technical analysis report: Three decades of student performance. nces 2005-484. *National Center for Education Statistics*.

Anaya, L. and G. Zamarro (2020). The role of student effort on performance in pisa: Revisiting the gender gap in achievement.

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.

Attali, Y., Z. Neeman, and A. Schlosser (2011). Rise to the challenge or not give a damn: differential performance in high vs. low stakes tests.

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.

Battaglia, M. and M. Hidalgo-Hidalgo (2018). Test performance and remedial education: Good news for girls.

Beede, D. N., T. A. Julian, D. Langdon, G. McKittrick, B. Khan, and M. E. Doms (2011). Women in stem: A gender gap to innovation.

Ben-Shakhar, G. and Y. Sinai (1991). Gender differences in multiple-choice tests: the role of differential guessing tendencies. *Journal of Educational Measurement 28*(1), 23–35.

Bertrand, M. (2019). The gender socialization of children growing up in nontraditional families. In *AEA Papers and Proceedings*, Volume 109, pp. 115–21.

Bolger, N. and T. Kellaghan (1990). Method of measurement and gender differences in scholastic achievement. *Journal of Educational Measurement 27*(2), 165–174.

Bowden, M., J. P. Bartkowski, X. Xu, and R. Lewis Jr (2018). Parental occupation and the gender math gap: Examining the social reproduction of academic advantage among elementary and middle school students. *Social Sciences 7*(1), 6.

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

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.

Card, D. and A. A. Payne (2017). High school choices and the gender gap in stem. *Economic Inquiry*.

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

Cho, S.-Y. (2017). Gender and confidence–evidence from the pisa math test.

Coffman, K. B., M. Collis, and L. Kulkarni (2019). *Stereotypes and belief updating*. Harvard Business School.

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*.

Conde-Ruiz, J., J. Ganuza, M. García, et al. (2020). Gender gap and multiple choice exams in public selection processes. Technical report, FEDEA.

Debeer, D., J. Buchholz, J. Hartig, and R. Janssen (2014). Student, school, and country differences in sustained test-taking effort in the 2009 pisa reading assessment. *Journal of Educational and Behavioral Statistics 39*(6), 502–523.

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education 13*(1), 55–77.

Duquennois, C. (2019). Fictional money, real costs: Impacts of financial salience on disadvantaged students.

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.

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

Freedle, R. (2003). Correcting the sat's ethnic and social-class bias: A method for reestimating sat scores. *Harvard Educational Review 73*(1), 1–43.

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

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

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

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

Liu, O. L. and M. Wilson (2009). Gender differences in large-scale math assessments: Pisa trend 2000 and 2003. *Applied Measurement in Education 22*(2), 164–184.

OECD (2014). Pisa 2012 technical report.

OECD (2017). Pisa 2015 technical report.

Oguzoglu, U. and S. Ozbeklik (2016). Like father, like daughter (unless there is a son): Parental occupational investment and stem field choice in college. Technical report, IZA Discussion Paper.

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.

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.

Schleicher, A. (2019). Pisa 2018: Insights and interpretations. *OECD Publishing*.

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.

Taylor, C. S. and Y. Lee (2012). Gender dif in reading and mathematics tests with mixed item formats. *Applied Measurement in Education 25*(3), 246–280.

van der Vleuten, M., E. Jaspers, I. Maas, and T. van der Lippe (2018). Intergenerational transmission of gender segregation: How parents' occupational field affects gender differences in field of study choices. *British Educational Research Journal 44*(2), 294–318.

von Schrader, S. and T. Ansley (2006). Sex differences in the tendency to omit items on multiple-choice tests: 1980–2000. *Applied Measurement in Education 19*(1), 41–65.

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

Wise, S. L. and L. Ma (2012). Setting response time thresholds for a cat item pool: The normative threshold method. In *annual meeting of the National Council on Measurement in Education, Vancouver, Canada*.

Wu, Q., D. Debeer, J. Buchholz, J. Hartig, and R. Janssen (2019). Predictors of individual performance changes related to item positions in pisa assessments. *Large-scale Assessments in Education 7*(1), 5.

Zamarro, G., C. Hitt, and I. Mendez (2016). When students don't care: Reexamining international differences in achievement and non-cognitive skills.

Table 1: Summary Statistics for Mathematics Questions in PISA 2015

Questions' Characteristics:	**Format of the Question**			
Content	**Multiple-Choice**	**Closed-Response**	**Open-Response**	**Total**
Change and Relationship	6	6	8	20
Quantity	11	8	2	21
Space and Shape	7	7	5	19
Uncertainty and Data	10	5	6	21
Context				
Occupational	6	8	6	20
Personal	7	4	2	13
Scientific	6	5	9	20
Societal	15	9	4	28
Process				
Employ	14	14	7	35
Formulate	6	10	7	23
Interpret	14	2	7	23
Total	34	26	21	81
Question Difficulty (% of international incorrect)	50.82	60.06	70.26	58.83

Table 2: Proportion of Different Formats of Questions and Students Characteristics

	Mathematics		
	(1) % of Multiple-Choice Questions	(2) % of Close-Response Questions	(3) % of Open-Response Questions
Female	0.030	-0.047	0.018
	(0.063)	(0.046)	(0.022)
Age in Months	-0.022	0.010	0.012
	(0.111)	(0.080)	(0.039)
Grade compared to modal grade in country	-0.005	0.002	0.003
	(0.059)	(0.043)	(0.020)
Immigration Status: First-Generation	-0.040	0.037	0.003
	(0.145)	(0.106)	(0.050)
Immigration Status: Second-Generation	0.071	-0.060	-0.010
	(0.143)	(0.104)	(0.049)
Mother's Highest Education	-0.026	0.016	0.010
	(0.025)	(0.018)	(0.009)
Father's Highest Education	0.001	0.002	-0.003
	(0.024)	(0.017)	(0.008)
Highest parental occupational status	0.002	-0.000	-0.001**
	(0.002)	(0.001)	(0.001)
Home Possession Index	0.021	-0.014	-0.007
	(0.039)	(0.028)	(0.014)
Obs	159,211	159,211	159,211
Mean Y	44.21	30.50	25.28
St Dev Y	11.17	8.12	3.91
School FE	Yes	Yes	Yes
F Statistics	0.30	0.32	0.69
P-Value for Model	0.975	0.969	0.717

Notes: Standard errors are in parenthesis. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. Observation are at student level. The percentage of multiple-choice, closed-response and open-response questions in the booklet is a value from 0 to 100. The explanatory variables include student's demographics characteristics, such as a dummy for female, the age of the student in months, the grade that the student is attending compares to the modal grade for 15-year-old students in the country, and dummies for whether the student is a first or second generation immigrant, as oppose to native. In addition, the model includes parental information, such as mother and father highest educational level (measured using ISCED), highest parental occupational status and a summary index for home possession (which among other includes information such as number of books in the house, whether the student has a desk to study, internet connection, etc.)

Table 3: Performance and Time Response by Gender and Format in PISA 2015

	Mathematics		
	(1)	(2)	(3)
	Correct Answer	Correct Answer Conditional on Answering	Time (Mins)
Female	-0.020***	-0.022***	0.038***
	(0.002)	(0.002)	(0.004)
Multiple-choice	0.011***	0.002*	-0.177***
	(0.001)	(0.001)	(0.003)
Female ×	-0.019***	-0.020***	-0.003
Multiple-choice	(0.001)	(0.001)	(0.004)
Open-response	0.012***	0.036***	0.360***
	(0.001)	(0.001)	(0.005)
Female ×	-0.006***	-0.008***	0.069***
Open-response	(0.002)	(0.002)	(0.006)
Obs	1,837,520	1,695,620	1,837,520
Mean Y	0.47	0.50	1.33
St Dev Y	0.50	0.50	1.12
School FE	Yes	Yes	Yes
Booklet FE	No	No	No
Student Controls	Yes	Yes	Yes
Question Difficulty	Yes	Yes	Yes
Question Controls	Yes	Yes	Yes

Notes: Observation are at student-question level. Each specification is estimated using linear-model and controls for question characteristics (difficulty, content, context and process) and 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). The omitted category is closed-response question. Standard errors, in parenthesis, are clustered at school level. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table 4: Proportion of Multiple-choice Questions and Gender Gap: PISA 2015

	Std. Raw Score in Mathematics		Av. Plausible Values in Mathematics	
	(1)	(2)	(3)	(4)
Female	-0.029	-0.057***	-9.133***	-11.760***
	(0.020)	(0.016)	(1.873)	(1.338)
Female ×	-0.272***	-0.264***	-11.781***	-11.188***
Prop. Math Multiple-choice Questions	(0.041)	(0.036)	(3.718)	(2.890)
Obs	159,211	159,211	159,211	159,211
Mean Y	0.00	0.00	472.82	472.82
St Dev Y	1.00	1.00	96.74	96.74
School FE	No	Yes	No	Yes
Booklet FE	Yes	Yes	Yes	Yes
Students' Controls'	Yes	Yes	Yes	Yes
R-sq	0.17	0.50	0.24	0.62
Raw Gender Gap	-0.14	-0.14	-12.65	-12.65

Notes: Observation are at student level. 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. The proportion of multiple-choice questions in each domain ranges from 0 to 1. Standard errors, in parenthesis, are clustered at school level. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table 5: Proportion of Multiple-Choice Questions and Gender Gap by Format

	Std. Raw Score in Mathematics			
	(1)	(2)	(3)	(4)
	All Questions	Multiple Choice	Close Response	Open Response
Female	-0.057***	-0.193***	0.038**	-0.022
	(0.016)	(0.018)	(0.019)	(0.018)
Female ×	-0.264***	0.017	-0.363***	-0.240***
Prop. Math Multiple-choice Questions	(0.036)	(0.039)	(0.045)	(0.040)
Obs	159,211	159,138	159,138	159,211
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
Student Controls	Yes	Yes	Yes	Yes
R-sq	0.50	0.41	0.43	0.44
Raw Gender Gap	-0.14	-0.13	-0.06	-0.08

Notes: Observation are at student level. 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. The proportion of multiple-choice questions in each domain ranges from 0 to 1. Standard errors, in parenthesis, are clustered at school level. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table 6: 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
Constant	13.151***	42.518***	4.138**	5.137**	3.875***
	(0.544)	(5.239)	(1.121)	(1.363)	(0.540)
Prop. of	-3.690**	-3.019	-1.324	-0.321	-2.045
Multiple-choice Questions	(1.191)	(11.467)	(2.454)	(2.984)	(1.181)
Obs	7	7	7	7	7
Mean Y	11.57	41.23	3.57	5.00	3.00
St Dev Y	0.79	4.46	0.98	1.15	0.58

Notes: Observation are at cluster level. The proportion of multiple-choice questions in each domain ranges from 0 to 1. The definition of easy, medium and hard question are computed from proficiency level provided by PISA (OECD, 2017). In particular, a question is define as easy if means to assess students proficiency level 1, and 2. A question is define as medium if aims to assess students proficiency level 3, and 4, while hard if aims to assess students proficiency level 5 and 6. Standard errors, are in parenthesis. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table 7: Correlation between Item Format and Item Position within Cluster

	Multiple-choice Question	
Constant	0.538***	0.400***
	(0.116)	(0.112)
Sequence in cluster	-0.018	
	(0.016)	
Question Order= 1st to 3th		0.124
(Omitted: Question Order= 10th to 13th)		(0.157)
Question Order= 4th to 6th		-0.050
(Omitted: Question Order= 10th to 13th)		(0.159)
Question Order= 7th to 9th		0.000
(Omitted: Question Order= 10th to 13th)		(0.159)
Obs	81	81

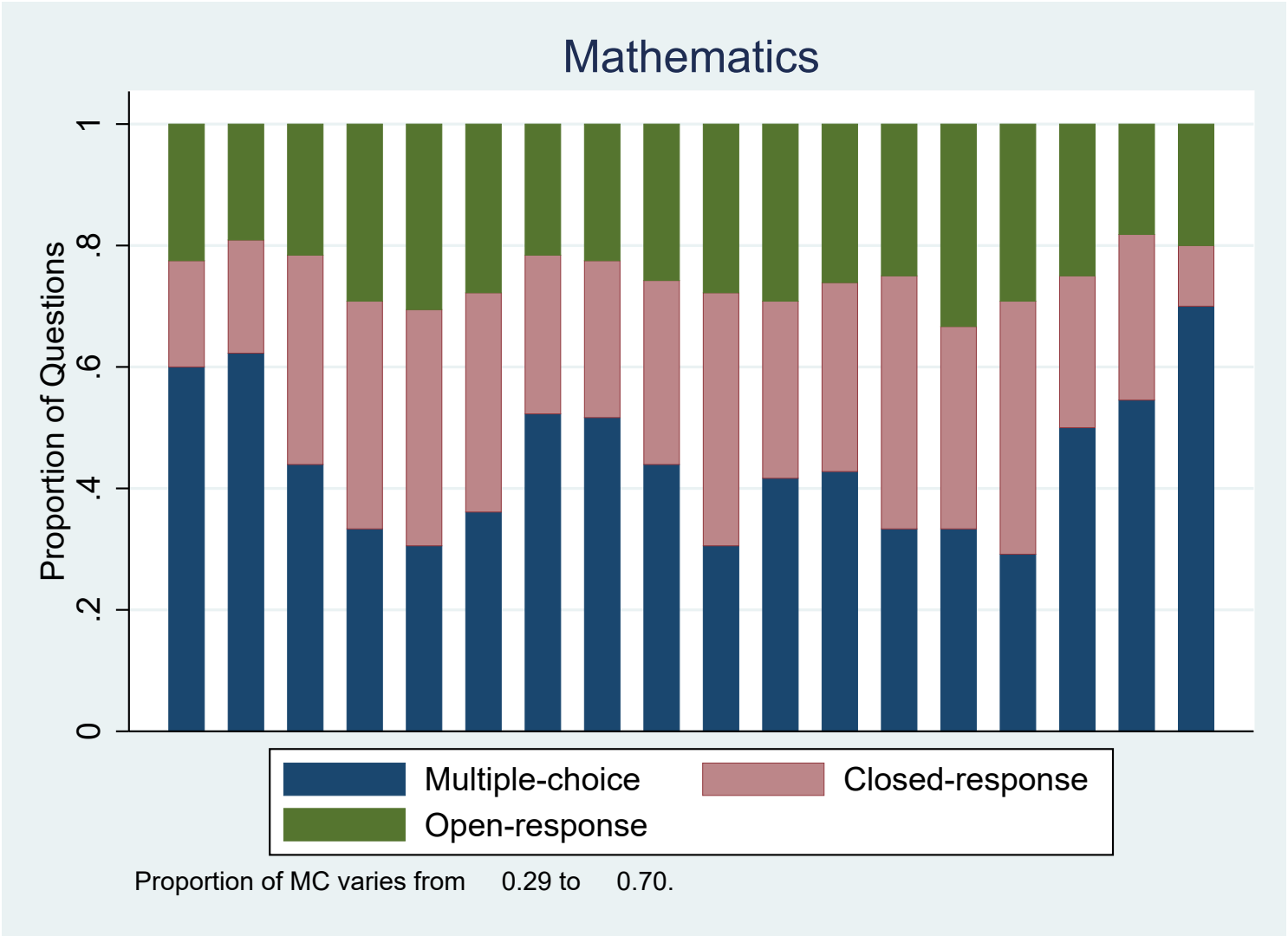
Notes: Observation are at question level. The outcome variable is a dummy variable indicating whether the question is a multiple-choice questions as opposed to close- or open-response one. Standard errors, are in parenthesis. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table 8: The Effect of the Proportion of Multiple-Choice Questions on Students' Engagement

	Low-effort Student		
	(1) OLS	(2) OLS	(3) Logit (dy/dx)
Female	-0.017***	-0.012***	-0.013***
	(0.004)	(0.004)	(0.004)
Female x Prop. of Multiple-choice Math Questions	0.031***	0.026***	0.024***
	(0.008)	(0.009)	(0.009)
Obs	155,636	155,619	155,636
Mean Y	0.04	0.04	0.04
St Dev Y	0.19	0.19	0.19
School FE	No	Yes	No
Booklet FE	Yes	Yes	Yes
Student Controls	Yes	Yes	Yes

Notes: Observation are at student level. 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. The proportion of multiple-choice questions in each domain ranges from 0 to 1. Columns 1 and 2 use OLS estimation, while column 3 estimate the specification 2 using Logit model and report the marginal effects. The outcome variable is a dummy variable equal to 1 if student is identify as low effort ones. A student is classified as low effort if he does not answer 3 or more questions, even if there is enough time remaining in the cluster (i.e. at least 5 minutes), or he 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. Standard errors, in parenthesis, are clustered at school level. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Figure 1: Variation in the Proportion of Question by Formats in PISA 2015



This figure shows the variation in the proportion of multiple-choice, closed-response, and open-response question in the 18 different combination of mathematics clusters. The proportion of multiple-choice questions varies from 0.29 to 0.70, with mean 0.45 and standard deviation 0.13.

Figure 2: Characteristics of Mathematics Question by Format in PISA 2015

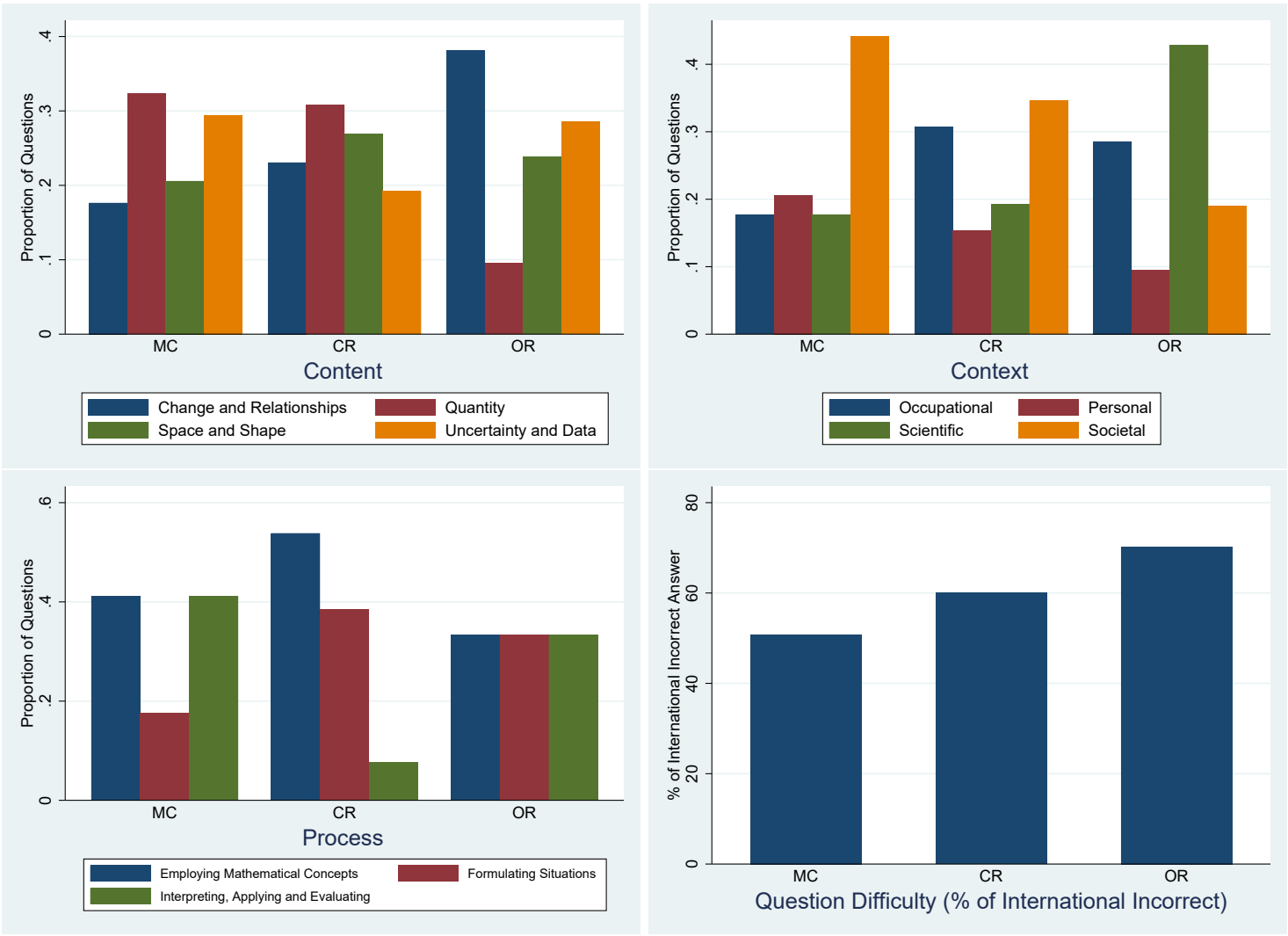
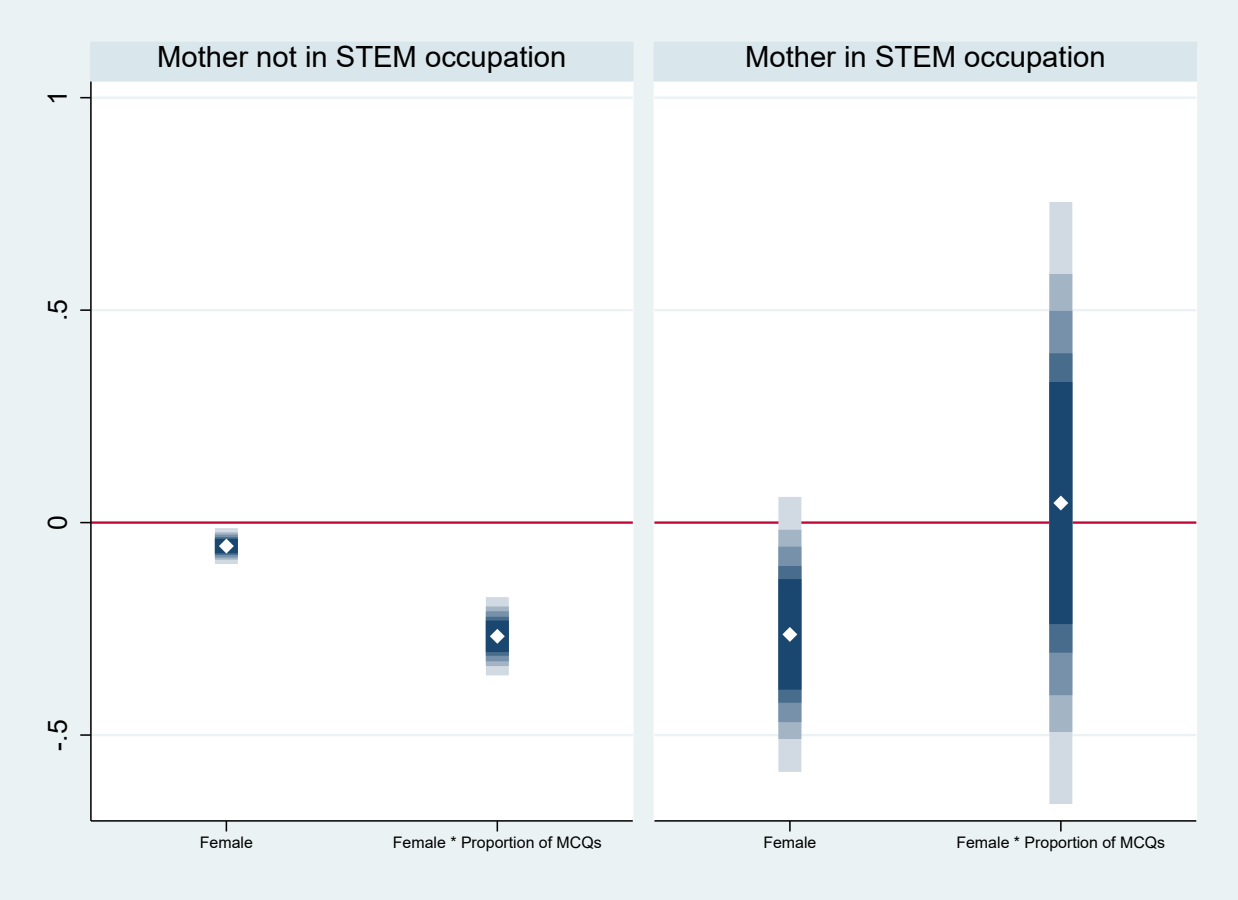


Figure 3: Heterogeneous Effects with Respect to Maternal Occupation



This figure shows the estimate for female and its interaction with proportion of mathematics question in model 2. The definition of STEM jobs follow the definition provided by the European Parliament (www.europarl.europa.eu/RegData/etudes/STUD/2015/542199/IPOL_STU(2015)542199_EN.pdf)

Appendix

Table A1: Proportions of Closed- and Open-response Questions and Gender Gap

	Std. Raw Score in Mathematics		Av. Plausible Values in Mathematics	
	(1)	(2)	(3)	(4)
Female	-0.294***	-0.289***	-22.000***	-20.893***
	(0.016)	(0.026)	(1.249)	(2.126)
Female ×	0.395***		17.349***	
Prop. Math Close-response Questions	(0.049)		(3.962)	
Female ×		0.455***		16.544**
Prop. Math Open-response Questions		(0.102)		(8.332)
Obs	159,211	159,211	159,211	159,211
Mean Y	0.00	0.00	472.82	472.82
St Dev Y	1.00	1.00	96.74	96.74
School FE	Yes	Yes	Yes	Yes
Booklet FE	Yes	Yes	Yes	Yes
Students' Controls'	Yes	Yes	Yes	Yes
R-sq	0.50	0.50	0.62	0.62
Raw Gender Gap	-0.14	-0.14	-12.65	-12.65

Notes: Observation are at student level. 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. The proportion of multiple-choice questions in each domain ranges from 0 to 1. Standard errors, in parenthesis, are clustered at school level. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table A2: The Effect of Proportion of Multiple-choice on Performance in Different Domains and Waves

	2012						2015			
	(1) Std. Raw Score in Maths	(2) Av. Plausible in Maths	(3) Std. Raw Score in Reading	(4) Av. Plausible in Reading	(5) Std. Raw Score in Science	(6) Av. Plausible in Science	(7) Std. Raw Score in Reading	(8) Av. Plausible in Reading	(9) Std. Raw Score in Science	(10) Av. Plausible in Science
Female	-0.052***	-13.060***	-0.033	24.779***	0.196***	0.998	0.174***	17.957***	-0.160***	-11.417***
	(0.009)	(0.710)	(0.023)	(1.970)	(0.031)	(2.636)	(0.029)	(2.470)	(0.022)	(1.940)
Female	-0.265***	-10.793***								
× Prop. of Maths Multiple-choice	(0.021)	(1.616)								
Female			0.481***	6.037			-0.064	-4.656		
× Prop. of Reading Multiple-choice			(0.050)	(4.261)			(0.064)	(5.375)		
Female					-0.431***	-14.440***			0.031	-1.341
× Prop. of Science Multiple-choice					(0.048)	(4.083)			(0.033)	(2.936)
Obs	341,291	341,291	242,911	242,911	234,069	234,069	156,986	156,986	375,507	375,507
Mean Y	0.00	480.47	0.00	479.41	-0.00	486.15	0.00	476.41	0.00	477.92
St Dev Y	1.00	100.35	1.00	98.55	1.00	98.61	1.00	98.76	1.00	98.23
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Booklet FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Students' Controls'										
R-sq	0.48	0.61	0.48	0.60	0.47	0.58	0.48	0.57	0.47	0.54
Raw Gender Gap	-0.13	-14.82	0.20	30.42	-0.06	-6.29	0.17	18.24	-0.11	-9.11

Notes: Observation are at student level. 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. The proportion of multiple-choice questions in each domain ranges from 0 to 1. Standard errors, in parenthesis, are clustered at school level. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table A3: **Descriptive Statistics: Proportions of Questions answered Too Fast by Gender and Formats**

	Male	Female	Difference	*p-value*
	(1)	(2)	(3)	(4)
Answered Too Fast (All Formats)	0.0497	0.0399	-0.0098	0.0000
Answered Too Fast (MC Questions)	0.0426	0.0354	-0.0072	0.0000
Answered Too Fast (CR Questions)	0.0475	0.0397	-0.0078	0.0000
Answered Too Fast (OR Questions)	0.0627	0.0479	-0.0148	0.0000

Notes: I report summary statistics for male and female students (columns 1 and 2, respectively); the gender difference between column (2) and (1) (column 3); and p-values for the t-test on the gender difference (column 4).

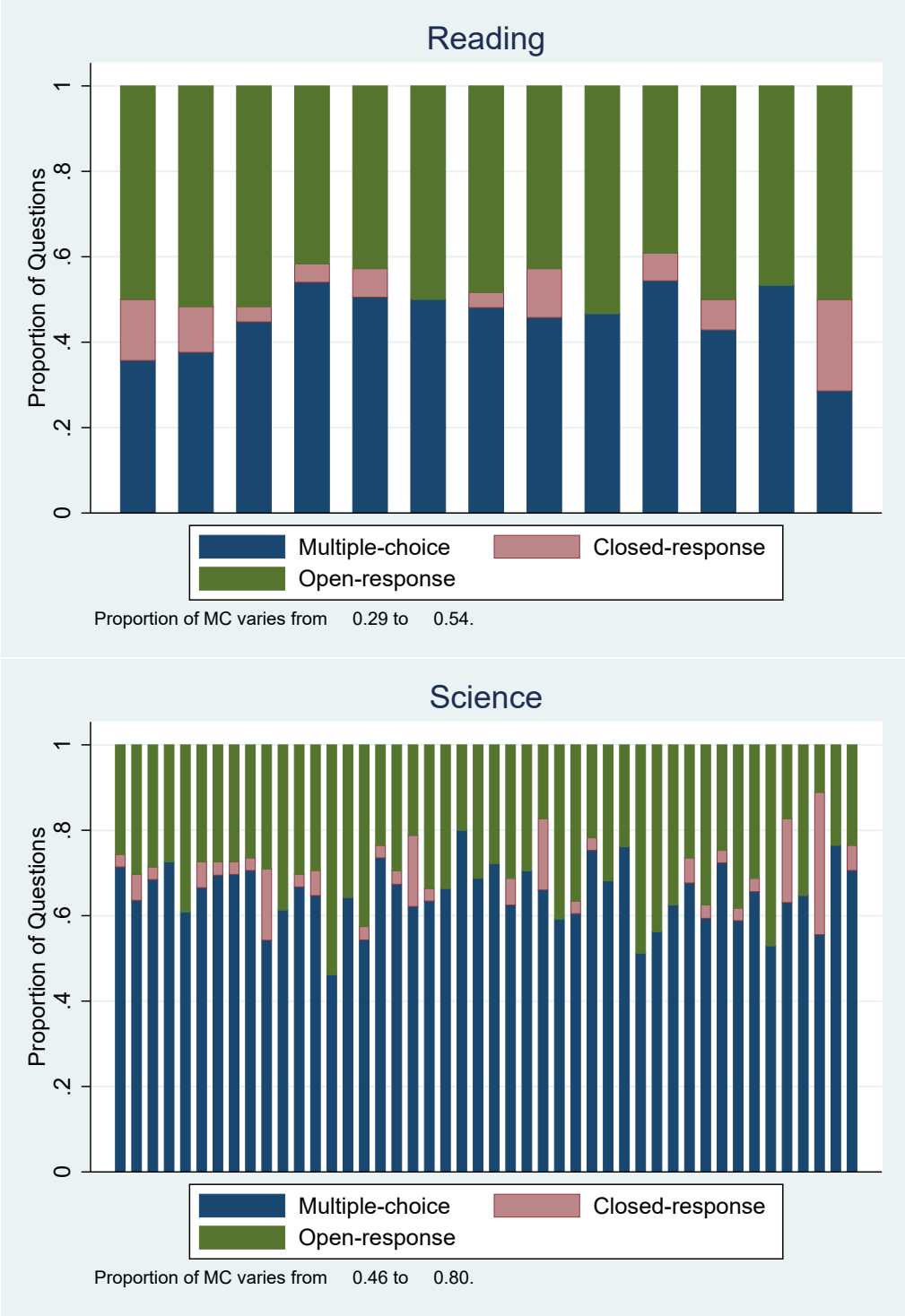
Table A4: **Descriptive Statistics: Low Effort Students**

	Male	Female	Difference	*p-value*
	(1)	(2)	(3)	(4)
Low-Efforts Student	0.0938	0.0851	-0.0087	0.0000

Notes: I report summary statistics for male and female students (columns 1 and 2, respectively); the gender difference between column (2) and (1) (column 3); and p-values for the t-test on the gender difference (column 4). A student is classified as low effort if he does not answer 3 or more questions, even if there is enough time remaining in the cluster (i.e. at least 5 minutes), or he 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.

Appendix Figures

Figure A1: Variation in the Proportion of Question by Formats in Reading and Science: PISA 2015



This figure shows the variation in the proportion of multiple-choice, closed-response, and open-response question in the different combination of reading and science clusters.

Figure A2: The Cluster Rotation Design Used to Form Standardized Test Booklets for PISA 2015

Forms	Cluster 1	Cluster 2	Cluster 3	Cluster 4
31	S	S	R01	R02
32	S	S	R02	R03
33	S	S	R03	R04
34	S	S	R04	R05
35	S	S	R05	R06ab
36	S	S	R06ab	R01
37	R01	R03	S	S
38	R02	R04	S	S
39	R03	R05	S	S
40	R04	R06ab	S	S
41	R05	R01	S	S
42	R06ab	R02	S	S
43	S	S	M01	M02
44	S	S	M02	M03
45	S	S	M03	M04
46	S	S	M04	M05
47	S	S	M05	M06ab
48	S	S	M06ab	M01
49	M01	M03	S	S
50	M02	M04	S	S
51	M03	M05	S	S
52	M04	M06ab	S	S
53	M05	M01	S	S
54	M06ab	M02	S	S
55	S	S	M01	R01
56	S	S	R02	M02
57	S	S	M03	R03
58	S	S	R04	M04
59	S	S	M05	R05
60	S	S	R06ab	M06ab
61	R01	M01	S	S
62	M02	R02	S	S
63	R03	M03	S	S
64	M04	R04	S	S
65	R05	M05	S	S
66	M06ab	R06ab	S	S
67	S	S	C01	M01
68	S	S	M02	C02
69	S	S	C03	M03
70	S	S	M04	C03
71	S	S	C02	M05
72	S	S	M06ab	C01
73	M01	C02	S	S
74	C03	M02	S	S
75	M03	C01	S	S
76	C01	M04	S	S
77	M05	C03	S	S
78	C02	M06ab	S	S
79	S	S	R01	C01
80	S	S	C02	R02
81	S	S	R03	C03
82	S	S	C03	R04
83	S	S	R05	C02
84	S	S	C01	R06ab
85	C02	R01	S	S
86	R02	C03	S	S
87	C01	R03	S	S
88	R04	C01	S	S
89	C03	R05	S	S
90	R06ab	C02	S	S
91	S	S	C01	C02
92	S	S	C02	C03
93	S	S	C03	C01
94	C02	C01	S	S
95	C03	C02	S	S
96	C01	C03	S	S

Source: https://www.oecd.org/pisa/data/2015-technical-report/PISA2015_TechRep_Final.pdf.

Figure A3: Base Form (CC) and Random number (S) Science cluster combination: PISA 2015

Base form (CC)	Random number (S)					
	1	2	3	4	5	6
31	1	13	6	9	22	25
32	2	16	12	10	31	32
33	11	5	17	14	26	29
34	35	4	7	19	23	30
35	34	15	8	20	24	28
36	3	36	18	21	27	33
37	35	4	7	19	23	30
38	34	15	8	20	24	28
39	3	36	18	21	27	33
40	1	13	6	9	22	25
41	2	16	12	10	31	32
42	11	5	17	14	26	29
43	1	13	6	9	22	25
44	2	16	12	10	31	32
45	11	5	17	14	26	29
46	35	4	7	19	23	30
47	34	15	8	20	24	28
48	3	36	18	21	27	33
49	35	4	7	19	23	30
50	34	15	8	20	24	28
51	3	36	18	21	27	33
52	1	13	6	9	22	25
53	2	16	12	10	31	32
54	11	5	17	14	26	29
55	1	13	6	9	22	25
56	2	16	12	10	31	32
57	11	5	17	14	26	29
58	35	4	7	19	23	30
59	34	15	8	20	24	28
60	3	36	18	21	27	33
61	35	4	7	19	23	30
62	34	15	8	20	24	28
63	3	36	18	21	27	33

Base form (CC)	Random number (S)					
	1	2	3	4	5	6
64	1	13	6	9	22	25
65	2	16	12	10	31	32
66	11	5	17	14	26	29
67	1	13	6	9	22	25
68	2	16	12	10	31	32
69	11	5	17	14	26	29
70	35	4	7	19	23	30
71	34	15	8	20	24	28
72	3	36	18	21	27	33
73	35	4	7	19	23	30
74	34	15	8	20	24	28
75	3	36	18	21	27	33
76	1	13	6	9	22	25
77	2	16	12	10	31	32
78	11	5	17	14	26	29
79	1	13	6	9	22	25
80	2	16	12	10	31	32
81	11	5	17	14	26	29
82	35	4	7	19	23	30
83	34	15	8	20	24	28
84	3	36	18	21	27	33
85	35	4	7	19	23	30
86	34	15	8	20	24	28
87	3	36	18	21	27	33
88	1	13	6	9	22	25
89	2	16	12	10	31	32
90	11	5	17	14	26	29
91	1	13	6	9	22	25
92	2	16	12	10	31	32
93	11	5	17	14	26	29
94	35	4	7	19	23	30
95	34	15	8	20	24	28
96	3	36	18	21	27	33

Source: https://www.oecd.org/pisa/data/2015-technical-report/PISA2015_TechRep_Final.pdf.

Figure A4: Computer-based assessment Science clusters combination: PISA 2015

Science cluster combination		
N	S	S
1	S01	S07
2	S01	S10
3	S02	S08
4	S03	S09
5	S03	S12
6	S04	S07
7	S04	S10
8	S05	S11
9	S06	S12
10	S07	S06
11	S08	S01
12	S08	S05
13	S09	S02
14	S09	S06
15	S10	S03
16	S11	S02
17	S11	S04
18	S12	S05

Science cluster combination		
N	S	S
19	S07	S08
20	S07	S09
21	S07	S11
22	S08	S10
23	S08	S12
24	S09	S08
25	S09	S11
26	S10	S07
27	S10	S09
28	S10	S12
29	S11	S08
30	S11	S10
31	S12	S07
32	S12	S09
33	S12	S11
34	S02	S04
35	S05	S01
36	S06	S03

Source: https://www.oecd.org/pisa/data/2015-technical-report/PISA2015_TechRep_Final.pdf.

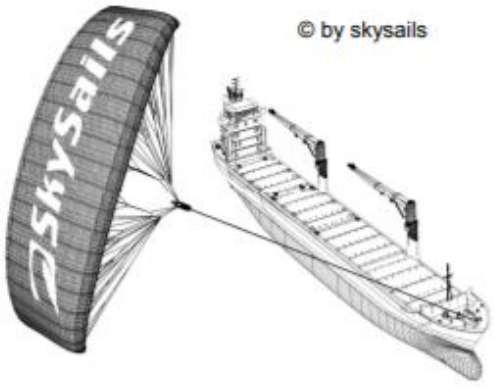
Figure A5: Example of Multiple-Choice Question

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

Source: https://www.oecd.org/pisa/pisaproducts/pisa2012-2006-rel-items-maths-ENG.pdf

Figure A6: Example of Closed-Response Question in PISA 2015

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

Source: https://www.oecd.org/pisa/pisaproducts/pisa2012-2006-rel-items-maths-ENG.pdf

Figure A7: Example of Open-Response Question in PISA 2015

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:.....

Source: https://www.oecd.org/pisa/pisaproducts/pisa2012-2006-rel-items-maths-ENG.pdf