DISCUSSION PAPER SERIES

DP13541

BRAVE BOYS AND PLAY-IT-SAFE GIRLS: GENDER DIFFERENCES IN WILLINGNESS TO GUESS IN A LARGE SCALE NATURAL FIELD EXPERIMENT

Nagore Iriberri and Pedro Rey-Biel

LABOUR ECONOMICS AND PUBLIC ECONOMICS

ISSN 0265-8003

BRAVE BOYS AND PLAY-IT-SAFE GIRLS: GENDER DIFFERENCES IN WILLINGNESS TO GUESS IN A LARGE SCALE NATURAL FIELD EXPERIMENT

Nagore Iriberri and Pedro Rey-Biel

Discussion Paper DP13541
Published 21 February 2019
Submitted 19 February 2019

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

BRAVE BOYS AND PLAY-IT-SAFE GIRLS: GENDER DIFFERENCES IN WILLINGNESS TO GUESS IN A LARGE SCALE NATURAL FIELD EXPERIMENT

Abstract

We study gender differences in willingness to guess using approximately 10,000 multiple-choice math tests, where for half of the questions, both wrong answers and omitted questions are scored 0, and for the other half, wrong answers are scored 0 but omitted questions are scored +1. Using a within-participant regression analysis, we find that female participants leave significantly more omitted questions than males when there is a reward for omitted questions. This gender difference, which is stronger among high ability and older participants, hurts female performance as measured by the final score and position in the ranking. In a subsequent survey, female participants showed lower levels of confidence and higher risk aversion, which may explain this differential behavior. When both are considered, risk aversion is the main factor explaining the gender differential in the willingness to guess. A scoring rule that is gender neutral must use non-differential scoring between wrong answers and omitted questions.

JEL Classification: C93, D81, I20, J16

Keywords: gender differences, willingness to guess, risk preferences, confidence, perceived ability in math, natural field experiment

Nagore Iriberri - nagore.iriberri@gmail.com
University of the Basque Country and CEPR

Pedro Rey-Biel - pedro.rey@esade.edu
ESADE

Acknowledgements
We thank researchers at various institutions and conferences for their helpful comments and feedback. We are thankful to the organizers of Concurso de Primavera de Matemáticas, who collaborated throughout the project. In particular, we would like to honor the memory of Joaquín Hernández, who put great effort and passion into this initiative and who unfortunately left us too soon. We also thank Adrià Bronchal, Lluis Puig and Ana Montiel for their excellent research assistance. Nagore Iriberri acknowledges financial support from the Departamento de Educación, Política Lingüística y Cultura del Gobierno Vasco (IT869- 13) and Ministerio de Economía y Competitividad y Fondo Europeo de Desarrollo Regional (ECO2015- 66027-P). Pedro Rey-Biel acknowledges financial support from Programa Ramón y Cajal and ESADE.

Brave Boys and *Play-it-Safe* Girls:

Gender Differences in Willingness to Guess in a Large Scale Natural Field Experiment*

Nagore Iriberri+ and Pedro Rey-Biel**

February 14 2019

Abstract

We study gender differences in willingness to guess using approximately 10,000 multiple-choice math tests, where for half of the questions, both wrong answers and omitted questions are scored 0, and for the other half, wrong answers are scored 0 but omitted questions are scored +1. Using a within-participant regression analysis, we find that female participants leave significantly more omitted questions than males when there is a reward for omitted questions. This gender difference, which is stronger among high ability and older participants, hurts female performance as measured by the final score and position in the ranking. In a subsequent survey, female participants showed lower levels of confidence and higher risk aversion, which may explain this differential behavior. When both are considered, risk aversion is the main factor explaining the gender differential in the willingness to guess. A scoring rule that is gender neutral must use non-differential scoring between wrong answers and omitted questions.

Keywords: gender differences, willingness to guess, risk preferences, confidence, perceived ability in math, natural field experiment.

JEL classification: C93, D81, I20, J16.

---

* We thank researchers at various institutions and conferences for their helpful comments and feedback. We are thankful to the organizers of *Concurso de Primavera de Matemáticas*, who collaborated throughout the project. In particular, we would like to honor the memory of Joaquín Hernández, who put great effort and passion into this initiative and who unfortunately left us too soon. We also thank Adrià Bronchal, Lluis Puig and Ana Montiel for their excellent research assistance. Nagore Iriberri acknowledges financial support from the Departamento de Educación, Política Lingüística y Cultura del Gobierno Vasco (IT869-13) and Ministerio de Economía y Competitividad y Fondo Europeo de Desarrollo Regional (ECO2015-66027-P). Pedro Rey-Biel acknowledges financial support from Programa Ramón y Cajal and ESADE.

+ University of the Basque Country, IKERBASQUE, Basque Foundation for Science. E-mail nagore.iriberri@gmail.com.

** Universitat Ramon Llull, ESADE. Email: pedro.rey@esade.edu.

Introduction

Multiple-choice tests are one of the most frequently used means to measure individuals' knowledge and aptitude. On top of their common use in everyday academic life, performance on some multiple-choice tests plays a crucial role in shaping labor market outcomes. For example, the Scholastic Aptitude Test (SAT) in the USA and the Graduate Record Examination (GRE), used worldwide, are standardized multiple-choice tests that play a key role in shaping students' future outcomes. Similarly, licensing exams for many professions, such as in Medicine and Law, are also based on multiple-choice tests: United States Medical Licensing Examination (USMLE), Medico Interno Residente (MIR) in Spain, and bar exams in Law. One crucial decision a multiple-choice test designer needs to make is whether wrong and omitted questions (questions with no answer) are scored the same. GRE and USMLE do not have differential grading, while MIR in Spain does. An interesting example is the SAT, which used to have differential grading but has changed recently to non-differential scoring rule.

The main motivation for scoring wrong answers and omitted questions differently is to prevent test takers from getting the question right by chance, which would add noise to the measure of knowledge and aptitude. However, one important concern is that multiple-choice tests in which incorrect answers are scored differently than omitted questions may lead individuals with different degrees of confidence and/or risk aversion to follow different strategies when answering, which might also misrepresent those students' knowledge and aptitude. An extensive body of literature has documented that women are on average more risk averse (Eckel and Grossman, 2008, Croson and Gneezy, 2009, and Filippin and Crosetto, 2016) and less confident (Beyer, 1999, and Barber and Odean, 2001) than men. Hence, an informed decision on the optimal scoring rule regarding omitted questions and wrong answers requires study of its effect on gender differences in willingness to guess and ultimately on performance.

In collaboration with the organizers of *Concurso de Primavera de Matemáticas,* we conduct a large-scale natural field experiment to test for and understand the mechanism behind gender differences in willingness to guess. *Concurso de Primavera de Matemáticas* is a regional math contest in which primary education, secondary education and high school students from the region of Madrid participate annually. In the 2016, 2017 and 2018 contests, which had a total of approximately 10,000 test takers, we designed tests with no differential score between omitted questions and wrong answers for the first 13 of 25 test questions (no reward for omitted), and for the last 12 test questions, wrong answers were scored 0 and omitted questions +1 (reward for omitted). We compare within-participant willingness to guess and overall performance across both parts of the test.

We find that, though the dominant strategy is to answer all questions when there is no reward, female participants leave on average more omitted questions than males in the no reward section (0.17 standard deviations of the mean). Most importantly, this difference increases (by an additional 0.14 standard deviations of the mean) in the reward section of the test. Moreover, the gender differential in willingness to guess has important consequences for the gender differences in the final scores and the ranking of the participants. Females tend to show lower performance on the math test (0.21 standard deviations of the mean), which leads females to lag approximately 51 positions behind in the ranking of test takers (with an average of approximately 1000 positions). This female underperformance increases by 0.05 standard deviations of the mean and they lose approximately 10 additional positions in the ranking under the differential scoring rule for omitted questions and wrong answers.

We explore two heterogeneity effects. First, using two different measures of ability (math grade at school and the number of correct answers when there is no reward), we test whether the gender gap in willingness to guess varies with ability. As expected, high-ability participants leave fewer omitted questions than low-ability ones. However, we find that the gender differential for willingness to guess is indeed stronger among high-ability participants (0.26 standard deviations of the mean), while we find no significant gender differential for the low-ability participants. Second, using four different age categories, we explore the differential gender effect of the scoring rule across different ages. Participants in their final years of high school (16-17 years old) show a significantly higher gender differential between the reward and no-reward parts of the test than younger participants (10-11 years old).

Motivated by the gender difference we found in the 2016 data and to understand the underlying mechanism, we designed a questionnaire that would allow us to measure the effects of confidence and risk aversion, which we administered with the 2017 and 2018 editions of the test. Regarding confidence, we use two measures: confidence in their perceived math ability and the difference between their guessed number of correct answers and the actual number of correct answers, which we label overconfidence. The following item is used to measure risk: "When omitting a question was worth 1 point, I answered the question ….", to which the participants could give 5 different answers: "when I was absolutely sure" (safest option), "when I was almost sure", "when I was considering 2 potential answers", "when I was considering 3 potential answers" and "always" (riskiest option). Female participants show lower confidence in their perceived math ability, lower overconfidence and higher risk aversion. When we control for these measures of confidence, overconfidence and risk, we find that gender differences in risk aversion explain most of the gender differences in the willingness to guess.

Previous literature has shown that women omit more questions than men when there is a differential scoring rule for wrong answers and omitted questions, mostly based on observational data (Swineford, 1941; Anderson, 1989; Atkins et al., 1991; Ramos and Lambating, 1996; Tannenbaum D., 2012, Akyol, Key, and Krishna, 2016). Only recently have there been important advances in pursuing randomized controlled trials in the laboratory (Baldiga, 2014), in the field (Ben-Shakhar and Sinai, 1991, Espinosa and Gardeazabal, 2013, and Funk and Perrone, 2016) and using before-after quasi-controlled studies (Coffman and Klinowski, 2018) to test for the causal effect of differential scoring rules on male and female test takers' willingness to guess and performance. Although all these studies find that female students leave more omitted questions than males when there is differential grading of wrong answers and omitted questions, there is disagreement over whether this differential grading hurts females or not. On the one hand, Funk and Perrone (2016) do not find any harmful effect for female students, arguing that females in their sample have higher average ability than males. Akyol, Key and Krishna (2016) estimate negative effects for females and for risk averse students but conclude that the effects are small, making a case for differential scoring of omitted questions and wrong answers. On the other hand, Baldiga (2014) and Coffman and Klinowski (2018) find that a differential scoring rule for omitted questions and wrong answers has a significant negative impact on the gender gap in performance.

Our study differs from existing randomized controlled trials in the following ways. First, the differential scoring rule *rewards* omitted questions rather than *penalizing* wrong answers. Espinosa and Gardeazabal (2013) show that these two approaches are only strategically equivalent under risk neutrality and that under risk aversion, penalties will lead to more omitted questions than rewards. Therefore, in our setting, the significant gender difference in willingness to guess and their non-negligible effects on performance show that they are still an important concern even though they reflect a milder and less non-favorable scoring rule for females than the use of penalties for wrong answers. Second, using a within-participant treatment assignment similar to that in Funk and Perrone (2016), we add a larger sample than those in existing natural field experiments, as well as evidence from a different setting. Funk and Perrone (2016) study the classroom behavior of undergraduate students in a large Microeconomics class, while we study behavior in a large math contest. The study of settings that involve a competitive component might be more informative about behavior on high-stakes tests that determine entry to university and the attainment of professional licenses, as they also have a competitive component. Third, similar in spirit to the laboratory experiment by Baldiga (2014), we also contribute to the understanding of the underlying mechanism and test how much the gender differences in willingness to guess are due to confidence, overconfidence and risk aversion. We indeed reach a similar conclusion that gender differences in risk aversion are the main factor but use a different approach to measure confidence, overconfidence and risk. Finally, most recent studies have shown interesting heterogeneity effects regarding ability differences (Funk and Perrone, 2016, and Akyol, Key and Krishna, 2016). We find that high-ability female participants are indeed more affected, which resonates with the results of Akyol, Key and Krishna (2016). We also explore heterogeneity effects regarding age, which to our knowledge no other study has. It is important to understand when gender differences appear and how they evolve with age.

The paper has the following structure. Section 2 describes the setting, the data and the main descriptive statistics. Section 3 shows the main results, heterogeneity results and the study of the underlying mechanism. Section 4 concludes.

2 The Data

2.1 The Setting: Mathematics Test

The Mathematics Department of Universidad Complutense de Madrid has been organizing annually since 1996 a regional math contest, *Concurso de Primavera de Matemáticas,* in the Madrid region of Spain.[1] As explained on the department's website,

[1] See the organization's website at https://www.concursoprimavera.es/#concurso for more details.

the contest has two main goals: to "motivate a large number of students by showing them that thinking and studying math can be fun," and "to promote thinking outside the box and textbooks when solving problems, using logical reasoning, class geometry, parity issues, the properties of numbers, and probability." It is a two-stage elimination math contest. Each year, approximately 40,000 students participate in the first stage math test and over 3,000 students in the second stage math test. Iriberri and Rey-Biel (forthcoming) analyzed gender differences between the two stages, which differ in terms of competitive pressure, using the 2014 data. In this study, we use data from one unique math test, the second-stage math test, from the 2016, 2017 and 2018 contests.

A large number of schools in Madrid participate in this initiative. As shown in Iriberri and Rey-Biel (forthcoming), the sample of participating schools ranges between 30% of the primary education schools and 50% of the secondary education schools in the region (see Table A.1 in Iriberri and Rey-Biel, forthcoming). Regarding the school characteristics, only a small proportion of the participating schools are public schools; they tend to have a relatively large number of students and, as expected, show better results in mathematics, as measured by the standardized test administered and evaluated by the Department of Education in the region of Madrid.[2]

The rules of the math test we study are clearly established. First, there are four different tests, one for each age group. These are referred to as levels 1 to 4 and are grouped such that students from two consecutive school years take the same math test. Thus, level 1 includes children in their fifth and sixth academic years of primary school, and participants are therefore aged 10 and 11. Similarly, level 2 includes 12-13-year-olds, level 3 includes 14-15-year-olds and level 4 includes 16-17-year-olds. Second, the math test takes place on the campus of Universidad Complutense de Madrid on a pre-specified day in April. Third, the top three contestants in each level obtain prizes. Additionally, the top 5% of participants receive a diploma and a small gift in a public ceremony.[3] Fourth, the test for each level consists of 25 multiple-choice questions, all of which are set by the organizers. The questions for each level are designed so that students in the lower school

[2] In particular, we use the standardized test called "Conocimientos y Destrezas Indispensables" (CDI – "Essential Knowledge & Skills"), which includes the subjects of Math, Spanish Language and General Culture. For more information, see http://www.educa2.madrid.org/web/cdi/pruebas-cdi

[3] As indicated on the website, what the main prizes will be is not revealed ex-ante. In past years, prizes were scientific calculators or iPads, and the gifts for the top 5% in stage 2 were books. The most important reward is the prestige associated with being among the top 5% of all contestants, which is publicly announced on the website and in a public award ceremony.

year in each level have already seen the material necessary to answer the questions correctly.

Each question has 5 possible answers, only one of which is correct. Up to 2015, the scoring rule was the same for all 25 questions: 0 for wrong answers, +1 point for omitted questions, and +5 points for correctly answered questions. For the 2016, 2017 and 2018 contests, we collaborated with the organizers to create a math test with two parts that would differ in terms of the scoring rule. For the first 13 questions, the grading system awards 0 points for both omitted questions and wrong answers and +5 points for questions answered correctly. For the remaining 12 questions, questions 14-25, the grading system awards 0 points for wrong answers, +1 point for omitted questions and +5 points for questions answered correctly. Figure A1 in the Appendix shows how the scoring rule was described to participants.

We explicitly instructed the organizers to keep other things the same, i.e., the content or difficulty of the questions. The mean values of correct answers per question for all the questions on the math test are presented in Figure A2, where we do not see differences between the two parts of the test (a *t*-test of the differences in the proportion of correct answers between the first and the second half of the test yields a *p*-value equal to 1). As additional evidence that the findings can be attributed to the change in the scoring rule and not to the first/sooner and second/later parts of the tests, the placebo test we performed using the 2013, 2014 and 2015 editions shows no differential scoring rule across the test. This allows us to rule out that males and females react differently to the first/sooner and second/later parts of the test (see the results in Table A2 in the Appendix, discussed in Section 3.1).

Finally, after studying the performance results for the 2016 contest and to better understand the underlying mechanism, we administered a questionnaire immediately after the end of the math test to the participants in 2017 and 2018. Figure A3 in the Appendix includes an English version of the questionnaire. The first five questions listed were used in Iriberri and Rey-Biel (forthcoming), as they focused on the differences between the stage 1 and stage 2 tests. We included questions 6 to 10 to understand whether gender differences in hours of preparation, confidence, overconfidence, risk preferences and perceived math ability can explain any of the gender differences observed in the number of omitted questions.

2.2 Descriptive Statistics

The database consists of the participants who took the 2016, 2017 and 2018 editions of the test. Table 1 shows the descriptive statistics of the main outcome and control variables, overall and by gender. The last column shows the *p*-values for the F-test of equality of variances across gender for the continuous variables and Fisher's exact test for categorical values.

[Table 1 about here]

Panel A shows the variables from the math test. This database contains a total of 9,906 math tests from 7,833 different participants. It is not a gender balanced sample, as 66% of the test takers are male. Looking at the control variables, we see that some students participate in multiple contests (*Participation Time*). In particular, 183 participants take the math test in all three years, 1,167 participate twice, and the remaining 7,023 take the math test just once. Female participants are less likely to participate more than once. The three different contests do not show large differences in overall participation or female participation. Regarding participation in different levels, level 2 is the most popular, and level 4 has the lowest number of participants. Female representation is also lowest in the last level, which is partially explained by female students being less likely to choose the math-science track in high school.

The performance data include the rank, score, and number of correct and omitted questions for each part of the test. When students register to take the math test, schools are asked to provide participants' math grade at school, which is available for approximately 90% of participants. For regression analysis, we will use the standardized math grade at school level in order to control for softer and more stringent schools. As expected, participants have on average high grades in math (*Math at School*), with an average of 8.40 out of 10, and female students indeed show higher performance than males (8.55 for females and 8.32 for males). However, the gender differences reverse when looking at the score on the math test we study, as on both parts of the test, female participants obtain a lower score than male participants. On the first part of the test, when there is no reward for omitted questions, male participants obtain an average score of 29.50 points, and females obtain an average score of 26.50 (out of the maximum score of 65). On the second part, when there is a reward for omitted questions, males obtain 23.30 points on average, while females obtain 20.67 points (out of the maximum core of 60).

The slight difference in score between the first and the second part of the test is because the first part has 13 questions, while the second part has 12 questions. This also carries over into the ranking between male and female participants. Females rank lower than males, on average approximately 51 positions behind (with an average of approximately 1000 positions), and this difference increases for the math test with the reward for omitted questions, where female participants rank on average 64 positions behind.

[Figure 1 about here]

The number of omitted questions, which is the focus of this paper, shows clear gender differences between the no reward and the reward parts of the test, which is consistent with the mainstream literature. Figure 1 shows the cumulative distribution of the *No. of Omitted* by gender when there is no reward (top) and when there is a reward for omitted questions (bottom), which complements the descriptive statistics in Table 1 well. Note that when there is no reward, the optimum behavior is to answer all questions, but when there is a reward for omitted questions, the optimum behavior depends on one's knowledge, confidence and risk aversion. Although participants should answer all questions when there is no reward for omitted questions, participants indeed omit on average 0.65 questions. In addition, women leave slightly more questions unanswered, at 0.86 questions; thus, while 80% of male participants indeed answer all questions, only 74% of female participants do. More importantly, when there is a reward, participants on average leave 4.82 questions unanswered, with male participants leaving 4.51 questions unanswered and females 5.40. In both panels of Figure 1, the distribution of female participants stochastically dominates that of male participants, and the differences on the reward part of the test are larger. In Table 1, we see that male participants also have a higher number of correct answers and a higher proportion of correct answers than females, but these differences are not large across the two parts.

[Figure 2 about here]

Finally, panel b in Table 1 shows the descriptive statistics of the control variables we collected in the survey administered with the 2017 and 2018 test editions. The variables of interest are *No. of Preparation Hours*, *Overconfidence*, *Perceived Math Ability*, *Perceived Gender Nature of Math* and *Risk*. All these variables show significant gender differences with one exception. Male and female participants show a very similar number of reported hours devoted to preparing for the test (see question 5 in Figure A3 in the Appendix).[4] Figure 2a shows the probability density function of the number of preparation hours by gender, truncated at the value of 30 because most participants' answers lie below that value, which again shows that both male and female participants devote a similar amount of time to preparing for the test.

Overconfidence is measured by the difference between the guessed number of correct answers (see question 7 in Figure A3 in the Appendix) and the actual number of correct answers; thus, the more positive the value, the higher the overconfidence. Figure 2b plots the observations; the x-axis shows the number of correct answers and the y-axis shows the number of guessed correct answers. Both male and female participants are overconfident (also found by Beyer, 1999). However, also consistent with other findings, as in Barber and Odean (2001), female participants in our setting show lower values of overconfidence than male participants. Note that overconfidence is measuring a lower bound of the gender difference, as it is restricted to the questions that were actually answered. Related to confidence, male participants also show higher agreement with the statement "I am good at math" than female participants, as shown in Figure 2c (see question 9 in Figure A3 in the Appendix), so perceived math ability is higher for male than for female participants. Finally, also related to confidence, we measure participants' perception of the gendered nature of the math task (see question 10 in Figure A3 in the Appendix). As shown in Figure 2d, the large majority of participants (94.44% of them) believe math to be gender neutral, such that men and women are equally good at/knowledgeable about math. However, both genders show some type of bias: a small fraction of male participants believe that men are better at math than women, and a small fraction of female participants believe that women are better at math than men.

Importantly, we measure risk by the following question (see question 8 in Figure A3 in the Appendix): "When omitting a question was worth 1 point I answered the question …." There are 5 possible answers (from 1 for "When I was Absolutely Sure" to 5 for "Always"); the higher the number, the more risk-loving the participant is. Figure 2d shows the histogram of all possible answers by gender. Clearly, more female participants than males answer the test question when absolutely or almost sure. Note that this risk

[4] Fifteen participants reported very high numbers of preparation hours. We replaced those values with missings to avoid outliers.

measure is also affected by confidence, as perceived probability of knowing the answer might also be affected by participants' ability and confidence in their own ability.

Finally, given that our setting is competitive, we also measure participants' non-competitiveness, using their answers to question 1 in the questionnaire: "It is more important for me to be selected for Stage 2 than being among the winners in Stage 2." The more they agree with this statement, the less competitive they are. As expected, and consistent with the literature on gender and competitiveness (Niederle and Vesterlund, 2011), female participants show lower degrees of agreement with this statement, as shown in Figure 2f.

One might be concerned about the correlation between the different measures of confidence, overconfidence and risk, as well as how all these three measures correlate with ability (standardized Math at school level). Regarding the correlations among confidence, overconfidence and risk, all three of them show low correlations, which suggests that they are indeed measuring different dimensions of personality.[5] Regarding their correlations with ability, as one would expect, confidence correlates with ability positively (0.20); overconfidence correlates negatively with ability (-0.14), as does risk-loving preferences, although the correlation is very low (-0.06). With the exception of the overconfidence measure, all other measures show low correlations with ability, which indicates that our risk and confidence measures are independent of ability.

3. Results

3.1. Do Female Participants Leave More Unanswered Questions than Males When There is a Reward for Omitting Questions Compared to When There is No Reward?

We start by looking at whether female participants react differently than males in their strategy to leave a question unanswered, comparing gender differences between the no reward and the reward parts of the test. The outcome variables of interest are the number of omitted questions, the proportion of correct answers and the final score and ranking. We use standardized values by contest year, level and part of the test for all outcome variables. Table 2 shows the estimation results.

[5] In particular, confidence correlates positively with overconfidence (0.0928) and negatively with risk (-0.0382). In addition, overconfidence correlates negatively with risk (-0.0211).

[Table 2 about here]

Columns 1-4 show the estimation results for the OLS specification with the standard errors clustered at the participant level. All regressions control for year, level and school fixed effects. The coefficients of interest are *Female* and, in particular, the interaction between *Female* and *Reward*. Female participants leave on average more omitted questions than males on the no reward part of the test (0.17 standard deviations of the mean), but most importantly, this difference increases 0.14 additional standard deviations of the mean on the reward part of the test. This is not the case for the proportion of correct answers. Though female participants show a lower proportion of correct answers (0.16 standard deviations of the mean), this difference does not increase on the reward part. Female participants leaving more questions unanswered than men on the reward than the no reward part has important consequences for how male and female participants perform under different reward systems. Female participants score on average worse than males (0.21 standard deviations of the mean) and receive lower rankings (51 positions behind) for the no reward part of the test. More importantly, this gap increases when there is a reward for omitted questions. Regarding the score, the gender gap increases by 0.05 standard deviations of the mean. Regarding the ranking, the gender gap increases by approximately 10 more positions. In other words, female underperformance increases when moving from the no reward to the reward part of the test, showing that differential grading of omitted questions and wrong answers hurts women more than non-differential grading for omitted questions and wrong answers.

We close this section by carrying out four robustness tests.

First, columns 5-8 and columns 9-12 in Table 2 show the equivalent estimation results for the random effects and individual fixed effects model specifications. Random effects and individual fixed effects models assume different specifications regarding the error term and therefore allow testing for robustness to the specification of the main effects. The variable of interest, the interaction between *Female* and *Reward*, maintains the same magnitude and significance levels. Hereafter, we will use the OLS estimation, with the standard errors clustered at the participant level.

Second, we comment on the effect of the two main control variables: *Math at School* for ability and *Participation Time* for experience with the math test. We find that the higher the math grade at school, as expected, the better the score and the higher the proportion of correct answers. Unexpectedly, the higher the math grade at school, the higher the number of omitted questions. However, note that in the fixed effects specification (column 9), the math grade is, as one would expect, negatively correlated with the number of omitted questions. In addition, the more experienced the participant, as one would expect, the higher the score, the lower the number of omitted questions and the higher the proportion of correct answers. As a robustness test, Table A1 in the Appendix shows the exact same table but with an alternative measure for ability; instead of *Math at School*, we control for individual ability by the number of correct answers on the no reward part of the test. The results for the main variable of interest, the interaction between *Female* and *Reward*, are very similar in terms of both the magnitude and the significance levels.

Third, choosing the non-differential scoring rule in the first part of the test and the differential scoring rule in the second and not varying this specific order is not ideal. It is possible that the observed effect is due to male and female participants reacting differently to fatigue over time when taking the test, although we could not find any evidence for this type of behavior. If female participants tend to get tired or lose interest in the math test *before* males do, the observed effect would be confounded with gender differences in performance due to fatigue over time. To rule this possibility out, we use the exact equivalent data from the 2013, 2014, and 2015 contests. Remember that in these three years, the scoring rule did not change across the test; thus, we can measure whether male and female participants show differential performance and willingness to guess between the first and second parts of the test. Table A2 in the Appendix shows the results: columns 1 to 4 for the 2013 contest, columns 5-8 for the 2014 contest and columns 9-12 for the 2015 contest. We find no evidence of any gender difference in performance between the different parts of 2013 and 2014 tests. In 2015, females tended to show better performance on the second half of the test. We therefore rule out that the identified effect is due to gender differences in fatigue when taking the test.

Fourth and finally, we already observed that about 25% of participants do not choose an optimal strategy in the no-reward part of the test leaving one or more questions unanswered. Furthermore, we observed that women tend to show higher frequencies for this non-optimal behavior. This poses the question: is the main finding on gender difference in willingness to guess driven mainly by gender differences in confusion and/or non-optimal reaction to incentives? We therefore proceed to replicate the analysis restricting the overall sample to those participants who showed optimal behavior in the no-reward part of the test, that is, those participants who left 0 questions omitted (7,740 out of 9,900 participants). Estimation results in Table A3 show that the effect is not driven by gender differences in confusion and that the magnitude of the effect is even larger (0.33 standard deviations of the man) if we restrict our main analysis to participants who showed optimal behavior in the first part of the test.

3.2. Analysis Along The Ability Distribution: Are High-Ability Female Participants Particularly Affected?

An important source of variation when looking at a large sample of math test takers is ability. There are two possible proxies for ability. First, if we take the number of correct answers on the part with no reward as a proxy for ability, we can observe in Figure 3 that there is large variation. The number of correct answers varies between 0 and 13, with a median of 6. Second, the math grade at school also shows some but definitely less variation, since those selected to participate in the second stage of the contest tend to be the best math students at each school and thus obtain similar grades. Although both measures show a positive correlation (0.22), due to its larger variation, we use the number of correct answers on the no reward part of the test as a proxy for ability and use the variation in standardized math grade at school level as a robustness test, which we will discuss at the end of the section.

[Figure 3 over here]

We now study whether the gender differential in the number of omitted questions between the no reward part of the test and the reward part of the test varies substantially by participants' ability.

[Figure 4 over here]

Figure 4 displays the gender differences by ability graphically. Figure 4a shows the number of omitted questions on the non-differential scoring rule and the differential scoring rule parts of the test by low and high ability and by gender. We define low ability as the standardized number of correct answers on the reward part being below 0 and high ability as the standardized number of correct answers being above 0. As expected, high-ability participants leave fewer omitted questions on both parts of the test. Additionally, female participants always leave more questions unanswered. However, the gender difference between the two parts is larger among the high-ability participants. Figures 4b for low ability and 4c for high ability take a closer look at the number of omitted questions on the reward part of the test by gender. Low- and high-ability female participants behave similarly, although as expected, high-ability females leave fewer questions unanswered. However, for male participants, low- and high-ability participants' behavior differs substantially, particularly with significantly more participants omitting no questions at all, which is less evident for female participants.

[Table 3 about here]

Table 3 shows the regression analysis results for the number of omitted questions. We take two complementary approaches.

First, as shown in columns 1 and 2, we consider a binary category for low and high ability using the standardized value of the number of omitted questions on the no reward part of the test. For the low-ability participants, the gender differential is not significantly different from zero, while for the high-ability participants, it is highly significant and the magnitude is much higher (0.26 standard deviations of the mean) than the average effect we found in Section 3.1 (0.14 standard deviations of the mean). As shown in column 3, the triple interaction of *Female*, *Reward* and *High Ability* is highly significant, and the magnitude corresponds to the difference between the female and reward coefficients in columns 1 and 2.

Second, we also consider a continuous variable of ability, looking at the actual number of omitted questions on the no reward part of the test. Column 4 shows the interaction among *Female*, *Reward* and the *No. of Correct Answers No Reward*, showing, consistent with the results in previous columns, that the gender differential when moving from the no reward part to the reward part is larger among the participants of higher ability (0.0691 standard deviations of the mean).

As a robustness test, we also perform the same exercise but use standardized math grade at school level as an alternative proxy for ability. Table A4 in the Appendix shows the results. The conclusions are very similar when looking at the interaction between *Female* and *Reward* for the low- and high-ability students, although the magnitudes are slightly lower when using standardized math grade at school level as a proxy for ability.

3.3. Analysis Regarding Age: Are Younger/Older Female Participants Equally Affected?

An interesting feature of our sample is that we can observe male and female participants from a young age (in their fifth and sixth academic years of primary school) up to their final two years of high school, right before going to university. Exploiting this variation, we test whether gender differences in willingness to guess vary with age.

[Table 4 about here]

Table 4 shows the results. Columns 1 to 4 show the regression analysis for each of the levels separately. The coefficient of interest, the interaction between *Female* and *Reward*, shows increasing magnitudes from the lowest academic level (0.02 standard deviations of the mean for the youngest participants in their 5^{th} and 6^{th} grade of primary school) to the highest academic level (0.236 among the oldest participants in high school). Column 5 shows the results when all levels are included in one regression to test how different the gender differences are across academic levels. The gender differential among high school participants is significantly different from the gender differential among the youngest participants, although the effect is only significant at the 10% level. We therefore conclude that the gender differential in willingness to guess increases as participants age.

3.4. Underlying Mechanism: Math Ability, Competitiveness, Confidence or Risk?

There are four possible underlying motivations for gender differences in willingness to guess. First, male and female participants may differ in their knowledge of math. We do not find any support for this when looking at the math grades from school, as female participants indeed outperform males in this domain (see Table 1). However, if we look at the number of correct responses on the no reward part of the test, we do see that while male participants obtain approximately 5.90 correct answers, females obtain approximately 5.29 correct answers. Note that, in all our analyses so far, we do control for math ability (using either math grade at school or the number of correct answers on the no reward part of the test), so any differences due to ability are being controlled for. As a further robustness test to account for ability differences between male and female participants, we have replicated the main regression analysis, shown in Table 2, using a subsample where we matched male and female participants with the same number of correct answers on the no reward part of the test. While column 1 in Table A5 in the

Appendix shows that in the overall sample, female participants show worse performance, both on the reward and no reward parts of the test, column 2, by construction, shows that in the matched sample, females are comparable to males in terms of performance on the no reward part of the test, although on the reward part of the test, the gender gap persists. Columns 3-6 in Table A5 in the Appendix replicate the main results for the rest of the outcome variables with the matched sample. Though these male and female participants do not differ in their ability, the gender differential is still significant for the reward part of the test.

Second, male and female participants also show differences in their competitiveness, as shown by Figure 2e, which uses the answers to question 1 in the questionnaire. Given that the setting is competitive by nature, we can look at whether the gender difference in competitiveness drives the gender difference in willingness to guess.

Third, male and female participants may differ in their perceived knowledge of math (confidence). We use three different variables to measure perceived math ability. First, question 9 on the questionnaire asks participants to rate how much they agree with the following statement: "I am good at math". Clearly, as shown in Figure 2c, female participants show lower levels of agreement with that statement. Second, question 10 asks participants to say whether they believe male participants are better/equally good/worse than female participants. As shown in Figure 2d, there seems to be a high degree of agreement among both male and female participants that both male and female participants are equally good. Third, we asked participants to guess the number of questions they answered correctly. Both male and female participants seem to be overconfident (Beyer, 1999), as they expect to get more correct than they actually get. However, again, female participants show lower confidence levels, which is also consistent with previous findings (Barber and Odean, 2001).

Finally, male and female participants might differ in their risk preferences, and again, consistent with previous findings (Eckel and Grossman, 2008, Croson and Gneezy, 2009, and Filippin and Crosetto, 2016) and using a different approach to elicit risk preferences, we indeed find that female participants show higher risk aversion than males.

[Table 5 about here]

We now proceed to test if any of these measures has explanatory power for the gender differential in the number of questions omitted between the no reward and the reward part of the test such that when these variables are controlled for, the female differential is still significant.[6] Table 5 shows the estimation results from this exercise. Note, however, that we collected all these measures in a questionnaire administered immediately after the test in 2017 and 2018, so we do not have these measures for all participants. Columns 1 and 2 show the main specification, as in Table 2, but in the sample for which we have control variables collected via the questionnaire. In column 2, we find that, as in the main sample, female participants omit more questions than males when moving from the no reward to the reward part of the test, although the magnitude is slightly lower than in the main specification. In column 3, we add the three main control variables: perceived math ability, overconfidence and risk, and the three of them have the expected sign. The more confident and the more risk-loving the participant is, the fewer the number of omitted questions. The female coefficient decreases, but the interaction of *Female* and *Reward* remains exactly the same as in column 2.

In columns 4 to 7, we interact each of the control variables with the variable *Reward*. As for competitiveness, the more competitive the participant is the fewer the omitted questions. Importantly, when adding these interactions with respect to the competitiveness measure, the main result on *Female* and *Reward* remains unchanged, which suggests that competitiveness plays no role in explaining the gender difference in willingness to guess (column 4). Similarly, when adding these interactions with respect to the two confidence measures, the main coefficient of interest, the interaction between *Female* and *Reward*, changes very little, suggesting that confidence does not explain why female participants leave more questions unanswered (column 5 and 6). However, when interacting the risk measure with reward, we clearly see that the coefficient of *Female* and *Reward* decreases substantially such that it is no longer significant. This shows that differences in risk preference between male and female participants are indeed the main factor explaining why male and female participants differ in their behavior in omitting questions (column 7).

Table A6 in the appendix shows the estimation results as in Table 5 but with the alternative measure for ability that uses the number of correct answers on the no reward

[6] Given our setting is competitive by nature, we also checked how much of the gender differential in willingness to guess might be explained by participants' attitudes toward competition. Using question 1 in the questionnaire, where participants assess how much they agree with the statement "It is more important to participate in the competition than to win the competition", we find that, although female participants show a less competitive attitude than males, this does not show any explanatory power in how male and female participants decide on their willingness to guess. These results are available upon request.

part of the test instead of the math grade at school. Although the main coefficient of interest becomes non-significant in column 2 when focusing on the sample administered the questionnaire, the results are qualitatively the same, as the interaction between *Female* and *Reward* is positive, though it is half the magnitude of the one we identified in Table 3 and Table A1 in the Appendix. More importantly, and consistent with the overall finding regarding the underlying mechanism, controlling for participants' risk preferences substantially changes (reduces to 1/3 of the original magnitude) the differential reaction of female participants regarding when they respond to questions. We conclude that gender differences in risk aversion are the main mediating factor when explaining gender differences in willingness to guess.

One issue that we have not considered so far is gender differences in time management during the test. It could be the case that female participants need more time and that the gender difference in the number of omitted questions is mostly driven by participants' behavior in the very last part of the test. Given we have performance in all questions in the test, we can indeed split the second part of the test, the las 12 questions, into two differences parts: questions 14-19 (labeled, Reward_14-19) and questions 20-25 (labeled, Reward_20-25). In Table A7 in the Appendix, we can see the estimation results, when we split the reward part into two. The gender difference in the number of omitted seems to be equally present in both the first half and the second half of the reward part of the test, such that we can rule out the gender difference in time management as a plausible underlying mechanism.

What about the gender differences found between the low- and high-ability participants and older participants? In Sections 3.2 and 3.3, we found that the gender difference in the number of omitted questions punished high-ability and older females in particular. Could it be that gender differences in risk and overconfidence are different between the low- and high-ability participants or among the younger/older participants?

We must first examine gender differences in confidence, overconfidence and risk by ability. Figure A4 shows the graphs. Gender differences are present among both the high- and low-ability participants, and they always follow the same pattern: female participants show lower perceived math ability, lower levels of overconfidence and higher risk aversion. However, the gender differences between the low- and high-ability participants are not striking.

[Table 6 about here]

We perform a similar exercise as we do in Table 5 but in two sub-samples, the low- and high-ability participants. Table 6 shows the results. Columns 1 and 2 reproduce the main results found in the first two columns in Table 3, and columns 3 and 4 replicate the same results for the sample of participants for whom we have questionnaire responses. The estimated values of the main variable of interest, the interaction between *Female* and *Reward*, are very similar in the overall sample and the sample for which we have questionnaire responses. Columns 5 and 6 add the main control variables of competitiveness, confidence and risk, and the results do not change significantly. However, when we add the interaction between each of the control variables of confidence and risk, we again see that the interaction between *Female* and *Reward* changes the most when the risk measure is interacted with *Reward*. This again suggests that gender differences in risk preferences underlie the greater gender differences seen among the high-ability participants. However, it is also important to note that, contrary to the main analysis in Table 5, in Table 6, the *Female* and *Reward* interaction does not become insignificant for the high-ability participants when adding risk measures, so some of the differences remain unexplained.

We then look at gender differences in confidence, overconfidence and risk by age, focusing on the most distant age groups: participants in levels 1 and 4. Figure A5 shows the graphs. Interestingly, older participants are less confident, more calibrated (less overconfident) and more risk-loving than the youngest participants. More importantly, gender differences are present both among the younger and older participants. Furthermore, these differences always follow the same pattern: female participants show lower perceived math ability, lower levels of overconfidence and higher risk aversion. However, the gender differences among youngest and oldest participants are not striking.

[Table 7 about here]

We perform a similar exercise as we do in Table 5 but in two sub-samples, the level 1 and level 4 participants. Table 7 shows the results. Columns 1 and 2 reproduce the main results found in the first two columns in Table 4, and columns 3 and 4 replicate the same results for the sample of participants for whom we questionnaire responses. The estimated values of the main variable of interest, the interaction between *Female* and *Reward*, are of very similar magnitude in the overall and the sample for which we have the questionnaire responses, although in the sample with questionnaire responses the value is not significantly different from zero. Columns 5 and 6 add the main control variables of competitiveness, confidence and risk, and the results do not change significantly. However, when we add the interaction between each of the control variables of confidence and risk, we again see that the interaction between *Female* and *Reward* changes the most when the risk measure is interacted with *Reward*. This again suggests that gender differences in risk preferences underlie the greater gender differences among the older participants.

4. Conclusions

Using performance data from a natural field experiment with approximately 10,000 observations, we test for gender differences in willingness to guess when there is differential grading for omitting questions and providing a wrong answer. We find that women always leave more omitted questions but that this behavior becomes even more prominent when there is differential grading for omitted questions and wrong answers. This has negative consequences for female participants, both in terms of their final score and their ranking, demonstrably hurting female performance on the math test.

We also find that this gender differential is stronger among high-ability and older participants. Finally, gender differences in risk aversion explain most of the gender differential in willingness to guess. Based on this evidence, we conclude that a gender neutral grading rule requires non-differential scoring for omitted questions and wrong answers, at least compared to the alternative of having a mild reward for omitting questions. It is still an open question whether gender differences persist, or even if they change sign or size, when the differential score for omitted questions and wrong answers is very large.

References:

Akyol, S.P., Key, J. and Krishna, K. (2016). "Hit or Miss? Test Taking Behavior in Multiple Choice Exams." NBER Working Paper Nr. 22401.

Anderson, J. (1989). "Sex-Related Differences on Objective Tests among Undergraduates." *Educational Studies in Mathematics*, 20(2):165–177.

Atkins, W.J., Leder, G.C., O'Halloran, P.J., Pollard, G.H. and Taylor, P. (1991). "Measuring Risk Taking." *Educational Studies in Mathematics*, 22(3), 297-308.

Baldiga, K. (2014). "Gender Differences in Willingness to Guess." *Management Science*, 60(2): 434-448.

Barber, B.M., and Odean, T. (2001). "Boys Will Be Boys: Gender, Overconfidence and Common Stock Investment." *Quarterly Journal of Economics*, 116(1), 261-292.

Beyer, S. (1999). "Gender Differences in the Accuracy of Grade Expectancies and Evaluations." *Sex Roles*, 41:314, 279-296.

Coffman, K. B., and Klinowski, D. (2018). "The Impact of Penalties for Wrong Answers on the Gender Gap in Test Scores." HBS Working Paper 19-017.

Croson R. and Gneezy U. (2009). "Gender Differences in Preferences." *Journal of Economic Literature*, 47(2): 448-474.

Eckel C. and Grossman P. (2008). "Men, Women and Risk Aversion: Experimental Evidence." *Handbook of Experimental Economics Results*.

Espinosa M.P. and Gardeazabal J. (2013). "Do Students Behave Rationally in Multiple Choice Tests? Evidence from a Field Experiment." *Journal of Economics and Management*, Vol. 9, No. 2, 107-135.

Filippin, A., and Crosetto P. (2016) "A Reconsideration of Gender Differences in Risk Attitudes." *Management Science* 62(11): 3138-3160.

Funk, P., and Perrone, H. (2016). "Gender Differences in Academic Performance: The Role of Negative Marking in Multiple-Choice Exams." Working Paper.

Iriberri, N. and Rey-Biel, P. (Forthcoming). "Competitive Pressure Widens the Gender Gap in Performance: Evidence from a Two-Stage Competition in Mathematics." Forthcoming in *The Economic Journal*.

Niederle, M. and Vesterlund, L., "Gender and Competition", *Annual Review in Economics*, 2011, 3, 601–30.

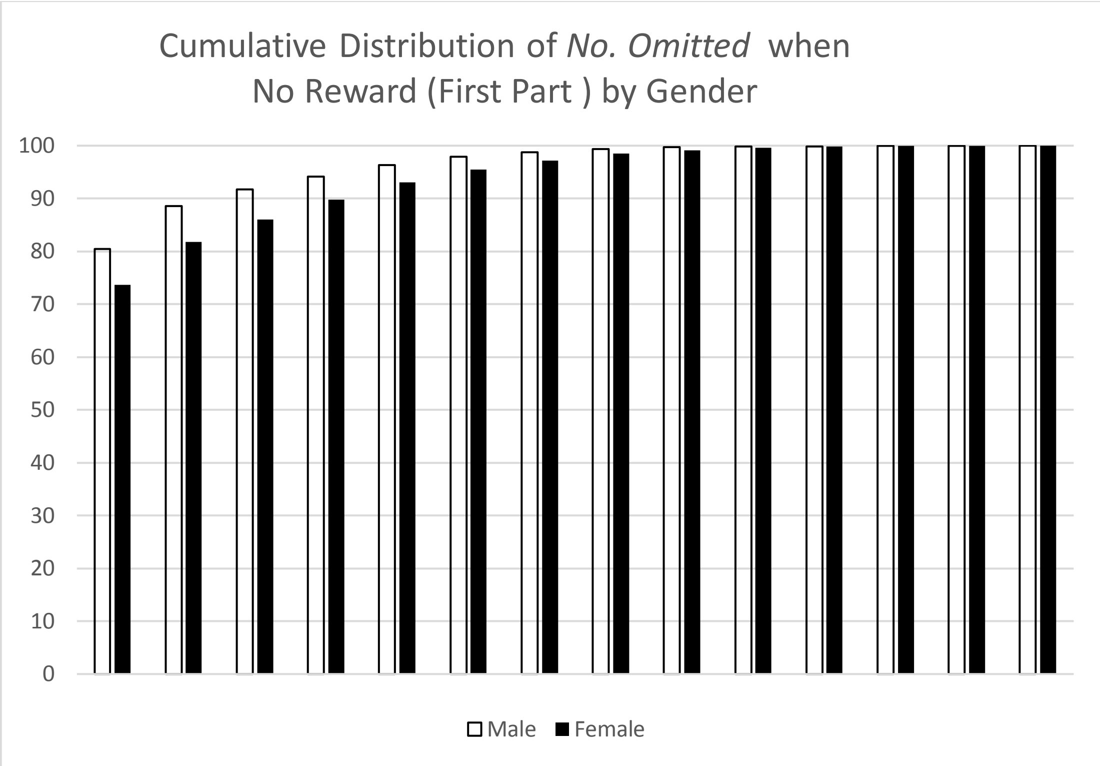
Ramos, I. and Lambating, J. (1996). "Gender Difference in Risk-Taking Behavior and their Relationship to SAT-Mathematics Performance." *School Science and Mathematics*, 96(4): 202-207.

Swineford, F. (1941). “Analysis of a Personality Trait.” *Journal of Educational Psychology*, 32(6):438–444.

Tannenbaum D. (2012). “Do Gender Differences in Risk Aversion Explain the Gender Gap in SAT Scores? Uncovering Risk Attitudes and the Test Score Gap". Unpublished paper, University of Chicago, Chicago.

Figures and Tables

Figure 1. ***No. Omitted*** **when No Reward and when Reward by Gender**



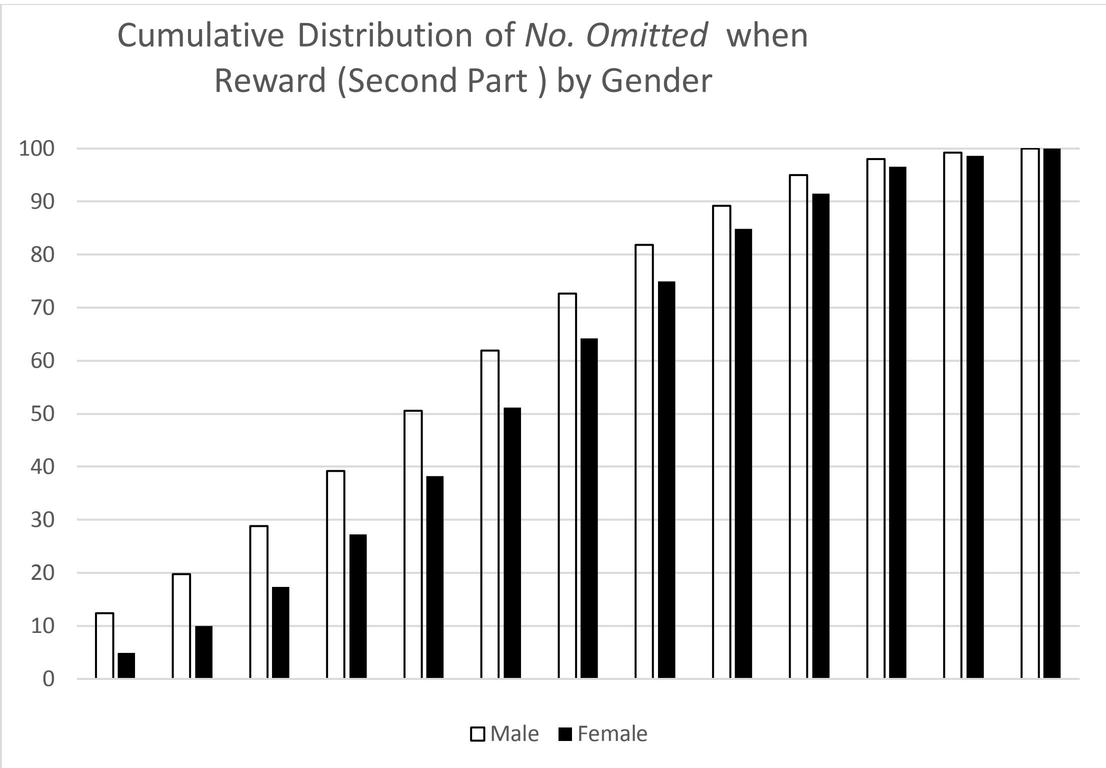
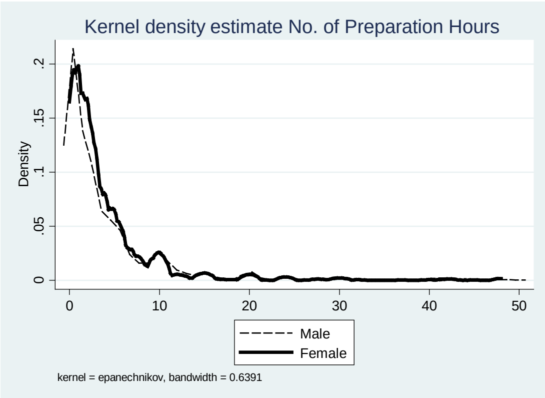
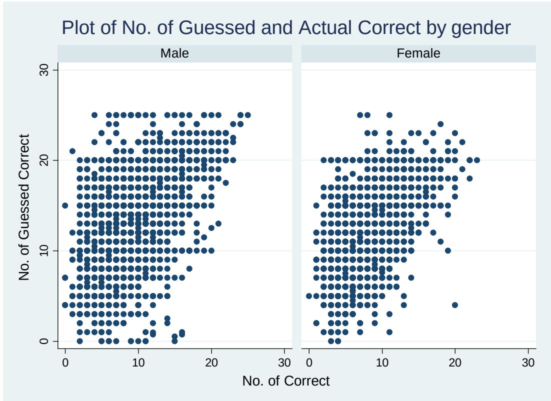


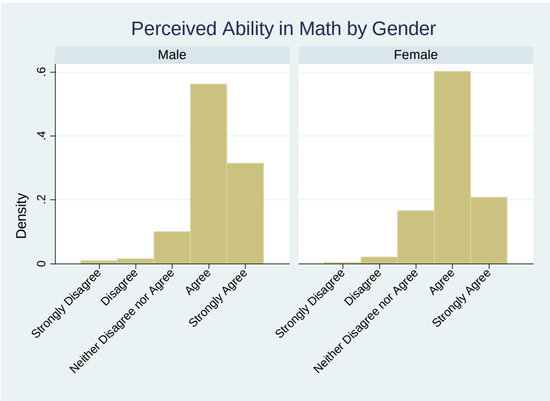
Figure 2. Descriptive Statistics on the Control Variables from the Questionnaire



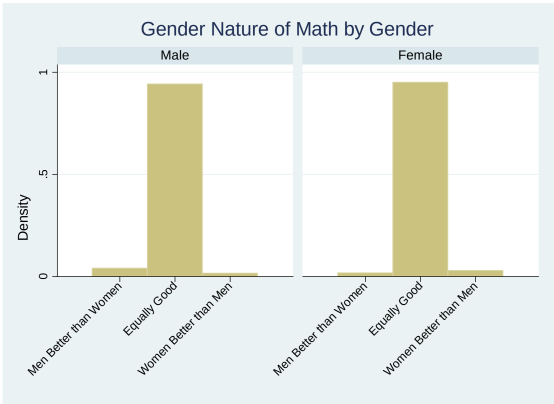
2a



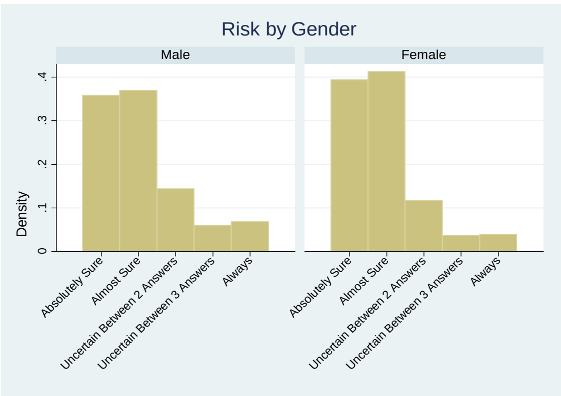
2b



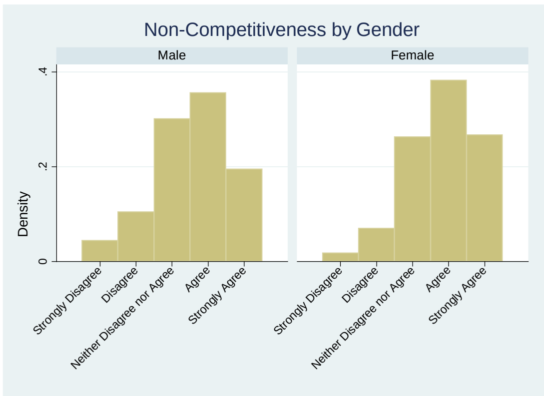
2c



2d



2e



2f

Figure 3. Variation in *No. of Correct No Reward Part* of the Test and in *Math at School*

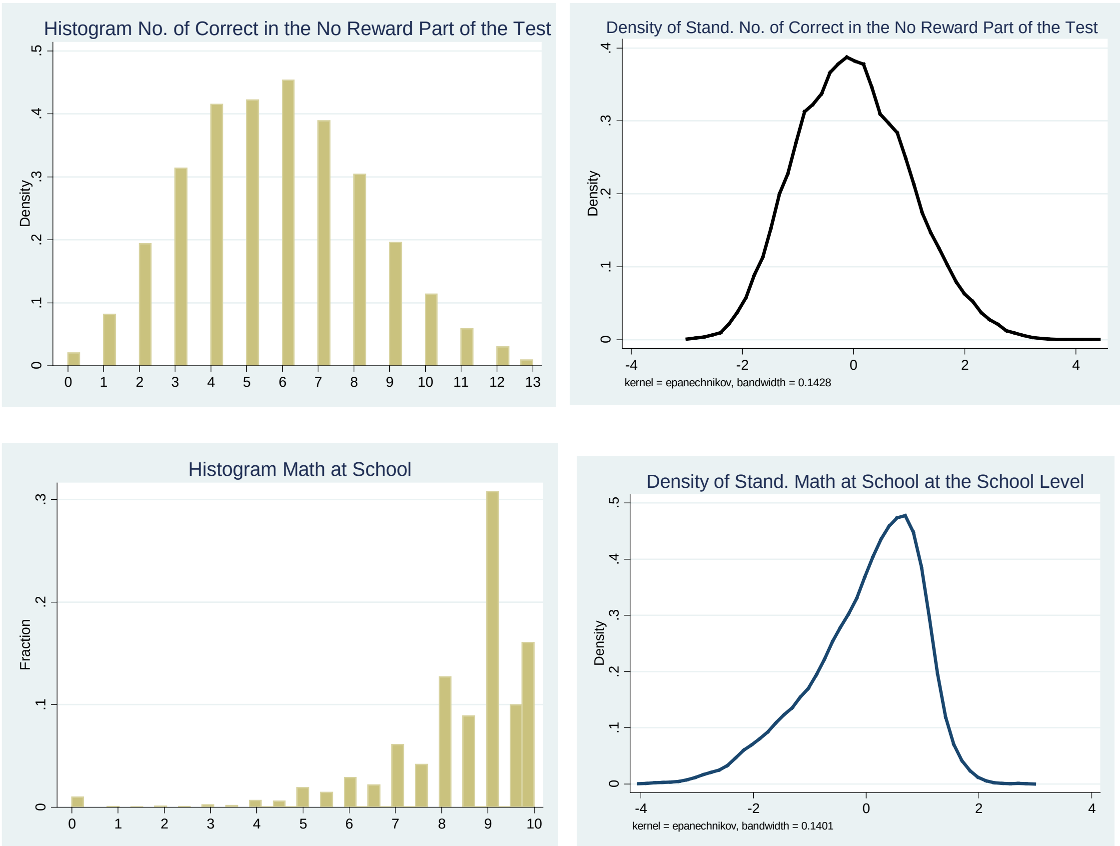
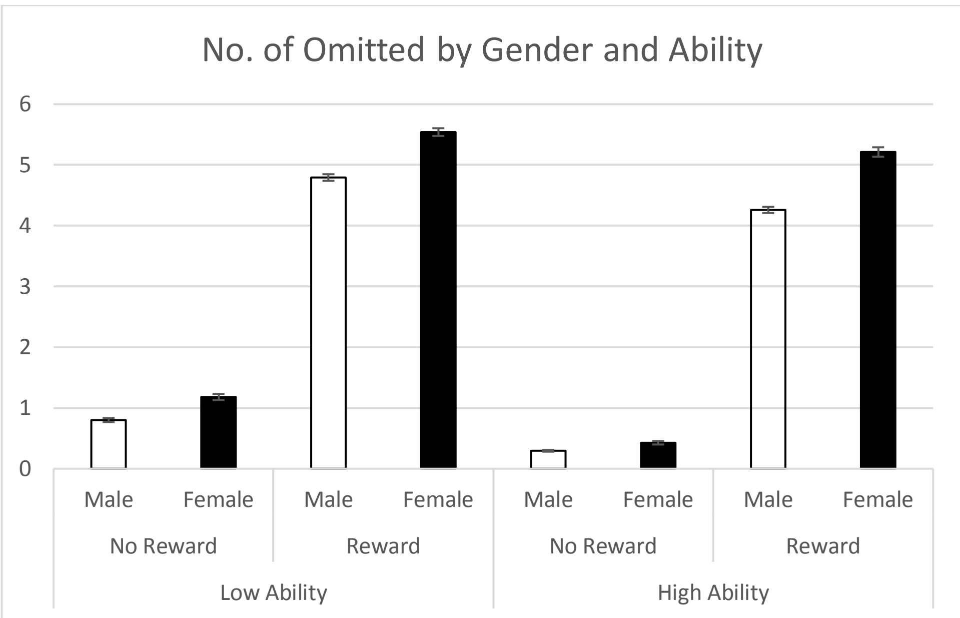
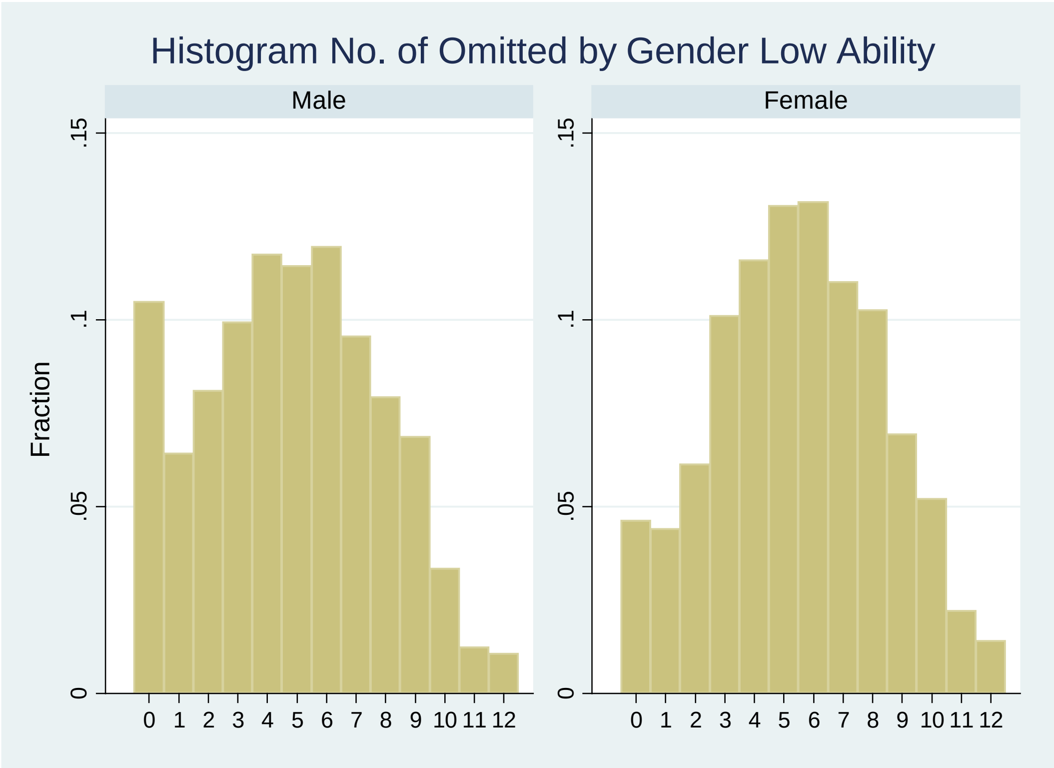


Figure 4. No. of Omitted by Gender and Ability: Low Ability: No. of Correct in No Reward<6 and High Ability: No. of Correct in No Reward>6

4a. No. of Omitted by Gender, Scoring Rule and by Ability



4b. No. of Omitted when Reward for Omitted by Gender for Low Ability



4b. No. of Omitted when Reward for Omitted by Gender for High Ability

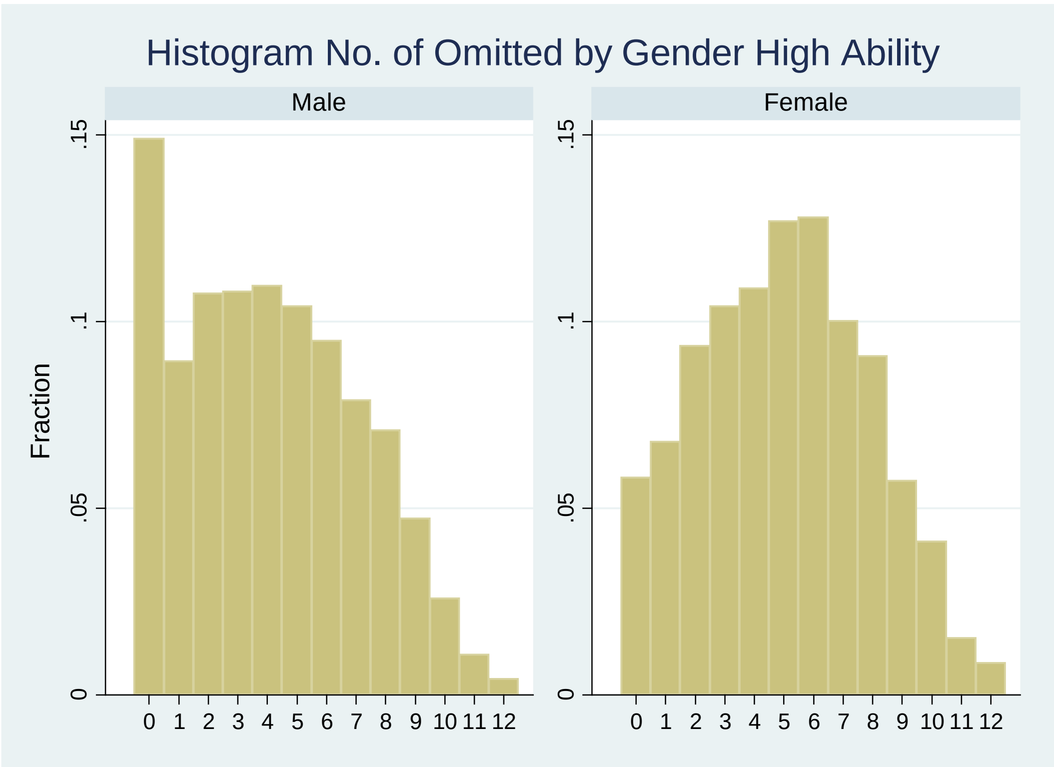


Table 1. Descriptive Statistics

a) Variables from the Math Test (2016-2017-2018)

	Overall					Men (6520)					Female (3386)					
No Reward for Omitted (First part)	Obs.	Mean	St. Dev	Min	Max	Obs.	Mean	St. Dev	Min	Max	Obs.	Mean	St. Dev	Min	Max	*p* -value
No. Omitted	9906	0.65	1.60	0	13	6520	0.53	1.43	0	13	3386	0.86	1.87	0	13	0.00
No. Correct	9906	5.69	2.49	0	13	6520	5.90	2.51	0	13	3386	5.29	2.42	0	13	0.00
Prop. Correct	9906	0.46	0.20	0	1	6520	0.48	0.20	0	1	3386	0.44	0.20	0	1	0.00
Score	9906	28.46	12.47	0	65	6520	29.50	12.53	0	65	3386	26.47	12.12	0	65	0.00
Rank	9906	387.531	273.43	1	1071	6520	405.93	274.13	1	1071	3386	352.11	268.59	1	1066	0.00
Reward for Omitted (Second part)	Obs.	Mean	St. Dev	Min	Max	Obs.	Mean	St. Dev	Min	Max	Obs.	Mean	St. Dev	Min	Max	*p* -value
No. Omitted	9906	4.82	2.99	0	12	6520	4.51	3.00	0	12	3386	5.40	2.87	0	12	0.00
No. Correct	9906	3.52	2.39	0	12	6520	3.76	2.44	0	12	3386	3.05	2.23	0	12	0.00
Prop. Correct	9906	0.48	0.27	0	1	6520	0.50	0.26	0	1	3386	0.46	0.27	0	1	0.00
Score	9906	22.40	10.43	0	60	6520	23.30	10.70	0	60	3386	20.67	9.66	1	60	0.00
Rank	9906	424.21	275.87	1	1072	6520	446.11	278.49	1	1072	3386	382.04	265.74	1	1067	0.00
Control Variables	Obs.	Mean	St. Dev	Min	Max	Obs.	Mean	St. Dev	Min	Max	Obs.	Mean	St. Dev	Min	Max	*p* -value
Math at School	8975	8.40	1.59	0	10	5899	8.32	1.64	0	10	3076	8.55	1.45	0	10	0.00
2016	9906	0.32				6520	0.33				3386	0.30				0.02
2017	9906	0.34				6520	0.34				3386	0.36				
2018	9906	0.34				6520	0.33				3386	0.34				
Level 1	9906	0.24				6520	0.23				3386	0.25				0.00
Level 2	9906	0.32				6520	0.32				3386	0.32				
Level 3	9906	0.28				6520	0.28				3386	0.30				
Level 4	9906	0.16				6520	0.17				3386	0.13				
Participation Time 1	9906	0.86				6520	0.85				3386	0.88				0.00
Participation Time 2	9906	0.12				6520	0.13				3386	0.10				
Participation Time 3	9906	0.02				6520	0.02				3386	0.02				

b) Variables from the Questionnaire (2017-2018)

	Overall					Men (6520)					Female (3386)					
	Obs.	Mean	St. Dev	Min	Max	Obs.	Mean	St. Dev	Min	Max	Obs.	Mean	St. Dev	Min	Max	*p* -value
No. of Preparation Hours	5924	4.36	8.65	0	100	3896	4.40	8.77	0	100	2028	4.28	8.42	0	100	0.63
Non-Competitiveness	6389	3.64	1.03	1	5	4132	3.55	1.05	1	5	2257	3.81	0.97	1	5	0.00
Overconfidence	4799	3.75	4.39	-16	21	3111	3.83	4.48	-15	21	1688	3.61	4.20	-16	18	0.11
Risk	5300	2.04	1.11	1	5	3399	2.11	1.16	1	5	1901	1.91	1.00	1	5	0.00
Perceived Math Ability	6104	4.10	0.72	1	5	3940	4.16	0.73	1	5	2164	3.99	0.70	1	5	0.00
Perceived Gender Nature of Math	6117	1.99	0.2355	1	3	3944	1.98	0.2413	1	3	2173	2.01	0.22	1	3	0.00

Notes : For all variables the table shows the number of observations, the mean, the standard deviation, the min and max values. The last column shows the *p* -value of the *F* -Test of equality of variable means across gender for the continuous variables and Fisher Exact test for categorical values. *No. of Omitted* , *No. of Correct* and *Prop. of Correct* measures the number of omitted, correct and proportion of correct by edition, level and test's parts level, respectively. *Score* measures the score in the Math test by edition, level and test's parts level. *Rank* measures the position in the rank by edition, level and test's parts level, where higher values represent better positions within the rank. *Math at School* measures the Math grade at school. *2016, 2017,* and *2018* take the value of 1 if the edition refers to 2016, 2017, and 2018, respectively. *Level 1, Level 2, Level 3,* and *Level 4* take the value of 1 if the level of the Math test refers to level 1, level 2, level 3 and level 4, respectively. *Participation Time 1, 2, 3* take the value of 1 if the it is the first, second or third time the same student takes part in the Math test, respectively. *No. of Preparation Hours* measures the total number of hours devoted to prepare the Math test. *Non-Competitiveness* contains the responses to question 1 in the questionnaire. *Perceived Math Ability* contains the responses to question 9 in the questionnaire. *Overconfidence* measures the difference between the guessed number of correct answers and the actual number of correct answers. And *Risk* contains the responses to question 8 in the questionnaire.

Table 2. Gender Differences between the No Reward and the Reward Parts of the Test

	OLS				RE				FE			
	zomitted	zprop_correct	zscore	rank	zomitted	zprop_correct	zscore	rank	zomitted	zprop_correct	zscore	rank
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Female	0.172***	-0.167***	-0.217***	-52.02***	0.173***	-0.170***	-0.217***	-52.71***				
	(0.0249)	(0.0211)	(0.0205)	(5.447)	(0.0252)	(0.0211)	(0.0204)	(5.441)				
Reward	-0.0458***	-0.00800	0.0126	39.33***	-0.0458***	-0.00800	0.0126	39.33***	-0.0458***	-0.00800	0.0126	39.33***
	(0.0162)	(0.0140)	(0.0131)	(3.470)	(0.0162)	(0.0140)	(0.0131)	(3.470)	(0.0160)	(0.0138)	(0.0129)	(3.416)
Female*Reward	0.145***	0.0194	-0.0439**	-10.62*	0.145***	0.0194	-0.0439**	-10.62*	0.145***	0.0194	-0.0439**	-10.62*
	(0.0295)	(0.0245)	(0.0218)	(5.942)	(0.0295)	(0.0245)	(0.0218)	(5.942)	(0.0291)	(0.0242)	(0.0215)	(5.850)
Math at School	0.0378***	0.235***	0.227***	55.36***	0.0365***	0.221***	0.210***	51.54***	-0.0433**	0.0627**	0.0721***	15.15**
	(0.00851)	(0.00810)	(0.00826)	(2.107)	(0.00840)	(0.00806)	(0.00815)	(2.113)	(0.0205)	(0.0245)	(0.0244)	(6.701)
Particiation Time	-0.162***	0.302***	0.381***	97.81***	-0.132***	0.211***	0.262***	67.64***	-0.128*	-0.0117	0.0465	1.254
	(0.0199)	(0.0205)	(0.0220)	(5.431)	(0.0186)	(0.0201)	(0.0208)	(5.304)	(0.0688)	(0.0760)	(0.0749)	(20.86)
Observations	17,822	17,822	17,822	17,822	17,822	17,822	17,822	17,822	17,822	17,822	17,822	17,822
R-squared	0.096	0.235	0.287	0.334					0.015	0.026	0.040	0.101
Number of participants					7,774	7,774	7,774	7,774	7,774	7,774	7,774	7,774

Notes: Observations are at the Math test's parts level. The first three outcome variables, *No. Omitted, Prop. of Correct,* and *Score* are standardized at the edition, level and part of the test levels. *Rank* measures the position in the rank by edition, level and test's parts level, where higher values represent better positions within the rank. *Female* takes the value of 1 if the participant is female. *Reward* takes the value of 1 if the outcome variable refers to the part of the test with reward for omitted questions. *Math at School* measures the standardized Math grade at school level and *Participation Time* takes the values of 1, 2, 3 if it is the first, second or third time that the participant does the Math test. Columns 1-4 show the OLS specification where the standard errors are clustered at the participant level. Columns 5-8 show the RE model specification and columns 9-12 show the FE specfication model. All specifications include edition, level and school fixed effects. Standard errors in parenthesis, where *** p<0.01, ** p<0.05, * p<0.1

Table 3. Gender Differences between the No Reward and the Reward Parts of the Test: Variation along the Ability Distribution

	Low Ability zomitted (1)	High Ability zomitted (2)	Interaction zomitted (3)	Continuous zomitted (4)
Female	0.191***	0.0847***	0.201***	0.404***
	(0.0387)	(0.0225)	(0.0380)	(0.0737)
Reward	-0.0928***	-0.00151	-0.0928***	-0.105**
	(0.0248)	(0.0194)	(0.0244)	(0.0447)
Female*Reward	0.0505	0.264***	0.0505	-0.228***
	(0.0425)	(0.0358)	(0.0418)	(0.0789)
High Ability			-0.0169	
			(0.0297)	
High Ability*Reward			0.0913***	
			(0.0307)	
Female*High Ability			-0.122***	
			(0.0439)	
Female*Reward*High Ability			0.213***	
			(0.0541)	
No. Of Correct No Reward	-0.0884***	-0.0603***	-0.0742***	-0.0711***
	(0.0117)	(0.00539)	(0.00596)	(0.00560)
Particiation Time	-0.0339	-0.0892***	-0.0738***	-0.0745***
	(0.0393)	(0.0198)	(0.0184)	(0.0184)
No. Of Correct No Reward*Reward				0.0102
				(0.00657)
Female*No. Of Correct No Reward				-0.0476***
				(0.0106)
Female*Reward*No. Of Correct No Reward				0.0691***
				(0.0122)
Observations	10,048	9,718	19,766	19,766
R-squared	0.123	0.153	0.114	0.115

Notes: Observations are at the Math test's parts level. *No. Omitted* is standardized at the edition, level and part of the test levels. *Female* takes the value of 1 if the participant is female. *Reward* takes the value of 1 if the outcome variable refers to the part of the test with reward for omitted questions. *No. of Correct No Reward* measures the number of correct questions in the part of the test without any reward for omitted questions, and *Participation Time* takes the values of 1, 2, 3 if it is the first, second or third time that the participant does the Math test. *High Ability* takes value 1 if the participant's standardized number of correct answers in the no reward part is>0. All columns show the OLS specification where the standard errors are clustered at the participant level and include edition, level and school fixed effects. Standard errors in parenthesis, where *** p<0.01, ** p<0.05, * p<0.1

Table 4. Gender Differences between the No Reward and the Reward Parts of the Test: Variation along Age

	Level 1 zomitted (1)	Level 2 zomitted (2)	Level 3 zomitted (3)	Level 4 zomitted (4)	Overall zomitted (5)
Female	0.214***	0.199***	0.0836*	0.177**	0.191***
	(0.0502)	(0.0453)	(0.0459)	(0.0700)	(0.0472)
Reward	-0.000281	-0.0623**	-0.0561*	-0.0608	-0.0458***
	(0.0321)	(0.0289)	(0.0312)	(0.0424)	(0.0162)
Female*Reward	0.0272	0.163***	0.181***	0.233***	0.0727
	(0.0588)	(0.0522)	(0.0549)	(0.0861)	(0.0509)
Math at School	-0.0124	0.00494	0.0959***	0.0278	0.0379***
	(0.0200)	(0.0168)	(0.0145)	(0.0204)	(0.00852)
Participation Time	-0.281***	-0.177***	-0.175***	-0.0776*	-0.162***
	(0.0574)	(0.0388)	(0.0346)	(0.0456)	(0.0199)
Level 2					0.0469
					(0.0320)
Level 3					0.0836**
					(0.0336)
Level 4					0.140***
					(0.0383)
Female*Level 2					0.0179
					(0.0622)
Female*Level 3					-0.0631
					(0.0629)
Female*Level 4					-0.0481
					(0.0822)
Level 2*Female*Reward					0.0741
					(0.0640)
Level 3*Female*Reward					0.0984
					(0.0651)
Level 4*Female*Reward					0.146*
					(0.0864)
Observations	4,248	5,748	5,040	2,786	17,822
R-squared	0.167	0.151	0.175	0.211	0.096

Notes: Observations are at the Math test's parts level. *No. Omitted* is standardized at the edition, level and part of the test levels. *Female* takes the value of 1 if the participant is female. *Reward* takes the value of 1 if the outcome variable refers to the part of the test with reward for omitted questions. *Math at School* measures the standardized Math grade at school level, and *Participation Time* takes the values of 1, 2, 3 if it is the first, second or third time that the participant does the Math test. All columns show the OLS specification where the standard errors are clustered at the participant level and include edition, level and school fixed effects. Standard errors in parenthesis, where *** p<0.01, ** p<0.05, * p<0.1

Table 5. Gender Differences between No Reward and the Reward Parts of the Test: Confidence, Overconfidence or Risk?

	Original Sample	Sample with Questionnaire	Sample with Questionnaire				
	zomitted	zomitted	zomitted	zomitted	zomitted	zomitted	zomitted
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	0.172***	0.173***	0.115***	0.117***	0.115***	0.117***	0.136***
	(0.0249)	(0.0348)	(0.0348)	(0.0350)	(0.0348)	(0.0347)	(0.0346)
Reward	-0.0458***	-0.0255	-0.0255	-0.100	0.000439	0.0291	0.427***
	(0.0162)	(0.0240)	(0.0240)	(0.0756)	(0.119)	(0.0304)	(0.0455)
Female*Reward	0.145***	0.101**	0.101**	0.0964**	0.100**	0.0981**	0.0590
	(0.0295)	(0.0430)	(0.0430)	(0.0435)	(0.0432)	(0.0429)	(0.0426)
Math at School	0.0378***	0.0380***	0.0293**	0.0293**	0.0293**	0.0293**	0.0293**
	(0.00851)	(0.0127)	(0.0125)	(0.0125)	(0.0125)	(0.0125)	(0.0125)
Particiation Time	-0.162***	-0.169***	-0.142***	-0.142***	-0.142***	-0.142***	-0.142***
	(0.0199)	(0.0263)	(0.0243)	(0.0243)	(0.0243)	(0.0243)	(0.0243)
Non-Competitiveness			0.0407***	0.0302*	0.0407***	0.0407***	0.0407***
			(0.0117)	(0.0159)	(0.0117)	(0.0117)	(0.0117)
Perceived Math Ability			-0.0640***	-0.0640***	-0.0609***	-0.0640***	-0.0640***
			(0.0167)	(0.0167)	(0.0214)	(0.0167)	(0.0167)
Overconfidence			-0.0137***	-0.0137***	-0.0137***	-0.00670**	-0.0137***
			(0.00265)	(0.00265)	(0.00265)	(0.00319)	(0.00265)
Risk			-0.202***	-0.202***	-0.202***	-0.202***	-0.0951***
			(0.00993)	(0.00993)	(0.00993)	(0.00993)	(0.0117)
Non-Competitiveness*Reward				0.0209			
				(0.0203)			
Perceived Math Ability*Reward					-0.00620		
					(0.0277)		
Overconfidence*Reward						-0.0141***	
						(0.00434)	
Risk*reward							-0.215***
							(0.0171)
Observations	17,822	8,302	8,302	8,302	8,302	8,302	8,302
R-squared	0.096	0.136	0.186	0.186	0.186	0.186	0.200

Notes: Observations are at the Math test's parts level. Column 1 shows the main estimation result from column 1 in Table 2 for the original sample. For the rest of the columns, observations are at the Math test's parts level in the edition of 2017 and 2018 for the participants whose questionnaire answers are available. *No. Omitted* is standardized at the edition, level and part of the test levels. *Female* takes the value of 1 if the participant is female. *Reward* takes the value of 1 if the outcome variable refers to the part of the test with reward for omitted questions. *Math at School* measures the standardized Math grade at school level and *Participation Time* takes the values of 1, 2, 3 if it is the first, second or third time that the participant does the Math test. *Non-Competitiveness* contains the responses to question 1 in the questionnaire. *Perceived Math Ability* contains the responses to question 9 in the questionnaire. *Overconfidence* measures the difference between the guessed number of correct answers and the actual number of correct answers. And *Risk* contains the responses to question 8 in the questionnaire. All columns show the OLS specification where the standard errors are clustered at the participant level and include edition, level and school fixed effects. Standard errors in parenthesis, where *** p<0.01, ** p<0.05, * p<0.1

Table A6. Gender Differences between the No Reward and the Reward Parts of the Test: Confidence, Overconfidence or Risk? With Alternative Control for Ability: No. Of Corret No Reward

	Original Sample	**Sample with Questionnaire**	**Sample with Questionnaire**				
	zomitted	zomitted	zomitted	zomitted	zomitted	zomitted	zomitted
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	0.153***	0.181***	0.112***	0.115***	0.112***	0.114***	0.134***
	(0.0234)	(0.0327)	(0.0322)	(0.0323)	(0.0322)	(0.0322)	(0.0321)
Reward	-0.0447***	-0.0184	-0.0184	-0.110	-0.0489	0.0300	0.439***
	(0.0154)	(0.0226)	(0.0226)	(0.0708)	(0.114)	(0.0286)	(0.0427)
Female*Reward	0.132***	0.0652	0.0652	0.0590	0.0664	0.0624	0.0219
	(0.0282)	(0.0409)	(0.0409)	(0.0413)	(0.0410)	(0.0408)	(0.0405)
No. Of Correct No Reward	-0.0704***	-0.0638***	-0.0925***	-0.0925***	-0.0925***	-0.0925***	-0.0925***
	(0.00401)	(0.00558)	(0.00609)	(0.00609)	(0.00609)	(0.00609)	(0.00609)
Particiation Time	-0.0738***	-0.0843***	-0.0449**	-0.0449**	-0.0449**	-0.0449**	-0.0449**
	(0.0184)	(0.0247)	(0.0223)	(0.0223)	(0.0223)	(0.0223)	(0.0223)
Non-Competitiveness			0.0164	0.00355	0.0164	0.0164	0.0164
			(0.0104)	(0.0142)	(0.0104)	(0.0104)	(0.0104)
Perceived Math Ability			-0.0230	-0.0230	-0.0267	-0.0230	-0.0230
			(0.0160)	(0.0160)	(0.0209)	(0.0160)	(0.0160)
Overconfidence			-0.0358***	-0.0358***	-0.0358***	-0.0295***	-0.0358***
			(0.00291)	(0.00291)	(0.00291)	(0.00340)	(0.00291)
Risk			-0.209***	-0.209***	-0.209***	-0.209***	-0.101***
			(0.00927)	(0.00927)	(0.00927)	(0.00927)	(0.0109)
Non-Competitiveness*Reward				0.0256			
				(0.0190)			
Perceived Math Ability*Reward					0.00733		
					(0.0265)		
Overconfidence*Reward						-0.0126***	
						(0.00411)	
Risk*reward							-0.217***
							(0.0161)
Observations	19,766	9,284	9,284	9,284	9,284	9,284	9,284
R-squared	0.112	0.146	0.209	0.209	0.209	0.210	0.224

Notes: Column 1 shows the main estimation result from column 1 in Table A2 for the original sample. For the rest of the columns, observations are at the Math test's parts level in the edition of 2017 and 2018 for the participants whose questionnaire answers are available. *Female* takes the value of 1 if the participant is female. *Reward* takes the value of 1 if the outcome variable refers to the part of the test with reward for omitted questions. *No. of Correct No Reward* measures the number of correct questions in the part of the test without any reward for omitted question and *Participation Time* takes the values of 1, 2, 3 if it is the first, second or third time that the participant does the Math test. *Non-Competitiveness* contains the responses to question 1 in the questionnaire. Perceived Math Ability contains the responses to question 9 in the questionnaire. *Perceived Math Ability* contains the responses to question 9 in the questionnaire. *Overconfidence* measures the difference between the guessed number of correct answers and the actual number of correct answers. And *Risk* contains the responses to question 8 in the questionnaire. All columns show the OLS specification where the standard errors are clustered at the participant level and include edition, level and school fixed effects. Standard errors in parenthesis, where *** p<0.01, ** p<0.05, * p<0.1

Table 7. Gender Differences between No Reward and the Reward Part of the Test along Age: Confidence, Overconfidence or Risk?

	Original Sample		Sample with Questionnaire		Sample with Questionnaire									
	Level 1 zomitted	Level 4 zomitted	Level 1 zomitted	Level 4 zomitted	Level 1 zomitted	Level 4 zomitted	Level 1 zomitted	Level 4 zomitted	Level 1 zomitted	Level 4 zomitted	Level 1 zomitted	Level 4 zomitted	Level 1 zomitted	Level 4 zomitted
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Female	0.214***	0.177**	0.150**	0.338***	0.131*	0.259**	0.123*	0.260**	0.135**	0.253**	0.131*	0.270**	0.140**	0.304***
	(0.0503)	(0.0700)	(0.0687)	(0.115)	(0.0687)	(0.111)	(0.0688)	(0.111)	(0.0687)	(0.111)	(0.0687)	(0.111)	(0.0686)	(0.110)
Reward	-0.000281	-0.0608	0.0293	-0.0941	0.0293	-0.0941	0.354**	-0.118	0.473*	-0.417	0.0328	-0.00438	0.229***	0.714***
	(0.0321)	(0.0424)	(0.0497)	(0.0662)	(0.0498)	(0.0663)	(0.178)	(0.193)	(0.262)	(0.348)	(0.0669)	(0.0860)	(0.0850)	(0.139)
Female*Reward	0.0272	0.233***	-0.0270	0.164	-0.0270	0.164	-0.0106	0.161	-0.0350	0.175	-0.0267	0.142	-0.0460	0.0727
	(0.0587)	(0.0861)	(0.0870)	(0.139)	(0.0871)	(0.140)	(0.0872)	(0.141)	(0.0871)	(0.139)	(0.0870)	(0.140)	(0.0868)	(0.136)
Math at School	-0.0133	0.0282	-0.0294	0.0286	-0.0181	-0.00549	-0.0181	-0.00549	-0.0181	-0.00549	-0.0181	-0.00549	-0.0181	-0.00549
	(0.0199)	(0.0203)	(0.0285)	(0.0379)	(0.0290)	(0.0340)	(0.0290)	(0.0340)	(0.0290)	(0.0340)	(0.0290)	(0.0340)	(0.0290)	(0.0340)
Particiation Time	-0.288***	-0.0819*	-0.197***	-0.0464	-0.181**	-0.00434	-0.181**	-0.00434	-0.181**	-0.00434	-0.181**	-0.00434	-0.181**	-0.00434
	(0.0563)	(0.0436)	(0.0752)	(0.0710)	(0.0726)	(0.0626)	(0.0726)	(0.0627)	(0.0726)	(0.0627)	(0.0726)	(0.0627)	(0.0726)	(0.0627)
Non-Competitiveness					-0.00796	0.0576	0.0348	0.0541	-0.00796	0.0576	-0.00796	0.0576	-0.00796	0.0576
					(0.0259)	(0.0369)	(0.0345)	(0.0496)	(0.0259)	(0.0369)	(0.0259)	(0.0369)	(0.0259)	(0.0369)
Perceived Math Ability					-0.137***	-0.0134	-0.137***	-0.0134	-0.0847*	-0.0532	-0.137***	-0.0134	-0.137***	-0.0134
					(0.0374)	(0.0487)	(0.0374)	(0.0487)	(0.0472)	(0.0672)	(0.0374)	(0.0487)	(0.0374)	(0.0487)
Overconfidence					-0.00497	-0.0262***	-0.00497	-0.0262***	-0.00497	-0.0262***	-0.00460	-0.0128	-0.00497	-0.0262***
					(0.00528)	(0.00809)	(0.00529)	(0.00810)	(0.00529)	(0.00810)	(0.00692)	(0.00953)	(0.00529)	(0.00810)
Risk					-0.119***	-0.268***	-0.119***	-0.268***	-0.119***	-0.268***	-0.119***	-0.268***	-0.0668**	-0.107***
					(0.0210)	(0.0285)	(0.0210)	(0.0285)	(0.0210)	(0.0285)	(0.0210)	(0.0285)	(0.0262)	(0.0353)
Non-Competitiveness*Reward							-0.0855*	0.00708						
							(0.0452)	(0.0555)						
Perceived Math Ability*Reward									-0.104*	0.0797				
									(0.0593)	(0.0845)				
Overconfidence*Reward											-0.000736	-0.0267**		
											(0.00839)	(0.0132)		
Risk*reward													-0.105***	-0.322***
													(0.0343)	(0.0435)
Observations	4,248	2,786	2,078	1,258	2,078	1,258	2,078	1,258	2,078	1,258	2,078	1,258	2,078	1,258
R-squared	0.167	0.211	0.213	0.304	0.232	0.363	0.234	0.363	0.234	0.364	0.232	0.366	0.235	0.400

Notes : Observations are at the Math test's parts level. Columns 1-2 show the main estimation results from columns 1-2 in Table 4 for the original sample. For the rest of the columns, observations are at the Math test's parts level in the edition of 2017 and 2018 for the participants whose questionnaire answers are available. *No. Omitted* is standardized at the edition, level and part of the test levels. *Female* takes the value of 1 if the participant is female. *Reward* takes the value of 1 if the outcome variable refers to the part of the test with reward for omitted questions. *Math at School* measures the Math grade at school, and *Participation Time* takes the values of 1, 2, 3 if it is the first, second or third time that the participant does the Math test. *Non-Competitiveness* contains the responses to question 1 in the questionnaire. Perceived Math Ability contains the responses to question 9 in the questionnaire. *Perceived Math Ability* contains the responses to question 9 in the questionnaire. *Overconfidence* measures the difference between the guessed number of correct answers and the actual number of correct answers. And *Risk* contains the responses to question 8 in the questionnaire. All columns show the OLS specification where the standard errors are clustered at the participant level and include edition, level and school fixed effects. Standard errors in parenthesis, where *** p<0.01, ** p<0.05, * p<0.1

Figures and Tables in the Appendix

Figure A1. Description of Grading System in the Math Test

concurso de primavera MATEMÁTICAS

XX CONCURSO DE PRIMAVERA DE MATEMÁTICAS

2ª FASE: 23 de abril de 2016

NIVEL I (5º y 6º Primaria)

¡¡¡ Lee detenidamente estas instrucciones!!!

Escribe tu nombre y los datos que se te piden en la hoja de respuestas. No pases la página hasta que se te indique.

La prueba tiene una duración de 1 HORA 30 MINUTOS.

No está permitido el uso de calculadoras, reglas graduadas, ni ningún otro instrumento de medida.

Es difícil contestar bien a todas las preguntas en el tiempo indicado. Concéntrate en las que veas más asequibles. Cuando hayas contestado a esas, inténtalo con las restantes.

PUNTUACIÓN

En los problemas 1 a 13:

Cada respuesta correcta te aportará	5 puntos
Cada pregunta en blanco o errónea	0 puntos

En los problemas 14 a 25:

Cada respuesta correcta te aportará	5 puntos
Cada pregunta que dejes en blanco	1 punto
Cada respuesta errónea	0 puntos

EN LA HOJA DE RESPUESTAS, MARCA CON UNA ASPA ☒ LA QUE CONSIDERES CORRECTA.

SI TE EQUIVOCAS, ESCRIBE "NO" EN LA EQUIVOCADA Y MARCA LA QUE CREAS CORRECTA.

CONVOCA
Facultad de Matemáticas de la UCM

ORGANIZA
Asociación Matemática
Concurso de Primavera

COLABORAN
Universidad Complutense de Madrid
Consejería de Educación de la Comunidad de Madrid
El Corte Inglés
Grupo ANAYA
Grupo SM
Smartick

Figure A2. Mean Values of Correct Per Question: First Part (Questions 1-13) and Second Part (Questions 14-25)

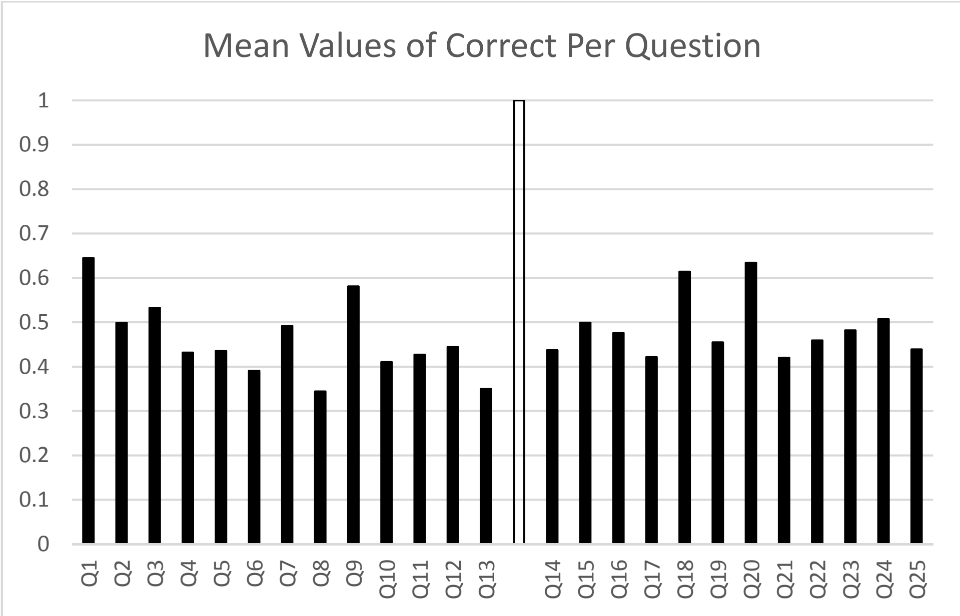


Figure A3. Questionnaire at the end of the Test

For the following statements please, say your agreement level (1 referring to Strongly Disagree and 5 to Strongly Agree):

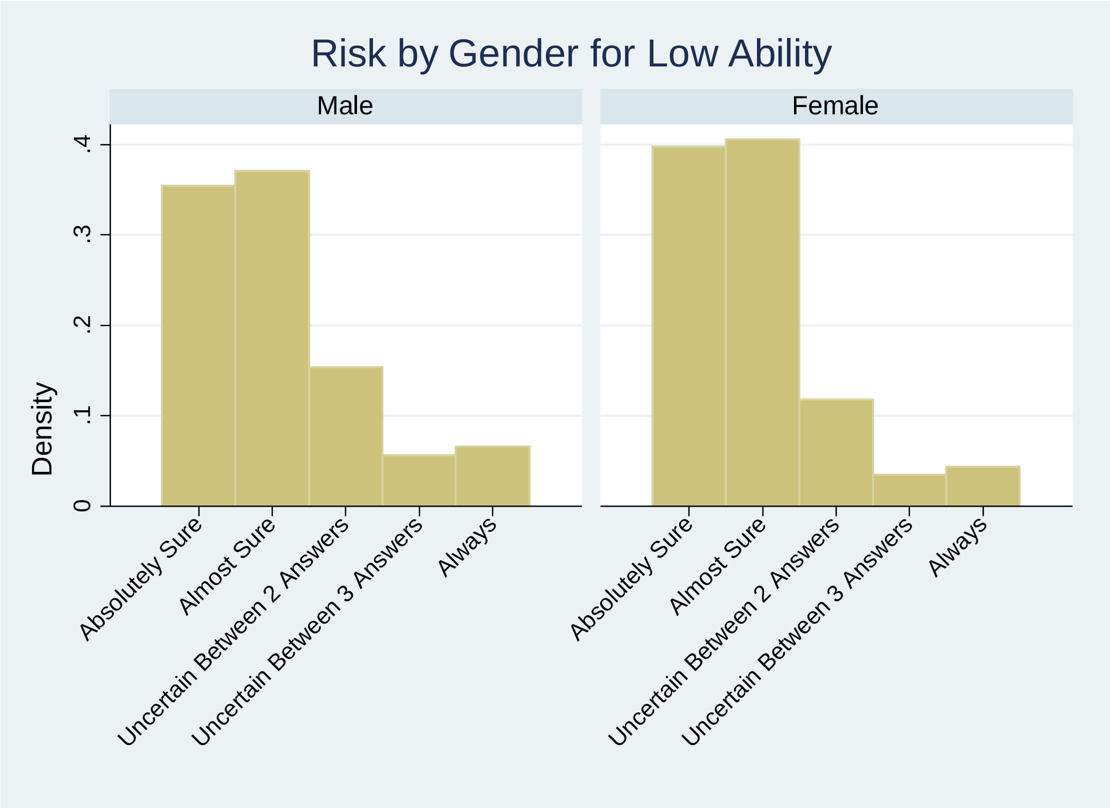
1. "It is more important for me to be selected for Stage 2 than being among the winners in Stage 2."
2. "It is more important to my parents being selected for Stage 2 than being among the winners in Stage 2."
3. "It is more important to do well in Stage 2 than in Stage 1."
4. "I have devoted more hours to prepare Stage 2 test than Stage 1 test."
6. "While doing the test I felt more pressure during Stage 2 than in Stage 1"

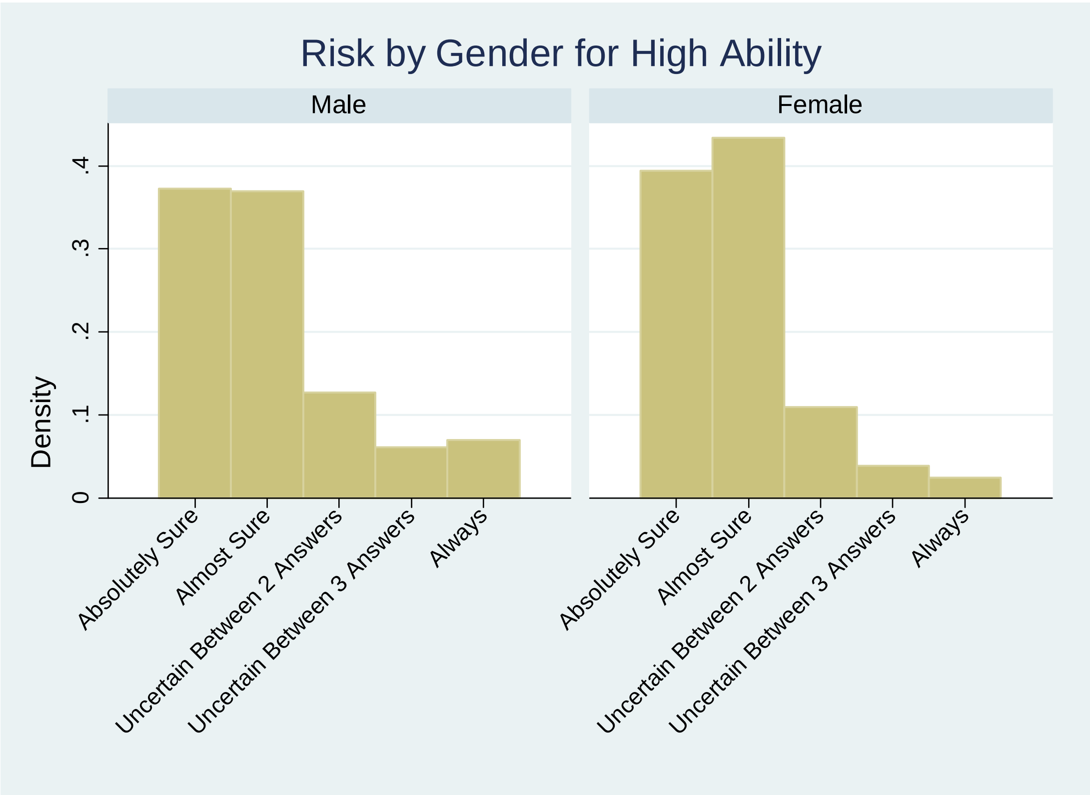
9. "I am good at Mathematics"

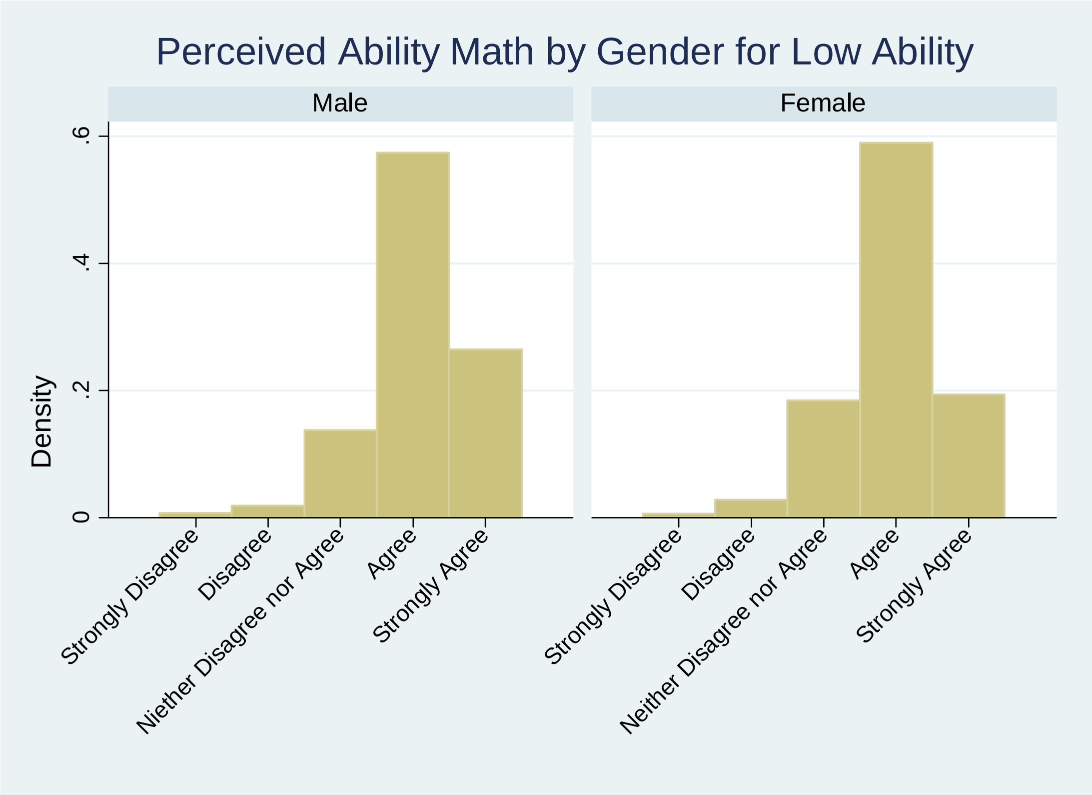
5. How many hours did you devote to prepare Stage 2 test?
7. How many questions do you expect to get right?
8. When omitting a question was worth 1 point I answered the question ______
   a. when I was absolutely sure.
   b. when I was almost sure.
   c. when I was uncertain between 2 answers.
   d. when I was uncertain between 3 answers.
   e. always.

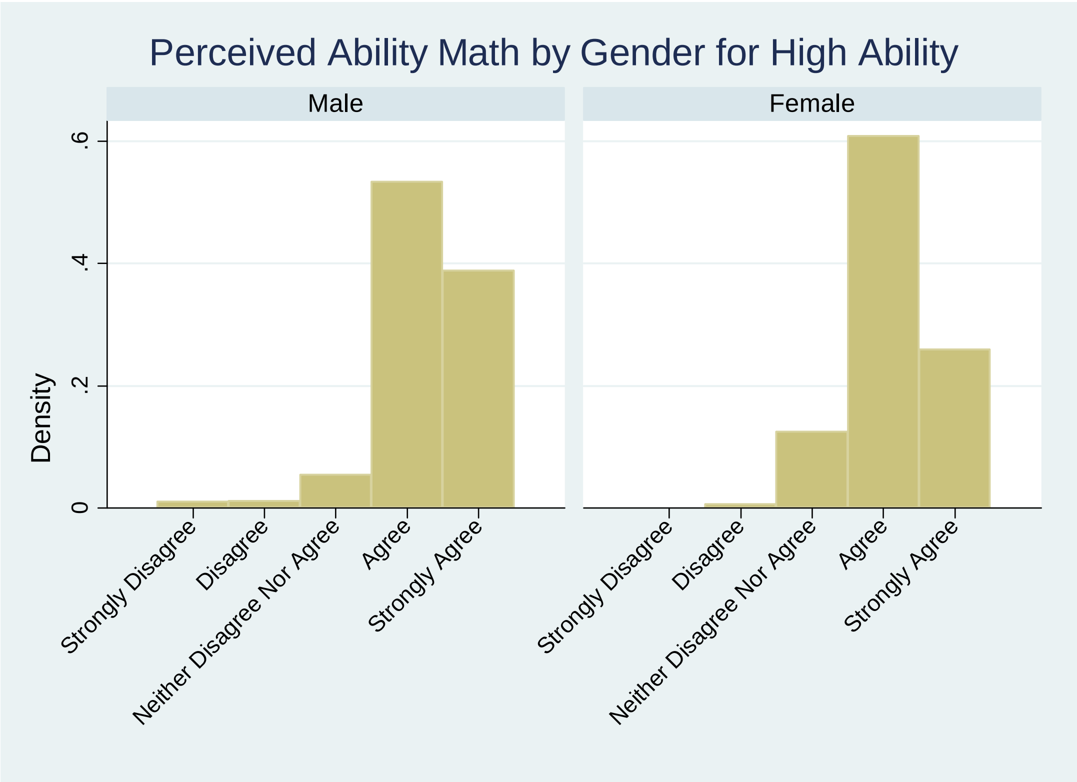
10. I believe _______ at Math
   a. men are better than women
   b. men and women are equally good
   c. women are better than men

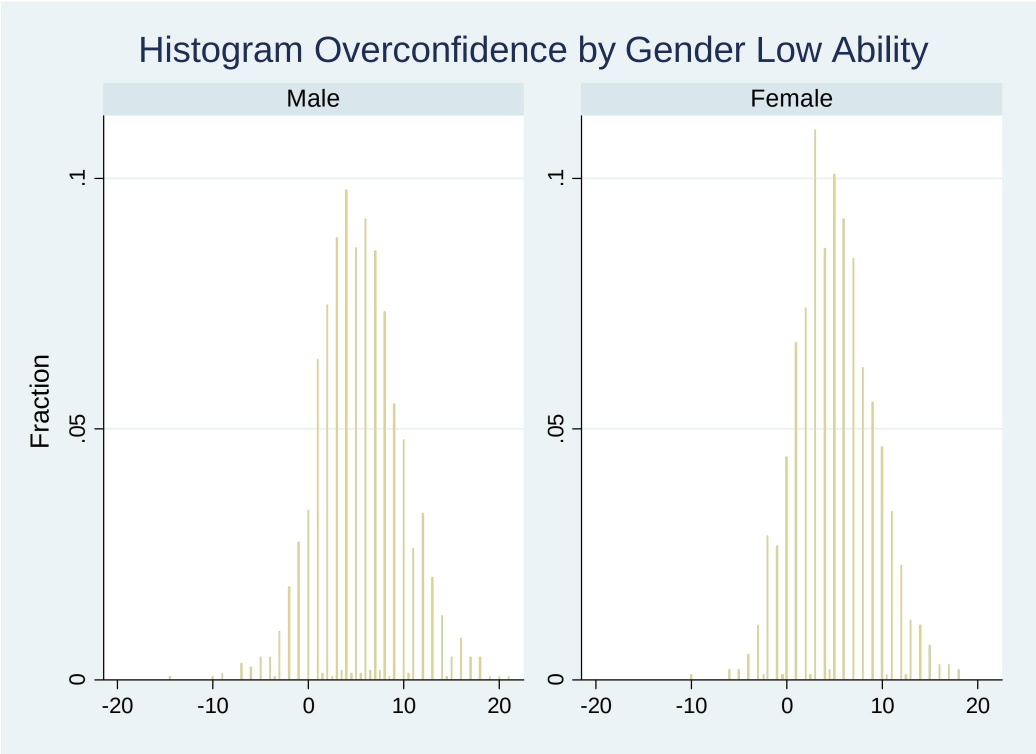
Figure A4. Risk, Confidence and Overconfidence by Gender: Low Ability: No. of Correct in No Reward<6 and High Ability: No. of Correct in No Reward>6



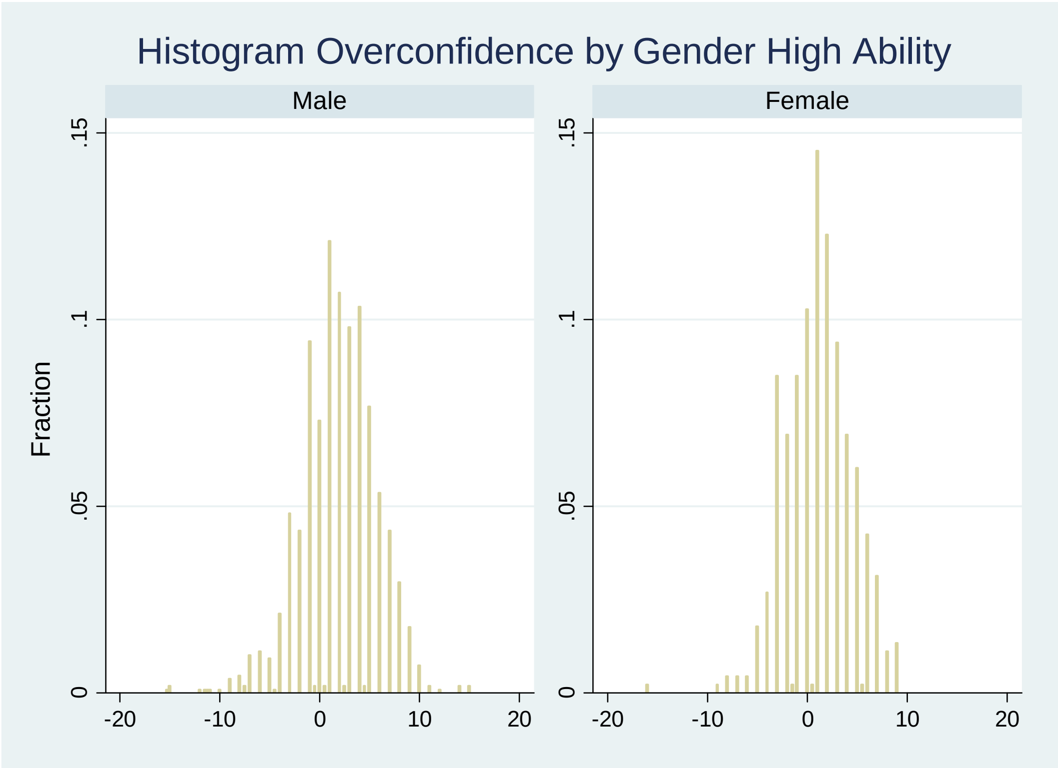



Perceived Ability Math by Gender for Low Ability
Male
Female
Density
0
.2
.4
.6
Strongly Disagree
Disagree
Niether Disagree nor Agree
Agree
Strongly Agree
Strongly Disagree
Disagree
Neither Disagree nor Agree
Agree
Strongly Agree


Perceived Ability Math by Gender for High Ability
Male
Female
Density
0
.2
.4
.6
Strongly Disagree
Disagree
Neither Disagree Nor Agree
Agree
Strongly Agree
Strongly Disagree
Disagree
Neither Disagree Nor Agree
Agree
Strongly Agree



Histogram Overconfidence by Gender Low Ability
Male
Female
Fraction
0
.05
.1
-20
-10
0
10
20



Histogram Overconfidence by Gender High Ability
Male
Female
Fraction
0
.05
.1
.15
-20
-10
0
10
20

Table A1. Gender Differences between No Reward and the Reward Parts of the Test with Alternative Control for Ability: No. Of Correct No Reward

	OLS				RE				FE			
	zomitted	zprop_correct	zscore	rank	zomitted	zprop_correct	zscore	rank	zomitted	zprop_correct	zscore	rank
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Female	0.153***	-0.0111	-0.0539***	-11.12***	0.153***	-0.0112	-0.0539***	-11.12***				
	(0.0234)	(0.0120)	(0.00808)	(2.609)	(0.0237)	(0.0120)	(0.00808)	(2.609)				
Reward	-0.0447***	-0.00383	0.0148	40.39***	-0.0447***	-0.00383	0.0148	40.39***	-0.0447***	-0.00383	0.0148	40.39***
	(0.0154)	(0.0134)	(0.0125)	(3.301)	(0.0154)	(0.0134)	(0.0125)	(3.301)	(0.0152)	(0.0132)	(0.0123)	(3.254)
Female*Reward	0.132***	0.0140	-0.0408**	-10.29*	0.132***	0.0140	-0.0408**	-10.29*	0.132***	0.0140	-0.0408**	-10.29*
	(0.0282)	(0.0233)	(0.0207)	(5.659)	(0.0282)	(0.0233)	(0.0207)	(5.659)	(0.0278)	(0.0230)	(0.0204)	(5.578)
No. Of Correct No Reward	-0.0704***	0.258***	0.284***	70.78***	-0.0679***	0.258***	0.284***	70.78***	-0.0318***	0.199***	0.212***	51.38***
	(0.00401)	(0.00245)	(0.00230)	(0.615)	(0.00400)	(0.00245)	(0.00230)	(0.615)	(0.0116)	(0.00727)	(0.00664)	(1.838)
Particiation Time	-0.0740***	0.0834***	0.134***	34.39***	-0.0757***	0.0833***	0.134***	34.39***	-0.219**	-0.0404	0.0211	-2.647
	(0.0184)	(0.0133)	(0.0128)	(3.512)	(0.0180)	(0.0133)	(0.0128)	(3.512)	(0.0872)	(0.0543)	(0.0486)	(13.95)
Observations	19,766	19,766	19,766	19,766	19,766	19,766	19,766	19,766	19,766	19,766	19,766	19,766
R-squared	0.112	0.474	0.591	0.583					0.016	0.070	0.100	0.150
Number of participants					8,537	8,537	8,537	8,537	8,537	8,537	8,537	8,537

Notes : Observations are at the Math test's parts level. The first three outcome variables, *No. Omitted, Prop. of Correct* and *Score* are standardized at the edition, level and part of the test levels. *Rank* measures the position in the rank by edition, level and test's parts level, where higher values represent better positions within the rank. *Female* takes the value of 1 if the participant is female. *Reward* takes the value of 1 if the outcome variable refers to the part of the test with reward for omitted questions. *No. of Correct No Reward* measures the number of correct in the part of the test with the reward and *Participation Time* takes the values of 1, 2, 3 if it is the first, second or third time that the participant does the Math test. Columns 1-4 show the OLS specification where the standard errors are clustered at the participant level. Columns 5-8 show the RE model specification and columns 9-12 show the FE specfication model. All specifications include edition, level and school fixed effects. Standard errors in parenthesis, where *** p<0.01, ** p<0.05, * p<0.1

Table A2. Gender Differences between the First and the Second Parts of the Test, Placebo test with 2013, 2014 and 2015 Editions

	Placebo 2013 Edition				Placebo 2014 Edition				Placebo 2015 Edition			
	zomitted	zprop_correct	zscore	rank	zomitted	zprop_correct	zscore	rank	zomitted	zprop_correct	zscore	rank
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Female	0.150***	-0.0258	-0.159***	-45.15***	0.267***	-0.164***	-0.278***	-38.53***	0.278***	-0.171***	-0.272***	-67.00***
	(0.0330)	(0.0330)	(0.0312)	(10.34)	(0.0500)	(0.0484)	(0.0443)	(7.756)	(0.0380)	(0.0372)	(0.0354)	(9.031)
Second Half	-0.000971	-0.0106	-0.0147	-8.088	-0.0151	-0.00447	-0.00825	-2.356	0.0257	-0.0272	-0.0249	-8.437
	(0.0124)	(0.0158)	(0.0175)	(5.527)	(0.0226)	(0.0279)	(0.0269)	(4.593)	(0.0184)	(0.0225)	(0.0221)	(5.523)
Female*Second Half	0.00280	0.0307	0.0424	13.28	0.0460	0.0137	0.0252	1.802	-0.0711**	0.0750*	0.0688*	18.63**
	(0.0219)	(0.0286)	(0.0297)	(9.551)	(0.0403)	(0.0502)	(0.0439)	(7.941)	(0.0314)	(0.0398)	(0.0360)	(9.332)
Observations	7,794	7,794	7,794	7,794	4,170	4,170	4,170	4,170	6,250	6,250	6,250	6,250
R-squared	0.219	0.224	0.259	0.346	0.256	0.262	0.317	0.374	0.220	0.234	0.279	0.354

Notes : Observations are at the Math test's parts level. The first three outcome variables, *No. Omitted* , *Prop. of Correct* and *Score* are standardized at the level and part of the test levels. Rank measures the position in the rank by level and test's parts level, where higher values represent better positions within the rank. *Female* takes the value of 1 if the participant is female. *Second Half* takes the value of 1 if the outcome variable refers to the questions 14-25. Columns 1-4 show the results for the edition 2013, columns 5-8 for the edition 2014 and columns 9-12 for the edition 2015. In editions 2013, 2014 and 2015 there was differential score for omitte questions and wrong answers for all questions in the test. All regressions include level and school fixed effects. Standard errors, clustered at the participant level, are shown in parenthesis, where *** p<0.01, ** p<0.05, * p<0.1

Table A3. Gender Differences between the No Reward and the Reward Parts of the Test for those Participants who Leave 0 Omitted Questions in the No Reward Part of the Test

	OLS				RE				FE			
	zomitted	zprop_correct	zscore	rank	zomitted	zprop_correct	zscore	rank	zomitted	zprop_correct	zscore	rank
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Female	-0.0127*	-0.195***	-0.194***	-47.66***	-0.0129*	-0.194***	-0.190***	-47.16***				
	(0.00762)	(0.0225)	(0.0230)	(6.149)	(0.00761)	(0.0226)	(0.0230)	(6.193)				
Reward	0.185***	0.0274*	-0.0218	29.06***	0.185***	0.0274*	-0.0218	29.06***	0.185***	0.0274*	-0.0218	29.06***
	(0.0158)	(0.0153)	(0.0149)	(3.901)	(0.0158)	(0.0153)	(0.0149)	(3.901)	(0.0155)	(0.0149)	(0.0146)	(3.821)
Female*Reward	0.333***	0.0462*	-0.0792***	-16.24**	0.333***	0.0462*	-0.0792***	-16.24**	0.333***	0.0462*	-0.0792***	-16.24**
	(0.0266)	(0.0270)	(0.0252)	(6.853)	(0.0266)	(0.0270)	(0.0252)	(6.853)	(0.0260)	(0.0264)	(0.0247)	(6.712)
Math at School	0.0219***	0.235***	0.2415***	58.63***	0.0220***	0.223***	0.225***	54.91***	-0.0372**	0.0627**	0.0697**	13.11*
	(0.00625)	(0.00878)	(0.0093)	(2.395)	(0.00624)	(0.00875)	(0.00924)	(2.389)	(0.0182)	(0.0272)	(0.0277)	(7.278)
Particiation Time	-0.0859***	0.315***	0.378***	94.68***	-0.0849***	0.233***	0.271***	68.86***	-0.0730	0.0422	0.0628	15.49
	(0.0149)	(0.0217)	(0.0240)	(5.885)	(0.0148)	(0.0213)	(0.0229)	(5.800)	(0.0540)	(0.0877)	(0.0897)	(24.48)
Observations	13,924	13,924	13,924	13,924	13,924	13,924	13,924	13,924	13,924	13,924	13,924	13,924
R-squared	0.139	0.271	0.308	0.354					0.105	0.028	0.040	0.095
Number of participants					6,134	6,134	6,134	6,134	6,134	6,134	6,134	6,134

Notes: Observations are at the Math test's parts level. The first three outcome variables, *No. Omitted, Prop. of Correct,* and *Score* are standardized at the edition, level and part of the test levels. *Rank* measures the position in the rank by edition, level and test's parts level, where higher values represent better positions within the rank. *Female* takes the value of 1 if the participant is female. *Reward* takes the value of 1 if the outcome variable refers to the part of the test with reward for omitted questions. *Math at School* measures the standardized Math grade at school level and *Participation Time* takes the values of 1, 2, 3 if it is the first, second or third time that the participant does the Math test. Columns 1-4 show the OLS specification where the standard errors are clustered at the participant level. Columns 5-8 show the RE model specification and columns 9-12 show the FE specfication model. All specifications include edition, level and school fixed effects. Standard errors in parenthesis, where *** p<0.01, ** p<0.05, * p<0.1

Table A4. Gender Differences between the No Reward and the Reward Parts of the Test along the Ability Distribution with Alternative Measure of Ability: Standardized Math at School Level

	Low Ability zomitted (1)	High Ability zomitted (2)	Interaction zomitted (3)	Continuous zomitted (4)
Female	0.173***	0.136***	0.196***	0.174***
	(0.0383)	(0.0330)	(0.0374)	(0.0249)
Reward	-0.0456*	-0.0459**	-0.0456*	-0.0452***
	(0.0241)	(0.0222)	(0.0235)	(0.0162)
Female*Reward	0.0759	0.191***	0.0759*	0.140***
	(0.0466)	(0.0390)	(0.0455)	(0.0296)
High Ability			-0.0279	
			(0.0353)	
High Ability*Reward			-0.000320	
			(0.0318)	
Female*High Ability			-0.0392	
			(0.0487)	
Female*Reward*High Ability			0.115*	
			(0.0591)	
Math at School	-0.0591	-0.0264	0.0469***	0.0300**
	(0.0471)	(0.0280)	(0.0152)	(0.0121)
Particiation Time	-0.0895**	-0.195***	-0.162***	-0.162***
	(0.0370)	(0.0244)	(0.0199)	(0.0199)
Math at School*Reward				0.0116
				(0.0166)
Female*Math at School				-0.0146
				(0.0252)
Female*Reward*Math at School				0.0414
				(0.0318)
Observations	7,708	10,114	17,822	17,822
R-squared	0.152	0.139	0.096	0.096

Notes: Observations are at the Math test's parts level. *No. Omitted* is standardized at the edition, level and part of the test levels. *Female* takes the value of 1 if the participant is female. Reward takes the value of 1 if the outcome variable refers to the part of the test with reward for omitted questions. *Math at School* measures the standardized Math grade at school level and *Participation Time* takes the values of 1, 2, 3 if it is the first, second or third time that the participant does the Math test. High Ability takes value 1 if the participant's standardized Math grade is>0. All columns show the OLS specification where the standard errors are clustered at the participant level and include edition, level and school fixed effects. Standard errors in parenthesis, where *** p<0.01, ** p<0.05, * p<0.1

Table A5. Gender Differences between the No Reward and the Reward Parts of the Test in the Matched Sample based on the No. Of Correct in the No Reward Part

	zcorrect (1)	zcorrect (2)	zomitted (3)	zprop_correct (4)	zscore (5)	rank (6)
Female	-0.216***	-0.0154	0.133***	0.0192	-0.0178	-4.667
	(0.0200)	(0.0229)	(0.0283)	(0.0236)	(0.0228)	(6.106)
Reward	0.0305**	0.180***	-0.0612***	0.142***	0.156***	76.02***
	(0.0134)	(0.0182)	(0.0227)	(0.0192)	(0.0178)	(4.847)
Female*Reward	-0.0975***	-0.247***	0.161***	-0.130***	-0.187***	-47.30***
	(0.0223)	(0.0256)	(0.0332)	(0.0279)	(0.0250)	(6.850)
Mat at School	0.205***	0.172***	0.0521***	0.214***	0.196***	50.11***
	(0.00800)	(0.00965)	(0.0109)	(0.00991)	(0.00972)	(2.573)
Particiation Time	0.381***	0.398***	-0.170***	0.327***	0.398***	106.5***
	(0.0215)	(0.0261)	(0.0263)	(0.0250)	(0.0259)	(6.573)
Observations	17,822	12,190	12,190	12,190	12,190	12,190
R-squared	0.278	0.275	0.107	0.240	0.286	0.340

Notes: Observations are at the Math test's parts level. The outcome variables, *No. of Correct*, *No. Omitted*, *Prop. of Correct*, and *Score* are standardized at the edition, level and part of the test levels. *Rank* measures the position in the rank by edition, level and test's parts level, where higher values represent better positions within the rank. *Female* takes the value of 1 if the participant is female. *Reward* takes the value of 1 if the outcome variable refers to the part of the test with reward for omitted questions. *Math at School* measures the standardized Math grade at school level and *Participation Time* takes the values of 1, 2, 3 if it is the first, second or third time that the participant does the Math test. Column 1 shows the estimation for the whole sample and columns 2 to 6 show the estimation results for the matched sample using the *No. of Correct* in the no reward part of the test with 3386 male and 3386 female participants. All regressions show the OLS specification where the standard errors are clustered at the participant level. All specifications include edition, level and school fixed effects. Standard errors in parenthesis, where *** p<0.01, ** p<0.05, * p<0.1

Table A6. Gender Differences between the No Reward and the Reward Parts of the Test: Confidence, Overconfidence or Risk? With Alternative Control for Ability: No. Of Corret No Reward

	Original Sample	Sample with Questionnaire	Sample with Questionnaire				
	zomitted	zomitted	zomitted	zomitted	zomitted	zomitted	zomitted
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	0.153***	0.181***	0.112***	0.115***	0.112***	0.114***	0.134***
	(0.0234)	(0.0327)	(0.0322)	(0.0323)	(0.0322)	(0.0322)	(0.0321)
Reward	-0.0447***	-0.0184	-0.0184	-0.110	-0.0489	0.0300	0.439***
	(0.0154)	(0.0226)	(0.0226)	(0.0708)	(0.114)	(0.0286)	(0.0427)
Female*Reward	0.132***	0.0652	0.0652	0.0590	0.0664	0.0624	0.0219
	(0.0282)	(0.0409)	(0.0409)	(0.0413)	(0.0410)	(0.0408)	(0.0405)
No. Of Correct No Reward	-0.0704***	-0.0638***	-0.0925***	-0.0925***	-0.0925***	-0.0925***	-0.0925***
	(0.00401)	(0.00558)	(0.00609)	(0.00609)	(0.00609)	(0.00609)	(0.00609)
Particiation Time	-0.0738***	-0.0843***	-0.0449**	-0.0449**	-0.0449**	-0.0449**	-0.0449**
	(0.0184)	(0.0247)	(0.0223)	(0.0223)	(0.0223)	(0.0223)	(0.0223)
Non-Competitiveness			0.0164	0.00355	0.0164	0.0164	0.0164
			(0.0104)	(0.0142)	(0.0104)	(0.0104)	(0.0104)
Perceived Math Ability			-0.0230	-0.0230	-0.0267	-0.0230	-0.0230
			(0.0160)	(0.0160)	(0.0209)	(0.0160)	(0.0160)
Overconfidence			-0.0358***	-0.0358***	-0.0358***	-0.0295***	-0.0358***
			(0.00291)	(0.00291)	(0.00291)	(0.00340)	(0.00291)
Risk			-0.209***	-0.209***	-0.209***	-0.209***	-0.101***
			(0.00927)	(0.00927)	(0.00927)	(0.00927)	(0.0109)
Non-Competitiveness*Reward				0.0256			
				(0.0190)			
Perceived Math Ability*Reward					0.00733		
					(0.0265)		
Overconfidence*Reward						-0.0126***	
						(0.00411)	
Risk*reward							-0.217***
							(0.0161)
Observations	19,766	9,284	9,284	9,284	9,284	9,284	9,284
R-squared	0.112	0.146	0.209	0.209	0.209	0.210	0.224

Notes: Column 1 shows the main estimation result from column 1 in Table A2 for the original sample. For the rest of the columns, observations are at the Math test's parts level in the edition of 2017 and 2018 for the participants whose questionnaire answers are available. *Female* takes the value of 1 if the participant is female. *Reward* takes the value of 1 if the outcome variable refers to the part of the test with reward for omitted questions. *No. of Correct No Reward* measures the number of correct questions in the part of the test without any reward for omitted question and *Participation Time* takes the values of 1, 2, 3 if it is the first, second or third time that the participant does the Math test. *Non-Competitiveness* contains the responses to question 1 in the questionnaire. Perceived Math Ability contains the responses to question 9 in the questionnaire. *Perceived Math Ability* contains the responses to question 9 in the questionnaire. *Overconfidence* measures the difference between the guessed number of correct answers and the actual number of correct answers. And *Risk* contains the responses to question 8 in the questionnaire. All columns show the OLS specification where the standard errors are clustered at the participant level and include edition, level and school fixed effects. Standard errors in parenthesis, where *** p<0.01, ** p<0.05, * p<0.1

Table A7. Gender Differences between the No Reward and the Reward Parts of the Test: Splitting the Reward part into two (Q14-Q19 and Q20-Q25)

	zomitted	zprop_correct	zscore
	(1)	(2)	(3)
Female	0.172***	-0.171***	-0.221***
	(0.0251)	(0.0211)	(0.0205)
Reward_Q14-Q19	-0.0409**	-0.0129	-0.000474
	(0.0162)	(0.0147)	(0.0138)
Reward_Q20-Q25	0.00851	-0.0164	-0.0122
	(0.0178)	(0.0156)	(0.0147)
Female*Reward_Q14-Q19	0.132***	0.0314	-0.00891
	(0.0298)	(0.0254)	(0.0231)
Female*Reward_Q20-Q25	0.0743**	0.0411	0.0281
	(0.0329)	(0.0270)	(0.0244)
Math at School	0.0374***	0.196***	0.199***
	(0.00820)	(0.00707)	(0.00740)
Particiation Time	-0.141***	0.264***	0.339***
	(0.0197)	(0.0180)	(0.0200)
Observations	26,733	26,733	26,733
R-squared	0.098	0.172	0.222

Notes: Observations are at the Math test's parts level. The first three outcome variables, *No. Omitted*, *Prop. of Correct*, and *Score* are standardized at the edition, level and part of the test levels. *Female* takes the value of 1 if the participant is female. *Reward_Q14-Q19* takes the value of 1 if the outcome variable refers to the part of the test with reward for omitted questions Q14 to Q19. *Reward_Q20-Q25* takes the value of 1 if the outcome variable refers to the part of the test with reward for omitted questions Q20 to Q25. *Math at School* measures the standardized Math grade at school level and *Participation Time* takes the values of 1, 2, 3 if it is the first, second or third time that the participant does the Math test. Columns 1-3 show the OLS specification where the standard errors are clustered at the participant level. All specifications include edition, level and school fixed effects. Standard errors in parenthesis, where *** p<0.01, ** p<0.05, * p<0.1