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[**]

December 26, 2018

Very preliminary, please do not circulate

Abstract

We study gender differences in willingness to guess in a multiple-choice math test with about 10,000 participants, where in half of the questions both wrong answers and omitted questions score 0, and in the other half wrong answers score 0 but omitted questions score +1. Using a within-participant regression analysis, we find that female participants leave more omitted questions than males under both types of scoring rules, but when there is a reward for omitted questions, the gender difference gets even larger. This gender difference, which is stronger among high ability and older participants, has negative consequences for females in the final score and ranking. In a subsequent survey, female participants show lower levels of confidence and higher risk aversion, which could potentially explain this differential behavior. When both are considered, risk aversion shows to be the main factor in explaining the gender differential in the willingness to guess. A scoring rule that is gender neutral begs for non-differential scoring between wrong answers and omitted questions.

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

JEL classification: J16, C93, D81, I20.

[*] We thank seminar participants at various institutions and conferences for helpful comments. We are particularly thankful to the organizers of *Concurso de Primavera de Matemáticas*, who collaborated throughout the project. We also thank Lluis Puig and Ana Montiel Rey for excellent research assistance. Nagore Iriberri acknowledges financial support from 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.

1 Introduction

Multiple choice tests are one of the most frequently used means to measure students' knowledge and aptitude. Performance in different multiple-choice tests plays a crucial role in shaping students' labor market outcomes, since they are extensively used to determine grade point averages and thus, university admission and postgraduate studies. Examples of standardized tests based on multiple-choice tests that play a key role in shaping students' future outcomes abound all over the world. Scholastic Aptitude Test (SAT) and Graduate Record Examination (GRE) are two important examples in the USA. One important decision a multiple choice test designer needs to take is whether wrong and omitted questions (questions with no answer) score the same or not. The main motivation for wrong answers and omitted questions to score differently is to avoid getting the question right by chance, which adds noise to the measure of knowledge and aptitude. However, one important concern is that multiple-choice tests in which wrong 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 misrepresent students' knowledge and aptitude. An extensive 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). Hence, an informed decision on the optimal scoring rule regarding omitted questions and wrong answers begs the study of its effect on gender differences in the willingness to guess and ultimately in performance.

In collaboration with the organizers of *Concurso de Primavera de Matemáticas*, we carry out a large-scale natural field experiment to test for and understand the mechanism behind gender differences in the 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 editions, with a total of about 10,000 participants, we designed tests with no differential score to omitted questions and wrongs answers for the first 13 of the 25 test questions (no reward for omitted) and where in the last 12 test questions wrong answers score 0 and omitted questions +1 (reward for omitted). We compare within-participant performance and behavior across both parts of the test.

We find that female participants leave on average more omitted questions than males in the no reward part (0.17 standard deviations of the mean), but most importantly, this difference increases in the reward part of the test (in 0.14 standard deviations of the mean). The gender differential in the willingness to guess has important consequences in the gender differences in the final score and the ranking of participants. The female underperformance increases in 0.05 standard deviations of the mean and they lose about 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 lower ability ones. However, we find that indeed the gender differential for the willingness to guess is stronger among high ability participants (0.26 standard deviations of the mean), while we find no gender differential for the low ability participants. Second, using the four different age categories in the exam, we explore the differential gender effect of the scoring rule across different ages. Participants in final years of their primary education show no gender differences between the reward and no-reward parts of the tests but students in secondary education and High School students do.

Motivated by the gender difference we found in the 2016 data and in order to understand the underlying mechanism behind it, we designed a questionnaire that would allow us to measure the effects of confidence and risk aversion, which we administered in the 2017 and 2018 editions of the test. Regarding confidence, we mainly use two measures: confidence on their perceived ability in math and the difference between their guessed number of correct answers and the actual number of correct answers ("overconfidence"). Regarding risk, we ask them: "When omitting a question was worth 1 point I answered the question …." where the participants can give 5 different answers: "when they are absolutely sure" (safest option), "when they are almost sure", "when they are doubting between 2 potential answers", "when they are doubting between 3 potential answers" and "always" (riskiest option). Female participants show lower confidence in their perceive math ability, lower overconfidence and higher risk aversion. When these measures of confidence, overconfidence and risk are controlled for 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 differential scoring rule for wrong answers and omitted questions, mostly using observational data (Swineford, 1941; Anderson, 1989; Atkins et al., 1991; Ramos and Lambating, 1996; Tannenbaum D., 2012, Akyol, Key, and Krishna, 2016). Only recently, there have been important advances in pursuing control-treatment studies in the laboratory (Baldiga, 2014), carrying out field-experiments (Ben-Shakhar and Sinai, 1991, Funk and Perrone, 2016) and using before-after type of quasi-controlled studies (Coffman and Klinowski, 2018) to test for the causal effect of differential scoring rule on male and female test takers' behavior and performance. Although all these studies find that female students leave more omitted than males when there is differential grading of wrong answers and omitted questions, there is divergence 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 mainly, which they rationalize because females in their setting show on average higher ability. Akyol, Key and Krishna (2016) estimate negative effects for females and for risk averse students but conclude 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 a significant impact on the gender gap in performance.

Our study differs from existing control-treatment studies in the following ways. First, the differential scoring rule rewards omitted questions rather than having penalties for wrong answers. Espinosa and Gardeazabal (2013) show that these two are only strategically equivalent under risk neutrality and that under risk aversion penalties will lead to more omitted than rewards. Therefore, in our setting, the significant gender difference in guessing and its non-negligible effects on performance and relative performance outcomes show that they are still an important concern even though they represent a milder and less non-favorable scoring rule for females than the use of penalties. Second, using a very similar within-participant treatment assignment as in Funk and Perrone (2016), we add a larger sample than the existing field natural experiments, as well as evidence from a different setting. Funk and Perrone (2016) study classroom behavior of undergraduate students in a large Microeconomics class, while we study behavior in a regional math contest. Third, similar in spirit to the laboratory experiment by Baldiga (2014), we can also contribute to the underlying mechanism and test how much of the gender differences in the willingness to guess are due to risk aversion and overconfidence, reaching a similar conclusion in that gender differences in risk aversion shows to be the main factor. Finally, most recent studies have shown interesting heterogeneity effects regarding ability differences (Funk and Perrone, 2016, and Akyol, Key and Krishna, 2016). We find high ability female participants are indeed more affected which resonates the results by Akyol, Key and Krishna (2016). We also add heterogeneity effects regarding age, which as far as we know no other studies does. It is important to understand when gender differences appear and how they evolver with age.

The paper has the following structure. Section 2 describes the setting and the data. 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 their website, 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. In every edition, about 40,000 students participate in the first stage math test and about 3,000 students in the second stage math test. Iriberri and Rey-Biel (forthcoming) analyzed gender differences between the two stage math tests, which differ in their competitive pressure, using the 2014 edition data. In this study, we use data from one unique math test (the second stage math test), from the editions of 2016, 2017 and 2018.

A large number of schools from Madrid participate in this initiative. As shown in Iriberri and Rey-Biel (forthcoming) the sample of schools that participate ranges between the 30% of 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, the participating schools contain a lower proportion of public schools, have larger numbers of students and, as expected, show better results in mathematics, as

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

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 set out. First, there are four different tests according to age groups, which we refer to as levels 1 to 4, 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, so participants are 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. Secondly, the math test takes place in the campus of Universidad Complutense de Madrid, in a pre-specified day in April. Thirdly, the top three contestants in each level obtain prizes. Additionally, the top 5% participants get a diploma and a small gift in a public ceremony.[3] Fourth, the test for each level is made up of 25 multiple-choice questions, all of them set by the organizers. The questions for each level are designed so that students in the lower school 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 the 2015 edition, the scoring rule was the same for all the 25 questions: 0 for wrong answers, +1 point for omitted questions, and +5 points for correctly answered questions. For the 2016, 2017 and 2018 editions, we collaborated with the organizers such that the math test had two parts that would differ in their scoring rule. For the first 13 questions, the grading system awards 0 points to both omitted questions and wrong answers, and +5 points to questions answered correctly. For the following 12 questions, question 14-25, the grading system awards 0 points to wrong answers, +1 point for omitted questions and +5 points to questions answered correctly. See Figure A1 in the Appendix to see how the scoring rule was described to the participants. We explicitly designed the test such that other things, i.e. the content or difficulty of the questions, did not vary between the first and the second parts of the math test. See the mean values of correct answers per question for all

[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 can be checked on the website, what the main prizes will be is not revealed ex-ante. In past editions, 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.

the questions in the math test in Figure A2, where we do not see any clear differences between the two parts of the test.

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

2.2 Descriptive Statistics

The database consists of the participants who take the math test in 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 variable means across gender for the continuous variables and Fisher 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 tests takers are male. Looking at control variables, we can see that students can participate in different editions (*Participation Time*) and we observe that 183 participants take the math test in the three editions, 1,167 repeat twice and the rest of 7,023 students we just observe them once. Female participants are less likely to participate more than once. The three different editions do not show big differences in overall participation or female participation. Regarding the participation in different levels, Level 2 is the most popular and the Level 4 the least popular, which has a lower number of participants. Female representation is lowest in the last level, which is partially explained by female students being less likely to choose the math-science track in High School.

Performance data includes the rank, score, the number of correct and omitted questions for each of the two parts 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 about 90% of participants. 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 show higher performance than male ones (8.55 for females and 8.32 for males). However, gender differences reverse when looking at the score in the math test we study in both parts of the test, as female participants obtain a lower score than male participants. In the first part of the test, when there is no reward for omitted, men obtain a score of 29.50 and women obtain a score of 26.50, and in the second part, when there is a reward for omitted, boys obtain 23.30 and girls obtain 20.67. 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. For regression analysis, all performance measures are standardized at the year, level and test part levels. This also translates into the ranking between male and female participants. Female participants rank lower than males, on average about 51 positions behind, and this difference gets larger in the math test with the reward for omitted questions, 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. Figure 1 shows the cumulative distribution of the *No. of Omitted* by gender when there is no reward (top) and when there is reward for omitted questions (bottom), which complements nicely the descriptive statistics in Table 1. Note that when no reward, the optimum behavior implies answering all questions, while when there is reward, the optimum behavior depends on one's knowledge, overconfidence and risk aversion. Although participants should answer all questions when there is no reward for omitted, participants indeed omit on average 0.65 questions. In addition, women leave slightly more questions unanswered, 0.86 questions: while 80% of male participants indeed answer all questions, only 74% of female participants do so. More importantly, when there is reward, participants on average leave 4.82 questions answered, male participants 4.51 and female participants leave 5.40. In both panels of Figure 1, the distribution of female participants stochastically dominates that of male participants, and the differences in the reward part are larger. Male participants also have higher number of correct answers and higher proportion of correct but these differences are not large across the two parts. In section 3.1, we will proceed to measure gender differences in the number of omitted questions between the reward and the no reward parts of the math test, which is the main focus of this study.

[Figure 2 about here]

Finally, panel b in Table 1 shows the descriptive statistics on the control variables we collected in the survey administered in the 2017 and 2018 editions. The variables of interest are the *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 very similar number of reported hours devoted to prepare 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 given most participants' answers lie below that value, which again shows that both male and female participants show very similar values for the number of preparation hours.

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, so the more positive the value the higher the overconfidence. Figure 2b plots the observations where the x-axis shows the number of correct answers and the y-axis the number of guessed correct answers. Both male and female participants are overconfident, however, consistent with other findings, female participants show lower values of overconfidence. Note that overconfidence is measuring a lower bound on the gender difference, as it restricts 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 by Figure 2c (see question 9 in Figure A3 in the Appendix), so perceived ability in math is higher for male than for female participants. Finally, somehow also related to confidence, we measure participants' perception of the gender nature of the math task (see question 10 in Figure A3 in the Appendix). As shown by Figure 2d, the large majority of participants believe math to be gender neutral, such that men and women are equally good/knowledgeable at math. However, both genders show some home-bias: a small fraction of male participants believes men are better at

[4] 15 participants provided very high numbers of preparation hours. We replaced those values with missing to avoid outliers.

math than women and a small fraction of female participants believe women are better at math than men.

Finally, 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 (1 for "When Absolutely Sure" to 5 for "Always") such that the higher the number of the answer the more risk loving the participant is. Figure 2d shows the histogram of all the possible answers by gender. Clearly, more female participants answer the question when absolutely and almost sure than males. Note that this risk measure is affected by overconfidence, as perceived probability of knowing the answer might also be affected by participants' confidence on own ability. In sections 3.4, we will proceed to understand the underlying mechanism for female participants leaving more omitted questions using all these measures for confidence, overconfidence and risk.

3. Results

3.1. Do Female Participants Leave More Omitted from Having No Reward To When Having A Reward For Omitting Questions?

We start by looking at whether female participants react differently from male participants in their strategy to leave a question omitted or not, 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 edition, level and part of the test, for all outcome variables. Table 2 shows the estimation results.

[Table 2 about here]

Columns 1-4 show the estimation results for the OLS specification where we cluster standard errors 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 in the no reward part (0.17 standard deviations of the mean), but most importantly, this difference increases in the reward part of the test (in 0.14 standard deviations of the mean). This is not the case for the proportion of correct answers. Despite female participants showing a lower proportion of correct answers (0.16 standard deviations of the mean), this difference does not increase in the reward part. Women leaving relatively more omitted than men in the reward part compared to the non-reward part has important consequences for how male and female participants perform under different reward systems. Female participants perform on average worse than males (0.21 standard deviations of the mean) and they get lower positions in the ranking (51 positions behind) in the no reward part of the test. More importantly, this gap gets larger when there is reward for omitted questions. Regarding the score, the gender gap increases in 0.05 standard deviations of the mean. Regarding the positions in the ranking, the gender gap increases in about 10 more positions. In other words, the female underperformance increases when moving from the no reward to the reward part of the test.

Columns 5-8 and columns 9-12 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, they allow testing for the robustness of the main effects. The variable of interest, the interaction between *Female* and *Reward* keeps the same magnitude and significance levels. From now on, we will use the OLS estimation, where we cluster the standard errors at the participant level.

We finally comment on the effect of the two main control variables: math at school and experience in 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. Somehow 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 negatively correlated to the number of omitted, which is more in line with what one would expect. 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. Further, 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 in the no reward part. The results in the main variable of interest, the interaction between *Female* and *Reward* are very similar in both the magnitude and the significance levels.

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 that participants vary substantially in their ability. We can think of two proxies for ability. First, if we take the number of correct answers in the part with no reward as a proxy for ability, we can observe in Figure 3 that there is a large variation. The number of correct answers vary between 0 and 13 where the median is at 6. Second, the math grade at school also shows some variation but definitely less. Given its larger variation, we use the number of correct answers in the no reward part of the test as a proxy for ability and use the variation in math at school as a robustness test, which we will discuss at the end of the section.

[Figure 3 over here]

We now study if the gender differential in the number of omitted questions from the no reward part of the test to the reward part of the test varies substantially along the distribution of participants' ability.

[Figure 4 over here]

Figure 4 shows graphical evidence on gender differences by ability. Figure 4a shows the number of omitted in the no reward and reward parts for the omitted questions, by low and high ability and by gender. We define low ability if the standardized number of correct answers in the reward part is below 0 and high ability if the standardized number of correct answers is above 0. As expected, higher ability participants leave fewer omitted questions, in both parts of the test. Also, female participants always leave more omitted. 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 takes a closer look at the number of omitted in the reward part of the test by gender. Lower and higher ability female participants behave similarly, although as expected higher ability participants leave fewer questions omitted. However, for male participants, lower and higher ability participants' behavior differs substantially, particularly with significantly more participants omitting no question at all, which is less evident for female participants.

[Table 3 about here]

Table 3 shows the results for the number of omitted questions. We take two complementary approaches. First, 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 in the no reward part of the test, such that the standardized value is equal or below 0 is labeled low ability and any value above 0 is labeled high ability. 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 high, as high ability female participants show a differential reaction to the reward part leaving more omitted compared to male participants, 0.26 standard deviations of the mean. As shown by column 3, the triple interaction between *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 in the no reward part of the test. Column 4 shows the interaction between *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.

As a robustness test, we also perform the same exercise but taking math as proxy for ability. Table A2 shows the results. The conclusions are very similar, when looking at the interaction between *Female* and *Reward* for the low and high ability. However, here the interaction is non-significant, probably due to the lower number of observations, when using math as 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 of young age (in their fifth and sixth academic years of primary school) and older age (in their final two years of High School). 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 lower academic levels (0.02 standard deviations of the mean for the youngest participants in their 5th grade in primary school) to the higher academic levels (0.236 among the oldest participants in High School).

Column 5 shows the results when all levels are included in one same regression, where we can test how significantly different the gender differences are across different 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. Therefore, we do find evidence that the gender differential in willingness to guess when there is reward for omitted is larger among older participants compared to younger participants.

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

Female participants leave more omitted questions than males and this is harmful for their performance outcomes in the math test. Furthermore, this negative effect seems to be larger for the high ability participants and older participants. Can we shed some light on the underlying mechanism?

In principle, there can be three underlying reasons for such an effect. First, male and female participants can 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 in the no reward part of the test, we do see that while male participants get about 5.90 correct answers, females get about 5.29 correct answers. However, in all our analysis so far we do control for ability in math (either using math grade at school or the number of correct answers in the no reward part of the test).

Second, male and female participants can differ in their perceived knowledge of math, to which we will refer as confidence. We have three different variables that measure their perceived ability in math. First, in question 9 in the questionnaire, we ask participants to rate how agreeable they are with the statement: “I am good at math”. Clearly, as shown by Figure 2c, female participants show lower levels of agreement with that statement. Second, in question 10 in the questionnaire, we ask participants to say whether male participants are better/equally good/worse than women. As shown by Figure 2d, there seems to be high degree of agreement among both male and female participants that both male and female participants are equally good. Third, we ask participants to guess the number of correctly answered questions. Both male and female participants seem to be overconfident (Beyer, 1999), as they expect to get more correct than what 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), we indeed find that female participants show higher risk aversion than males. Remember that we ask participants when they decide to answer a question, to which they can provide 5 possible answers (from safest strategy: only answer when absolutely sure, to the riskiest strategy: answer always).

[Table 5 about here]

We now proceed to test if any of these measures indeed has explanatory power for the gender differential in the number of omitted from the no reward to the reward part of the test, such that, when these variables are controlled for, whether the female differential is still significant. Table 5 shows the estimation results from this exercise. Notice, however, we collect all these measures in the questionnaire administered right after the test in the editions of 2017 and 2018 so we do not have these measures for all our participants. Column 1 and 2 show the main specification, as in Table 2, but in the sample for which we have control variables collected in the questionnaire. In column 2, we find that, as in the main sample, female participants leave more omitted questions than males when moving from the no reward to the reward part of the test. The magnitude is slightly lower. In column 3, we add the three main control variables: perceived ability in math, 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 number of omitted questions. The female coefficient goes down but the interaction of *Female* and *Reward* is exactly the same as in column 2. In columns 4, 5 and 6, we interact each of the control variables with the variable *Reward*. When adding these interactions with respect to the two overconfidence measures, the main coefficient of interest, the interaction between *Female* and *Reward*, hardly changes, suggesting overconfidence is not explaining why female participants leave more omitted. However, when interacting the risk measure with the reward, we clearly see that the coefficient of *Female* and *Reward*, goes down substantially such that it is no longer significant. This shows that risk differences between male and female participants are indeed the main factor in explaining why male and female participants differ in their behavior in omitting questions.

Table A3 in the appendix shows the estimation results with the alternative measure of the number of correct answers in the no reward part of the test. Although the main coefficient of interest becomes non-significant the results are qualitatively the same, as when controlling for participants' risk preferences changes substantially the differential reaction of female participants as to when to respond questions.

We have observed that gender differences in risk aversion seem to be explaining most of gender differences in number of omitted questions. What about the gender differences found between the low and high ability participants? In Section 3.2, we found that the gender difference in the number of omitted question was harming high ability participants in particular. Could it be that gender differences in risk and overconfidence are different between the low and high ability participants? How much of the gender differences in the number of omitted among the low and high ability can be explained by overconfidence and risk?

We first have a look at gender differences in confidence, overconfidence and risk by ability. Figure A4 show the graphs. Gender differences are present both among the high and low ability participants and they always go in the same direction, female participants show lower perceived ability in math, lower levels of overconfidence and higher risk aversion. However, the gender differences between the low and the high ability participants do not seem to show significant differences.

[Table 6 about here]

We therefore 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 the answers to the questionnaire. 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 the questionnaire answers. Columns 5 and 6 add the main control variables on 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 are behind the higher gender differences among the high ability participants. However, it is also important to notice 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 part of the differences remain unexplained.

4. Conclusions

Using performance data from a natural field experiment we test for gender differences in the willingness to guess when there is differential grading from omitting questions and providing a wrong answer. We find that women always leave more omitted questions but that this behavior gets more prominent when there is differential grading for omitted questions and wrong answers, having negative consequences for female participants, both in the final score and ranking. We also find that this gender differential is stronger among the high ability participants and older participants. Finally, gender differences in risk aversion explain most of the gender differential in the willingness to guess. Based on this evidence, we conclude that a gender neutral grading rule begs for non-differential scoring of omitted and wrong answers.

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. (2018). “Competitive Pressure Widens the Gender Gap in Performance: Evidence from a Two-Stage Competition in Mathematics.” Forthcoming in The Economic Journal.

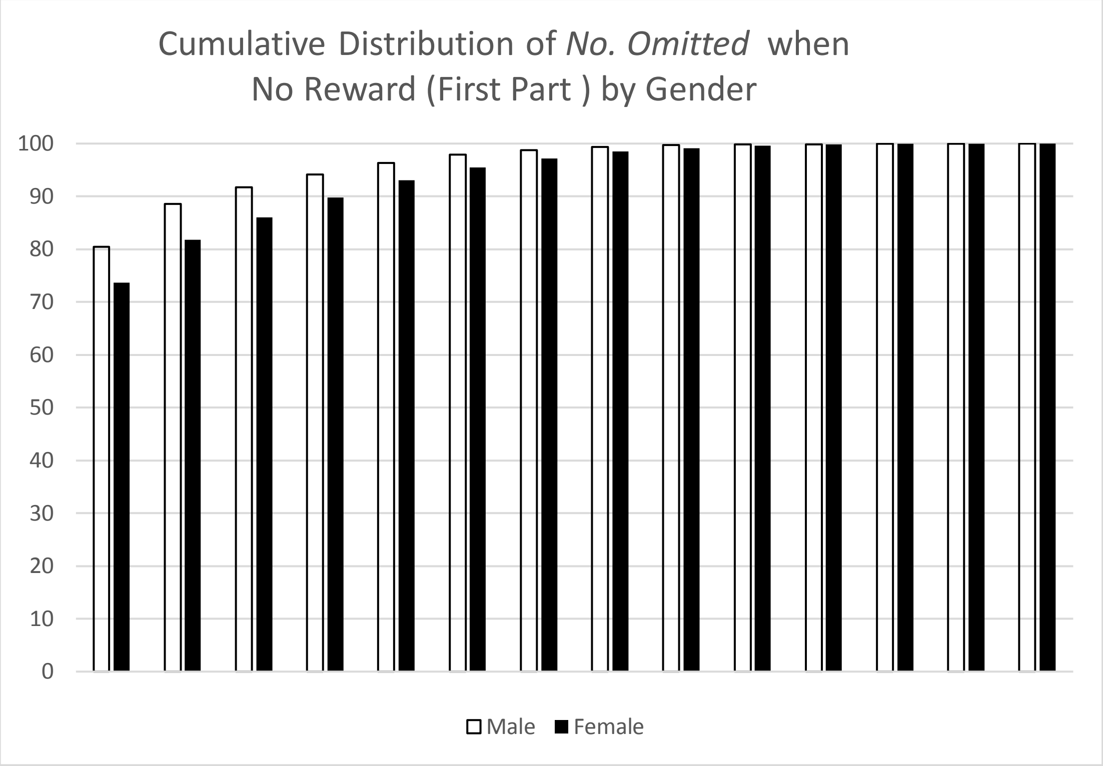
Ramos, I. and Lambating, J. (1996). “Gender Difference in Risk-Taking Behavior and their Relationship to SAT-Mathematics Performance.” School Science and Mathematics, 96(4): 202-207.

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**



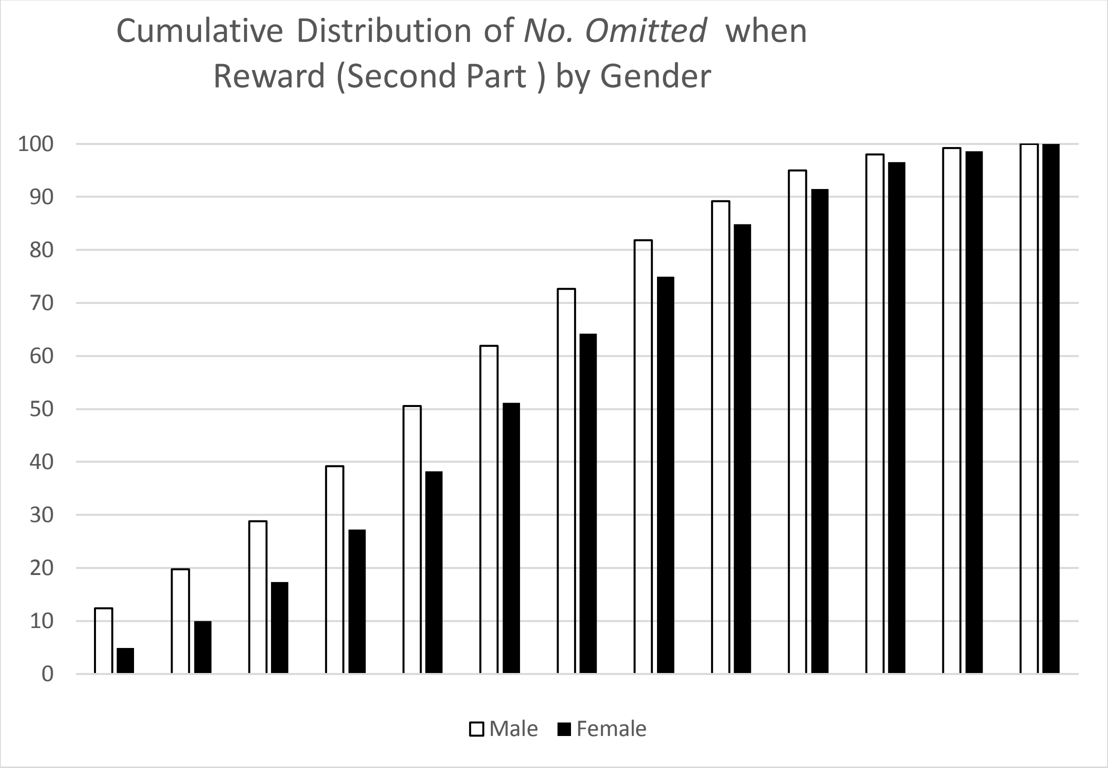
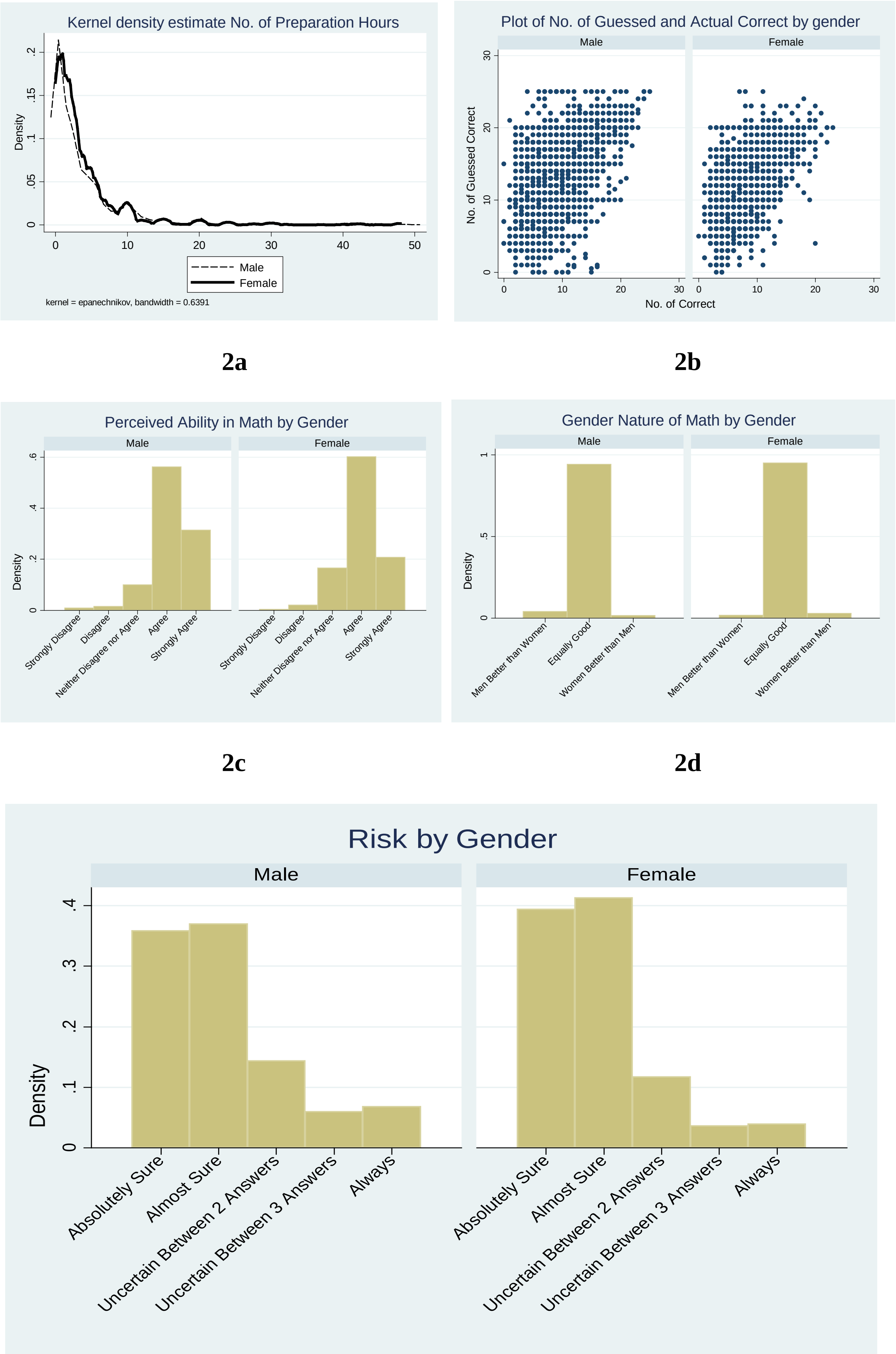


Figure 2. Descriptive Statistics on the Control Variables from the Questionnaire



2a

2b

2c

2d

2e

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

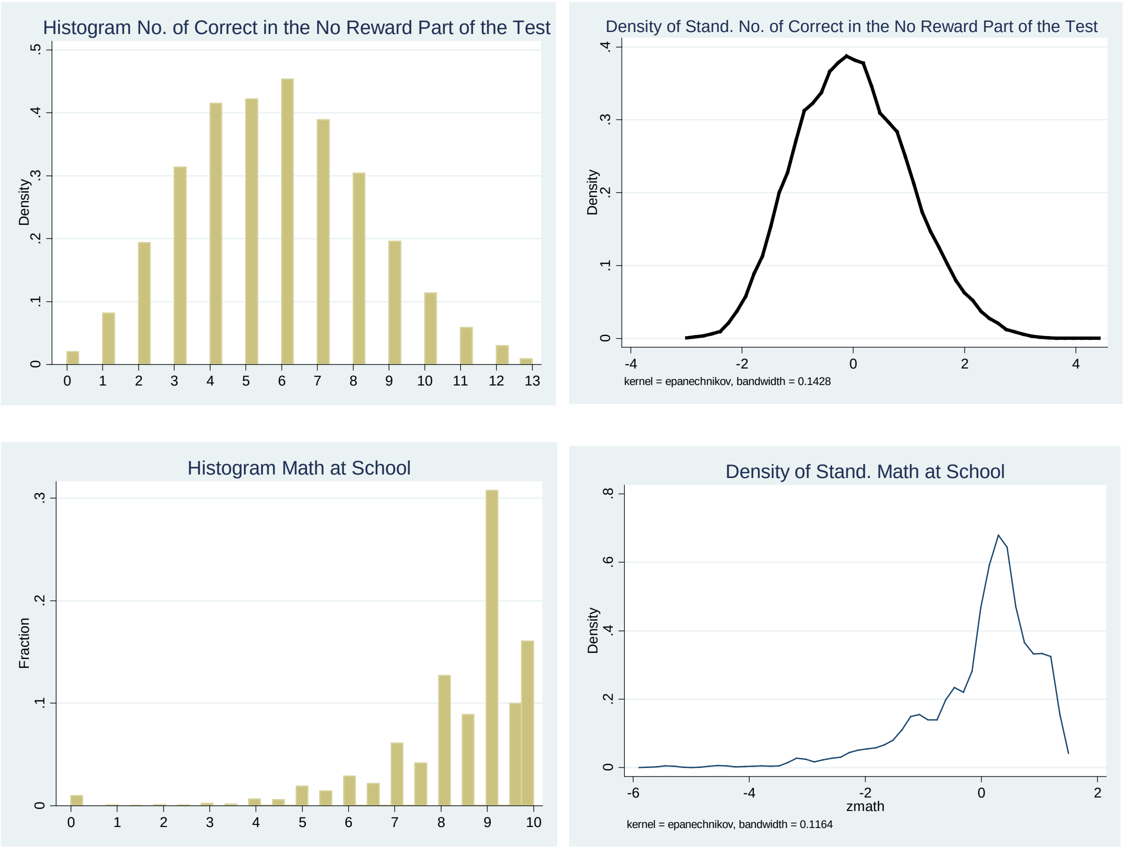
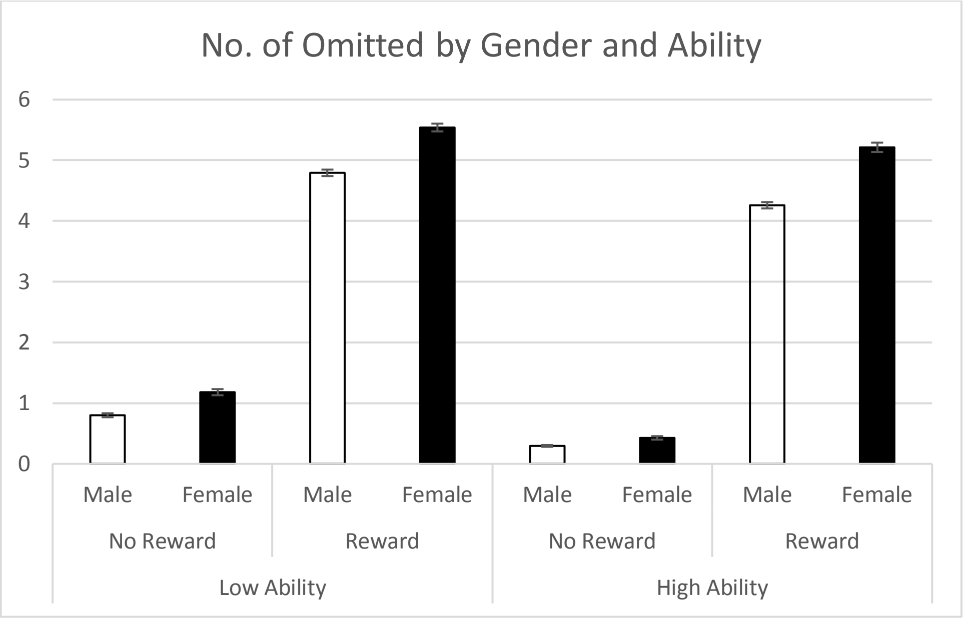
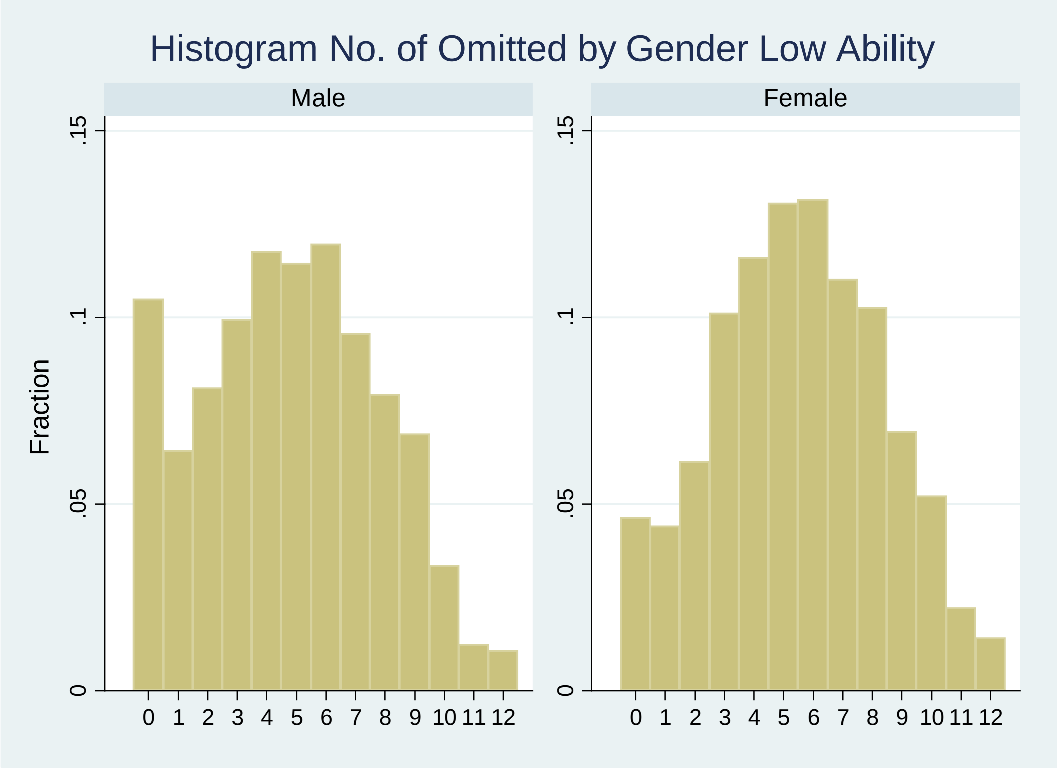


Figure 4. No. of Omitted by Gender and Ability: Low Ability: No. of Correct in No Reward<6 and High Ability: No. of Correct in No Reward>6

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

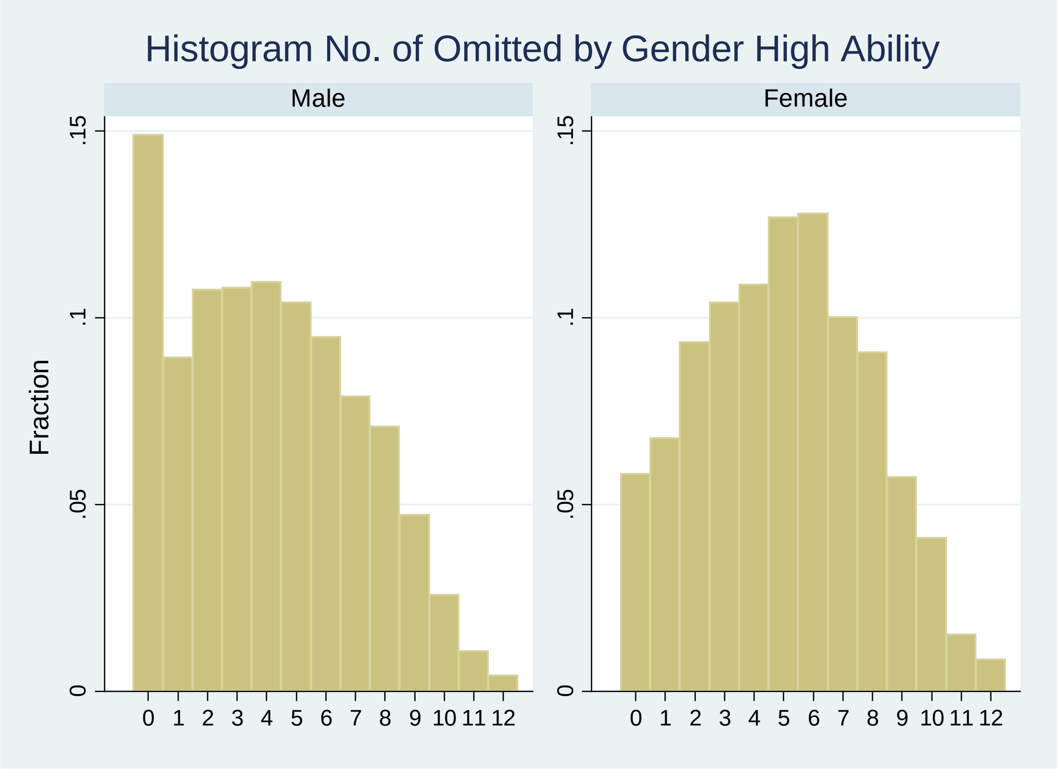


Table 1. Descriptive Statistics

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
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. *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	0.167***	-0.165***	-0.213***	-51.33***	0.168***	-0.168***	-0.213***	-51.98***				
	(0.0248)	(0.0212)	(0.0206)	(5.466)	(0.0251)	(0.0212)	(0.0205)	(5.458)				
Reward	-0.0462***	-0.00868	0.0133	39.38***	-0.0462***	-0.00868	0.0133	39.38***	-0.0462***	-0.00868	0.0133	39.38***
	(0.0162)	(0.0140)	(0.0131)	(3.456)	(0.0162)	(0.0140)	(0.0131)	(3.456)	(0.0159)	(0.0138)	(0.0129)	(3.401)
Female*Reward	0.144***	0.0209	-0.0433**	-10.32*	0.144***	0.0209	-0.0433**	-10.32*	0.144***	0.0209	-0.0433**	-10.32*
	(0.0294)	(0.0245)	(0.0218)	(5.920)	(0.0294)	(0.0245)	(0.0218)	(5.920)	(0.0289)	(0.0241)	(0.0214)	(5.826)
Math at School	0.0380***	0.155***	0.146***	36.05***	0.0360***	0.143***	0.131***	32.68***	-0.0314**	0.0343**	0.0405**	7.444*
	(0.00633)	(0.00654)	(0.00660)	(1.682)	(0.00634)	(0.00661)	(0.00666)	(1.744)	(0.0154)	(0.0150)	(0.0160)	(4.305)
Particiation Time	-0.167***	0.320***	0.401***	102.1***	-0.138***	0.221***	0.273***	69.86***	-0.126*	-0.0107	0.0479	0.955
	(0.0197)	(0.0207)	(0.0225)	(5.493)	(0.0185)	(0.0202)	(0.0210)	(5.342)	(0.0685)	(0.0763)	(0.0751)	(20.89)
Observations	17,950	17,950	17,950	17,950	17,950	17,950	17,950	17,950	17,950	17,950	17,950	17,950
R-squared	0.098	0.229	0.280	0.328					0.015	0.025	0.040	0.101
Number of participants					7,833	7,833	7,833	7,833	7,833	7,833	7,833	7,833

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 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. 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	High Ability	Interaction	Continuous
	zomitted	zomitted	zomitted	zomitted
	(1)	(2)	(3)	(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 across Age

	Level 1	Level 2	Level 3	Level 4	Overall
	zomitted	zomitted	zomitted	zomitted	zomitted
	(1)	(2)	(3)	(4)	(5)
Female	0.212***	0.195***	0.0773*	0.171**	0.187***
	(0.0498)	(0.0453)	(0.0461)	(0.0700)	(0.0468)
Reward	0.00156	-0.0621**	-0.0587*	-0.0634	-0.0462***
	(0.0317)	(0.0288)	(0.0313)	(0.0423)	(0.0162)
Female*Reward	0.0209	0.163***	0.184***	0.236***	0.0686
	(0.0582)	(0.0521)	(0.0549)	(0.0861)	(0.0504)
Math at School	0.0134	0.0163	0.0751***	0.0254	0.0382***
	(0.0145)	(0.0123)	(0.0105)	(0.0159)	(0.00634)
Participation Time	-0.296***	-0.182***	-0.176***	-0.0879*	-0.168***
	(0.0562)	(0.0388)	(0.0345)	(0.0455)	(0.0198)
Level 2					0.0457
					(0.0319)
Level 3					0.0913***
					(0.0335)
Level 4					0.147***
					(0.0383)
Female*Level 2					0.0178
					(0.0619)
Female*Level 3					-0.0673
					(0.0626)
Female*Level 4					-0.0514
					(0.0819)
Level 2*Female*Reward					0.0783
					(0.0636)
Level 3*Female*Reward					0.103
					(0.0648)
Level 4*Female*Reward					0.150*
					(0.0862)
Observations	4,338	5,770	5,046	2,796	17,950
R-squared	0.169	0.151	0.181	0.213	0.099

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 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. 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: Overconfidence or Risk?

	Original Sample	**Sample with Questionnaire**	**Sample with Questionnaire**			
	zomitted	zomitted	zomitted	zomitted	zomitted	zomitted
	(1)	(2)	(3)	(4)	(5)	(6)
Female	0.167***	0.168***	0.119***	0.119***	0.120***	0.139***
	(0.0248)	(0.0347)	(0.0346)	(0.0347)	(0.0346)	(0.0345)
Reward	-0.0462***	-0.0262	-0.0262	-0.000954	0.0273	0.426***
	(0.0162)	(0.0239)	(0.0239)	(0.119)	(0.0303)	(0.0452)
Female*Reward	0.144***	0.0999**	0.0999**	0.0989**	0.0968**	0.0583
	(0.0294)	(0.0428)	(0.0428)	(0.0430)	(0.0427)	(0.0424)
Math at School	0.0382***	0.0431***	0.0355***	0.0355***	0.0355***	0.0355***
	(0.00633)	(0.0100)	(0.00969)	(0.00969)	(0.00969)	(0.00969)
Particiation Time	-0.164***	-0.178***	-0.153***	-0.153***	-0.153***	-0.153***
	(0.0190)	(0.0261)	(0.0242)	(0.0242)	(0.0242)	(0.0242)
Perceived Math Ability			-0.0680***	-0.0650***	-0.0680***	-0.0680***
			(0.0168)	(0.0215)	(0.0168)	(0.0168)
Overconfidence			-0.0126***	-0.0126***	-0.00572*	-0.0126***
			(0.00264)	(0.00264)	(0.00317)	(0.00264)
Risk			-0.205***	-0.205***	-0.205***	-0.0977***
			(0.00993)	(0.00993)	(0.00993)	(0.0117)
Perceived Math Ability*Reward				-0.00604		
				(0.0276)		
Overconfidence*Reward					-0.0138***	
					(0.00431)	
Risk*reward						-0.214***
						(0.0170)
Observations	17,950	8,370	8,370	8,370	8,370	8,370
R-squared	0.098	0.138	0.186	0.186	0.187	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 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. *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

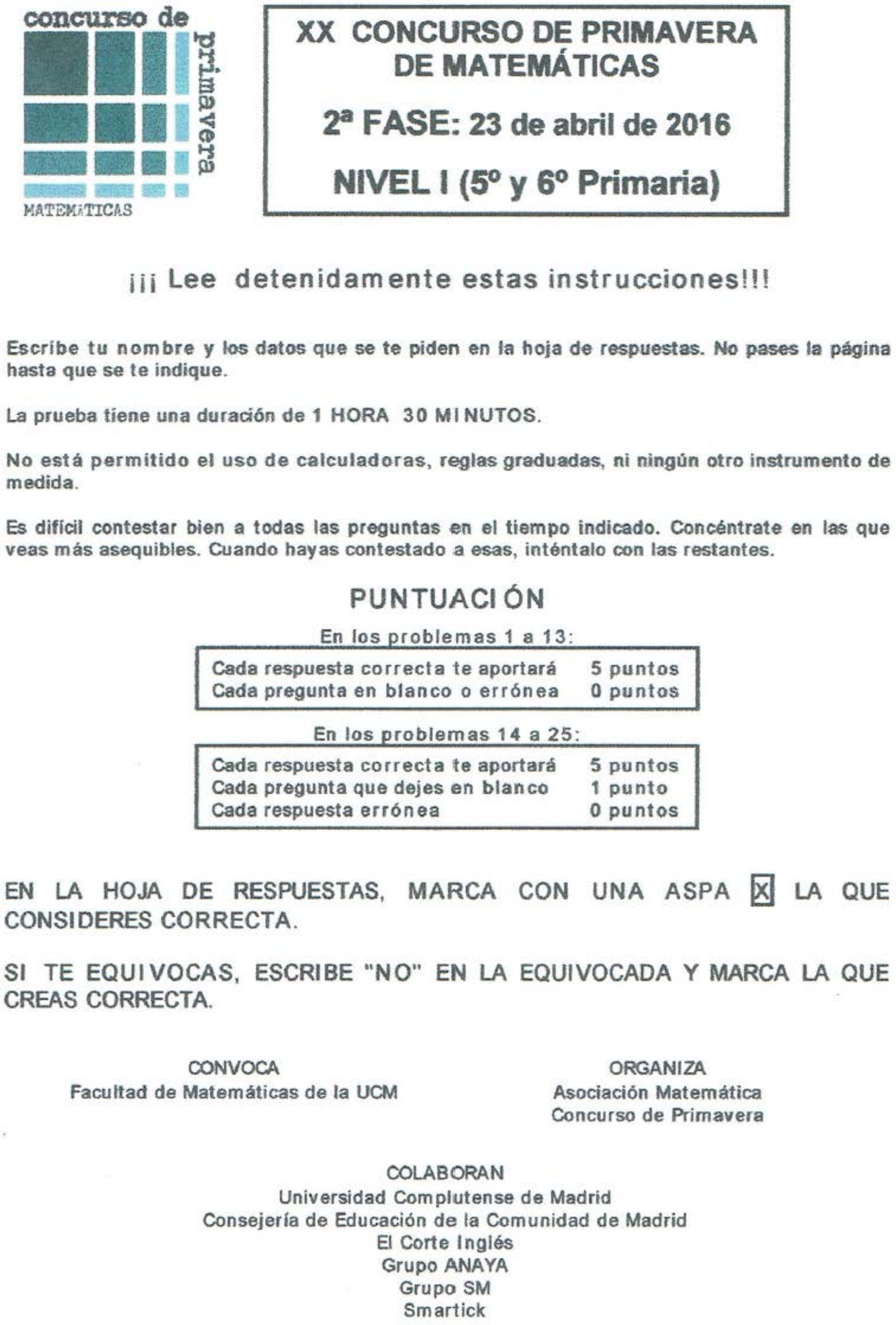
Table 6. Gender Differences between No Reward and the Reward Part of the Test along the Distribution of Ability: Overconfidence or Risk?

	Original Sample		ample with Questionnair		Sample with Questionnaire							
	Low Ability	High Ability	Low Ability	High Ability	Low Ability	High Ability	Low Ability	High Ability	Low Ability	High Ability	Low Ability	High Ability
	zomitted	zomitted	zomitted	zomitted	zomitted	zomitted	zomitted	zomitted	zomitted	zomitted	zomitted	zomitted
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Female	0.191***	0.0847***	0.180***	0.142***	0.138**	0.0720**	0.136**	0.0753**	0.138**	0.0782**	0.154***	0.103***
	(0.0387)	(0.0225)	(0.0551)	(0.0361)	(0.0538)	(0.0357)	(0.0538)	(0.0358)	(0.0537)	(0.0356)	(0.0536)	(0.0352)
Reward	-0.0928***	-0.00151	-0.0488	0.0104	-0.0488	0.0104	-0.154	0.190	-0.0515	0.0558	0.348***	0.549***
	(0.0248)	(0.0194)	(0.0367)	(0.0295)	(0.0367)	(0.0295)	(0.174)	(0.159)	(0.0559)	(0.0342)	(0.0690)	(0.0546)
Female*Rew	0.0505	0.264***	-0.0181	0.228***	-0.0181	0.228***	-0.0149	0.221***	-0.0179	0.215***	-0.0505	0.165***
	(0.0425)	(0.0358)	(0.0628)	(0.0542)	(0.0628)	(0.0542)	(0.0628)	(0.0546)	(0.0628)	(0.0542)	(0.0625)	(0.0528)
No. Of Corre	-0.0809***	-0.0577***	-0.0810***	-0.0441***	-0.123***	-0.0593***	-0.123***	-0.0593***	-0.123***	-0.0593***	-0.123***	-0.0593***
	(0.0110)	(0.00509)	(0.0177)	(0.00853)	(0.0177)	(0.00839)	(0.0177)	(0.00839)	(0.0177)	(0.00839)	(0.0177)	(0.00839)
Particiation T	-0.0525	-0.0976***	-0.0155	-0.119***	0.0375	-0.0981***	0.0375	-0.0981***	0.0375	-0.0981***	0.0375	-0.0981***
	(0.0384)	(0.0187)	(0.0529)	(0.0279)	(0.0481)	(0.0249)	(0.0481)	(0.0249)	(0.0481)	(0.0249)	(0.0481)	(0.0249)
Perceived M					-0.0215	-0.0213	-0.0345	-0.000162	-0.0215	-0.0213	-0.0215	-0.0213
					(0.0253)	(0.0184)	(0.0350)	(0.0199)	(0.0253)	(0.0184)	(0.0253)	(0.0184)
Overconfide					-0.0461***	-0.0210***	-0.0461***	-0.0210***	-0.0464***	-0.0114***	-0.0461***	-0.0210***
					(0.00454)	(0.00358)	(0.00454)	(0.00358)	(0.00588)	(0.00374)	(0.00454)	(0.00358)
Risk					-0.230***	-0.192***	-0.230***	-0.192***	-0.230***	-0.192***	-0.136***	-0.0647***
					(0.0159)	(0.0117)	(0.0159)	(0.0117)	(0.0159)	(0.0117)	(0.0199)	(0.0108)
Perceived M							0.0259	-0.0423				
							(0.0419)	(0.0362)				
Overconfide									0.000520	-0.0191***		
									(0.00691)	(0.00616)		
Risk*reward											-0.189***	-0.254***
											(0.0250)	(0.0214)
Constant	0.809*	-0.0952	-0.281***	-0.301	0.751***	0.328**	0.804***	0.238	0.752***	0.306*	0.553***	0.0591
	(0.438)	(0.107)	(0.0858)	(0.190)	(0.168)	(0.167)	(0.195)	(0.170)	(0.171)	(0.167)	(0.170)	(0.167)
Observations	10,048	9,718	4,746	4,180	4,746	4,180	4,746	4,180	4,746	4,180	4,746	4,180
R-squared	0.122	0.153	0.191	0.210	0.249	0.269	0.249	0.269	0.249	0.271	0.258	0.301

Notes : Observations are at the Math test's parts level. Columns 1-2 show the main estimation results from columns 1-2 in Table 3 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. *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. 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)

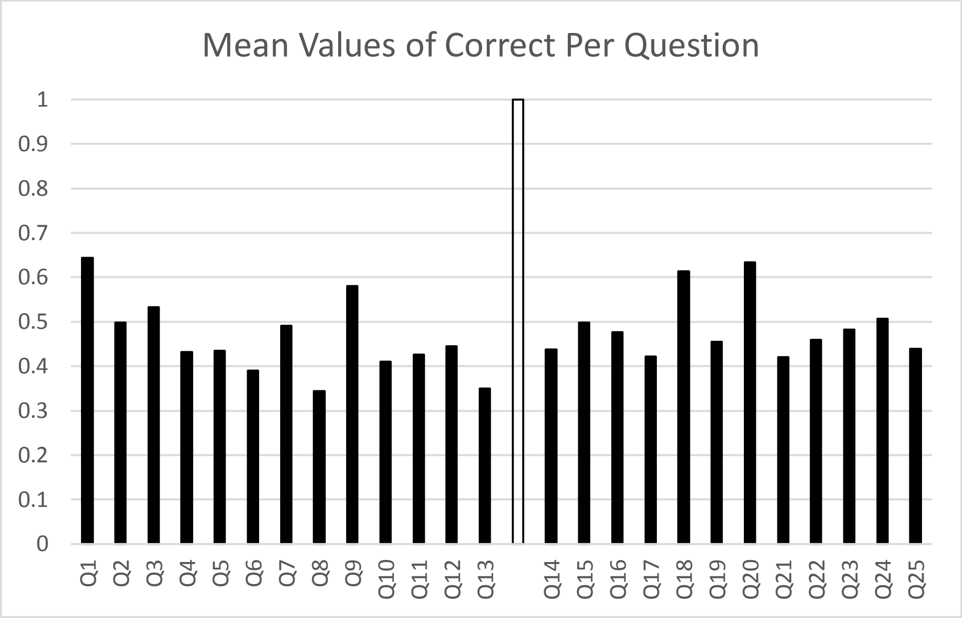


Figure A3. Questionnaire at the end of the Test

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

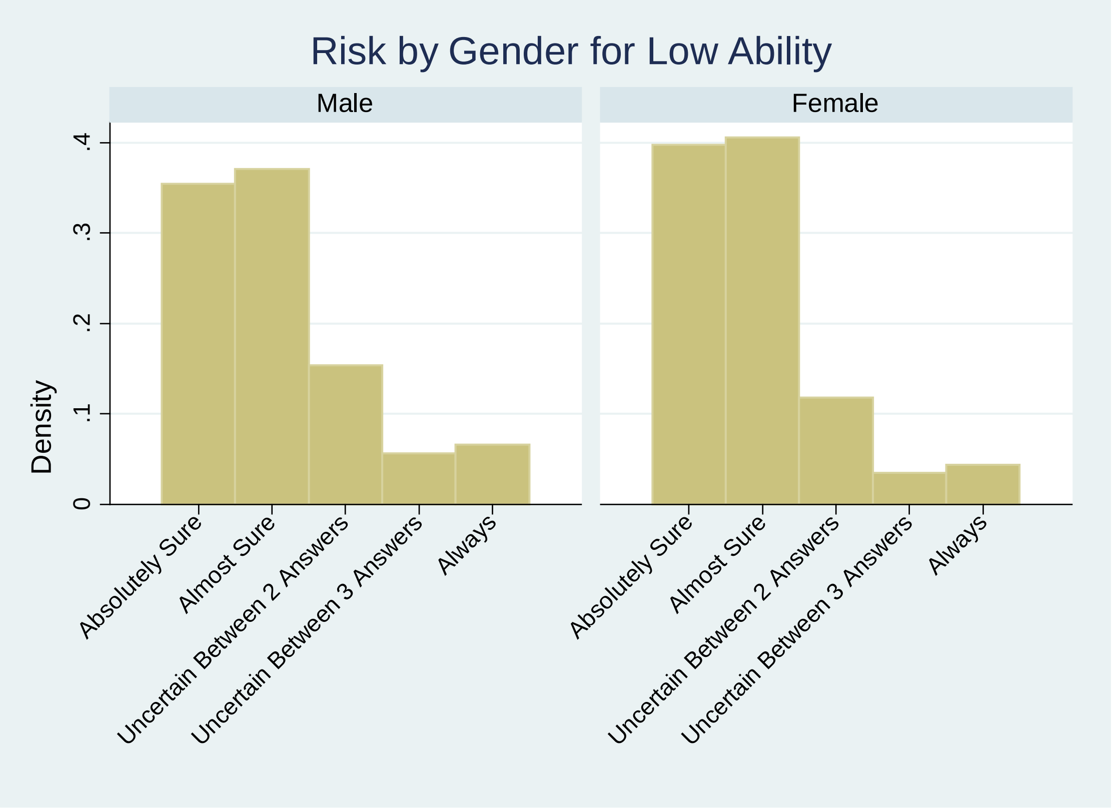
1. "It is more important to me being selected for Stage 2 than being among the winners in Stage 2."
2. "It is more important to my parents being selected for Stage 2 than being among the winners in Stage 2."
3. "It is more important to do well in Stage 2 than in Stage 1."
4. "I have devoted more hours 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"

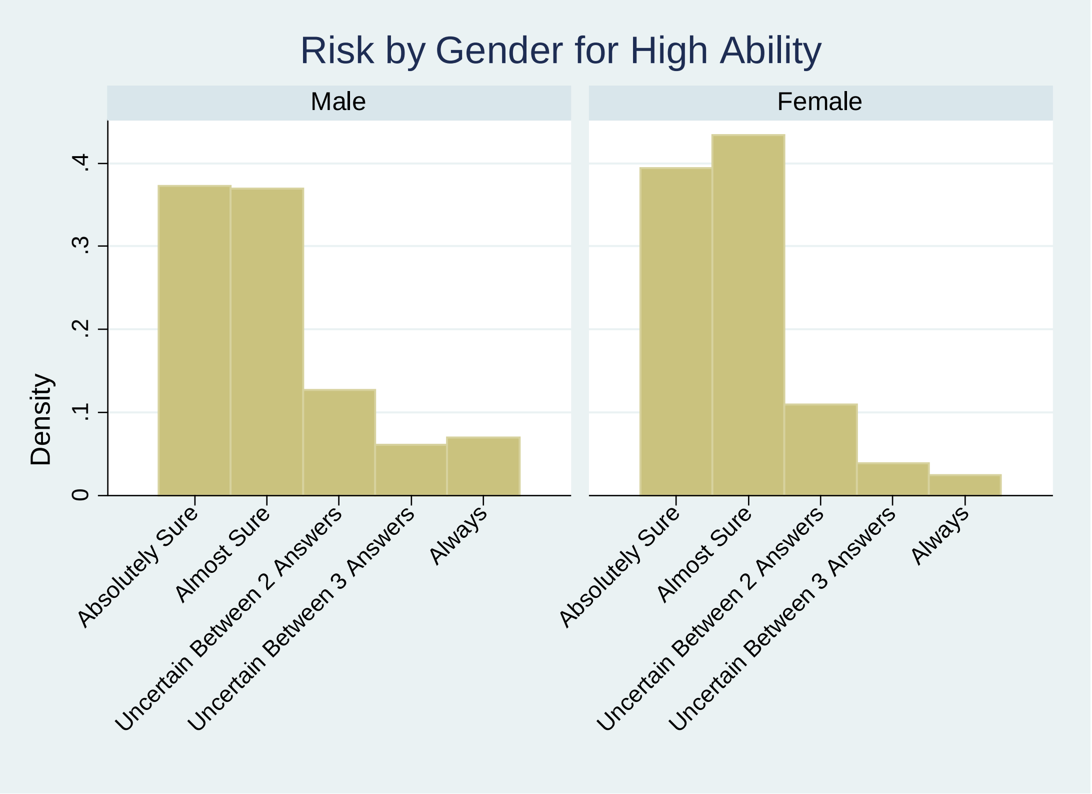
9. "I am good at Mathematics"

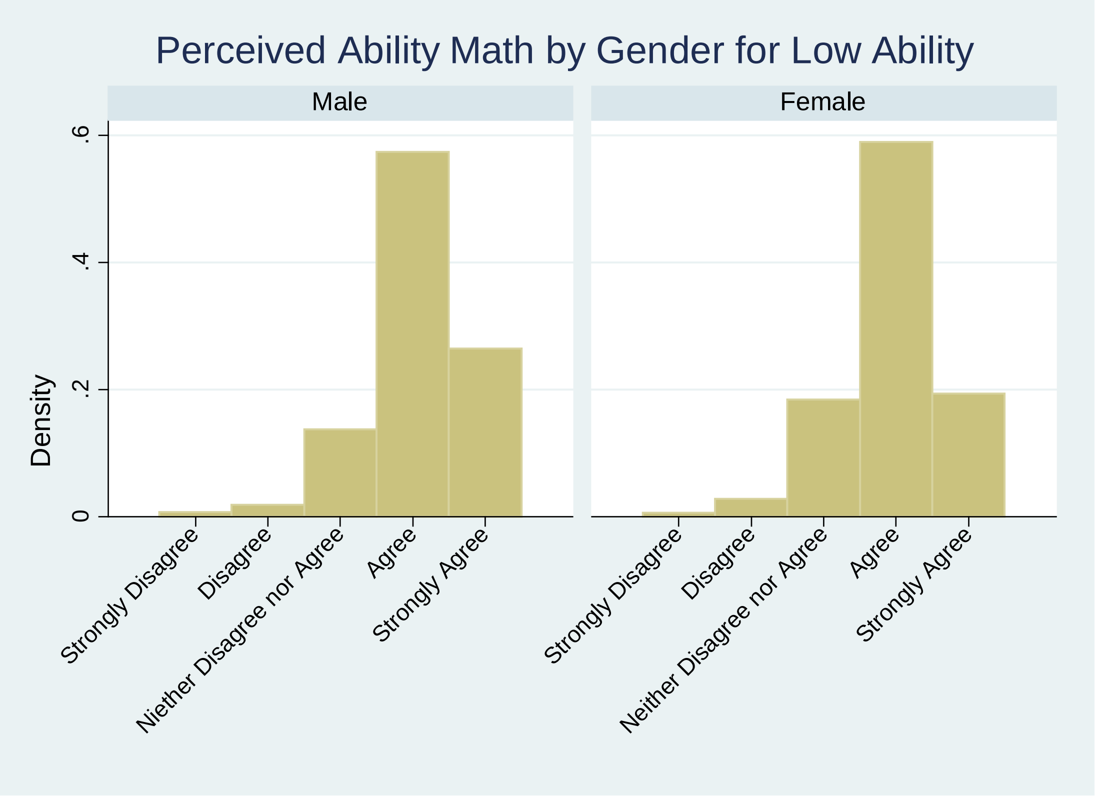
5. How many hours did you devote to prepare Stage 2 test?
7. How many questions do you expect to get right?
8. When omitting a question was worth 1 point I answered the question ______
   a. when I was absolutely sure.
   b. when I was almost sure.
   c. when I was uncertain between 2 answers.
   d.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

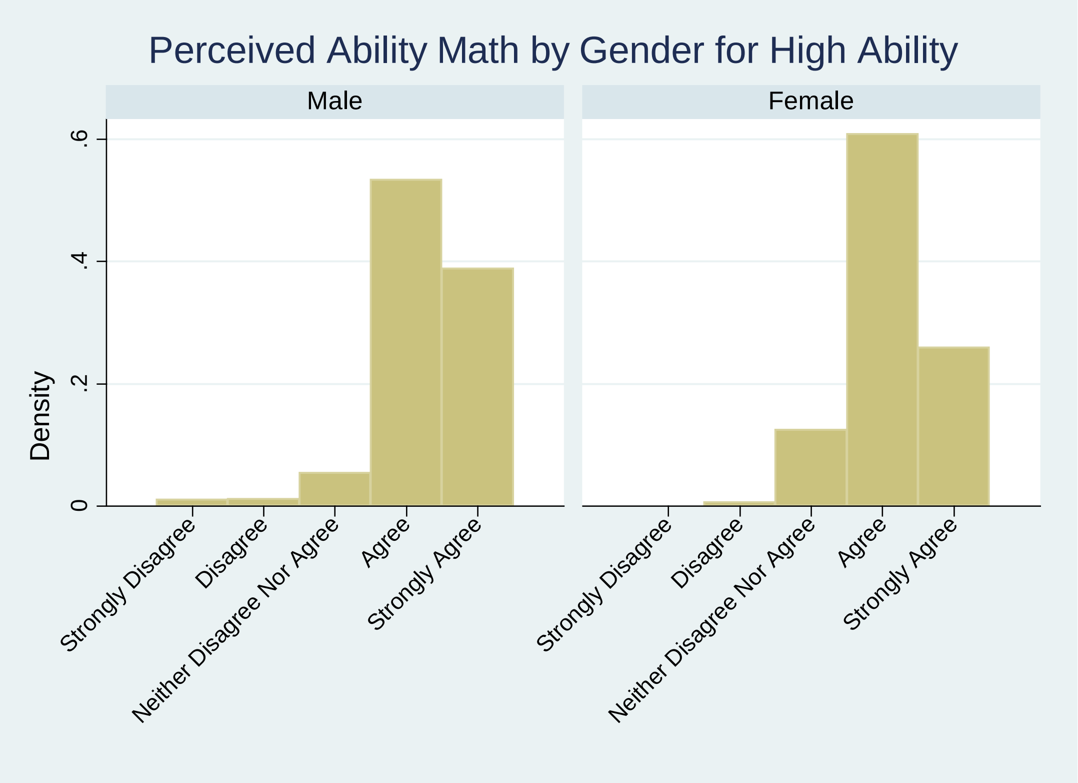
10. I believe _______ at Math
    a. men are better than women
    b. men and women are equally good
    c. women are better than men

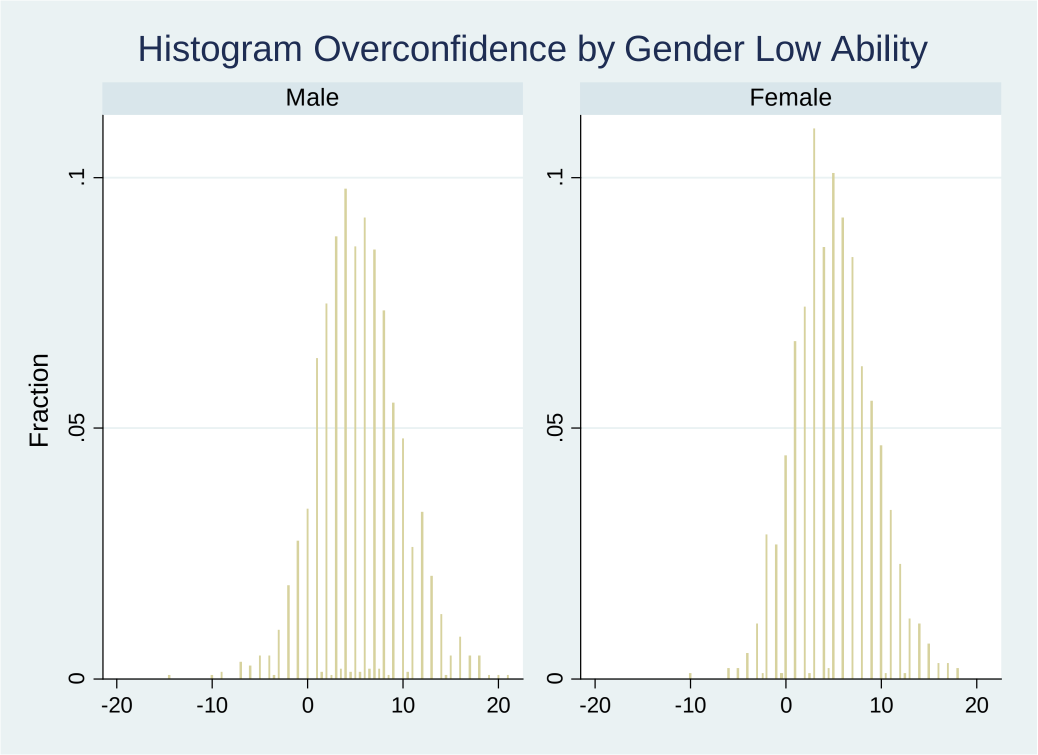
Figure A4. Risk, Confidence and Overconfidence by Gender: Low Ability: No. of Correct in No Reward<6 and High Ability: No. of Correct in No Reward>6






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

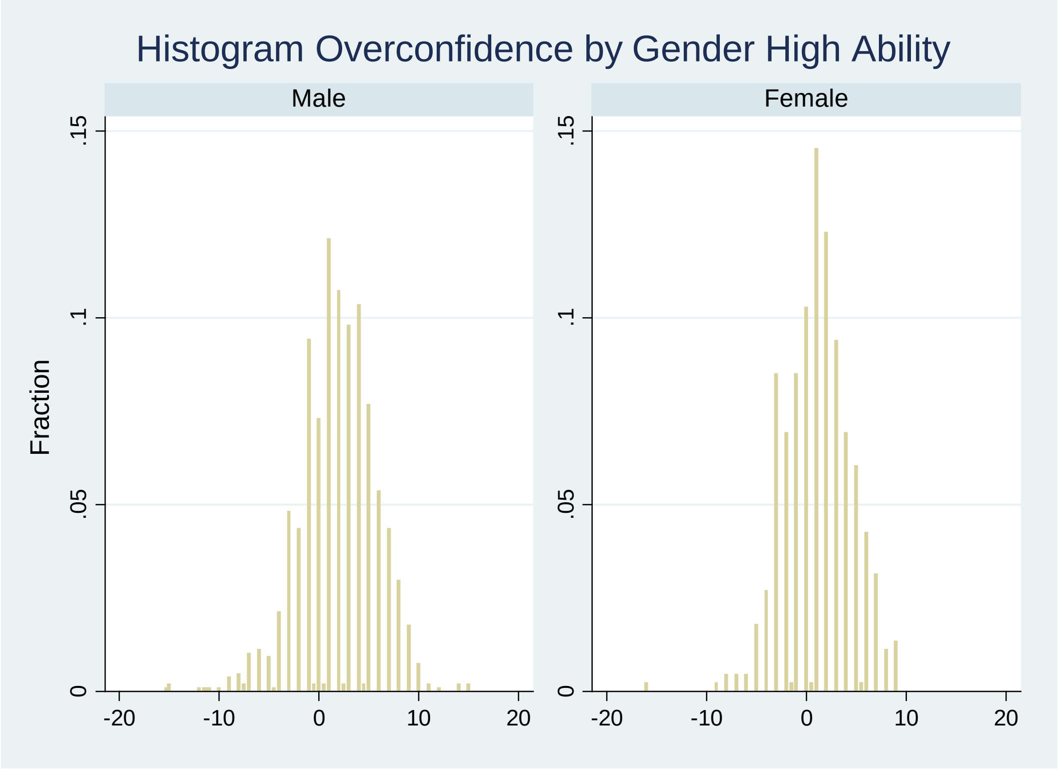

Histogram Overconfidence by Gender High Ability
Male
Female
Fraction
0
.05
.1
.15
-20
-10
0
10
20

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 No Reward and the Reward Parts of the Test along the Ability Distribution with Alternative Measure of Ability: Math at School

	Low Ability	High Ability	Interaction	Continuous
	zomitted	zomitted	zomitted	zomitted
	(1)	(2)	(3)	(4)
Female	0.173***	0.141***	0.171***	0.140
	(0.0324)	(0.0399)	(0.0319)	(0.126)
Reward	-0.0390*	-0.0580**	-0.0390*	-0.129
	(0.0203)	(0.0266)	(0.0201)	(0.0873)
Female*Reward	0.106***	0.200***	0.106***	-0.0898
	(0.0384)	(0.0464)	(0.0379)	(0.164)
High Ability			-0.0694***	
			(0.0266)	
High Ability*Reward			-0.0190	
			(0.0321)	
Female*High Ability			-0.0115	
			(0.0490)	
Female*Reward*High Ability			0.0945	
			(0.0582)	
Math at School	0.0499***	0.0431*	0.0518***	0.0280***
	(0.00886)	(0.0230)	(0.00763)	(0.00868)
Particiation Time	-0.132***	-0.200***	-0.164***	-0.167***
	(0.0280)	(0.0296)	(0.0197)	(0.0198)
Math at School*Reward				0.00995
				(0.0103)
Female*Math at School				0.00336
				(0.0147)
Female*Reward*Math at School				0.0271
				(0.0191)
Observations	10,924	7,026	17,950	17,950
R-squared	0.123	0.162	0.099	0.098

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 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. 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 A3. Gender Differences between No Reward and the Reward Parts of the Test:
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
	(1)	(2)	(3)	(4)	(5)	(6)
Female	0.153***	0.179***	0.114***	0.113***	0.115***	0.135***
	(0.0234)	(0.0326)	(0.0321)	(0.0321)	(0.0320)	(0.0319)
Reward	-0.0447***	-0.0189	-0.0189	-0.0498	0.0285	0.439***
	(0.0154)	(0.0226)	(0.0226)	(0.113)	(0.0285)	(0.0426)
Female*Reward	0.132***	0.0666	0.0666	0.0678*	0.0639	0.0234
	(0.0282)	(0.0408)	(0.0408)	(0.0409)	(0.0407)	(0.0404)
No. Of Correct No Reward	-0.0704***	-0.0638***	-0.0934***	-0.0934***	-0.0934***	-0.0934***
	(0.00401)	(0.00558)	(0.00610)	(0.00610)	(0.00610)	(0.00610)
Particiation Time	-0.0740***	-0.0853***	-0.0462**	-0.0462**	-0.0462**	-0.0462**
	(0.0184)	(0.0246)	(0.0222)	(0.0222)	(0.0222)	(0.0222)
Perceived Math Ability			-0.0224	-0.0261	-0.0224	-0.0224
			(0.0161)	(0.0209)	(0.0161)	(0.0161)
Overconfidence			-0.0358***	-0.0358***	-0.0296***	-0.0358***
			(0.00290)	(0.00290)	(0.00338)	(0.00290)
Risk			-0.211***	-0.211***	-0.211***	-0.103***
			(0.00926)	(0.00926)	(0.00926)	(0.0109)
Perceived Math Ability*Reward				0.00742		
				(0.0265)		
Overconfidence*Reward					-0.0123***	
					(0.00410)	
Risk*reward						-0.217***
						(0.0160)
Observations	19,766	9,310	9,310	9,310	9,310	9,310
R-squared	0.112	0.146	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. *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