Gender Differences in Academic Performance: The Role of Negative Marking in Multiple-Choice Exams

Patricia Funk

Università della Svizzera italiana

Helena Perrone

Universitat Pompeu Fabra & Barcelona GSE

First Draft: December 2016

Abstract

We investigate whether penalizing wrong answers on multiple-choice tests ("negative marking") makes females relatively worse off compared to males (the comparison being no penalties for wrong answers). With a cohort of more than 500 undergraduate students at a major Spanish university, we conducted a field experiment in the Microeconomics course. We created a final exam, which was composed of two parts: one with penalties for wrong answers and one without. Students were randomly allocated to different exam permutations, which differed in the questions that carried penalties for wrong answers. We find that the penalties did not harm female students. Females performed better than males on both parts of the exam and did so to a greater extent on the part with penalties. Whereas risk aversion did not affect overall scores (despite affecting answering behavior), ability did. High-ability students performed relatively better with negative marking, and these were more likely to be women.

*Correspondence: Patricia Funk, IdEP, Università della Svizzera italiana, Email: patricia.funk@usi.ch. Helena Perrone, Universitat Pompeu Fabra, Department of Economics and Business, email: helena.perrone@upf.edu. We would like to thank seminar participants at the Barcelona GSE Summer Forum, Freie Universitaet Berlin, University of Amsterdam, Universitat Pompeu Fabra and Università della Svizzera italiana for helpful comments. Financial support from the Barcelona GSE Seed Grants is gratefully acknowledged.

1 Introduction

In recent years, there have been frequent discussions on whether the SAT, the US college entrance exam, has design features that disfavor women. Whether due to the SAT's emphasis on speed, its multiple-choice format, or the grade penalties for wrong answers, various studies have found that SAT test scores underpredict women's college grades relative to men's (Leonard & Jiang, 1999). Pressured by scientific studies and the press, the SAT was recently revised, and the new SAT, in place since March 2016, no longer penalizes wrong answers.[1]

To date, we do not yet have a clear answer on whether dropping penalties for wrong answers ("negative marking") improves the relative performance of women. One explanation for the adverse impact on women of penalizing wrong answers is risk aversion: women tend to be more risk averse than men, and this could induce them to guess too little if wrong answers are penalized (Baldiga, 2014). It seems essential to shed light on this topic, as several countries use university admission tests similar to the SAT (e.g., Turkey, China, Japan). The results of these tests are therefore crucial determinants of subsequent academic and professional achievement. Furthermore, at the university level, there is discussion on the consequences of negative marking, and some universities have recently changed policies with respect to that.[2] Using test designs that do not disfavor females is important, as women are underrepresented at top-level jobs and typically earn lower wages than men (Bertrand & Hallock, 2001; ONeill, 2003).

The goal of this article is to investigate whether negative marking on multiple-choice tests decreases women's performance (relative to men) compared to a testing design with no grade penalty for wrong

[1] Various studies compare performance on multiple-choice tests relative to essays and typically find that women perform relatively better on the latter (Ferber et al., 1983; Walstad and Robson, 1997). Press mentions include the following: http://www.opposingviews.com/i/gender-bias-in-college-admissions-tests, http://www.forbes.com/sites/kimelsesser/2016/07/01/is-the-college-board-making-the-sat-more-difficult-for-women/7982949d4100, and http://www.fairtest.org/sat-math-gender-bias-causes-consequences. The changes in the new SAT are described here: https://collegereadiness.collegeboard.org/sat/inside-the-test/compare-old-new-specifications.

[2] See the following examples in Belgium and India:
http://www.ugent.be/en/education/degree/practical/studentadmin/OEREnglish/multiplechoice.htm;
http://timesofindia.indiatimes.com/city/nagpur/37800-claimants-for-2800-medical-seats-highlight-positives-of-negative-marking/articleshow/47697095.cms

answers. To obtain a valid counterfactual for performance under negative marking, we conduct a field experiment in the Microeconomics undergraduate final exam of a major Spanish university. Students solve two comparable parts of a multiple-choice test: one with penalties for wrong answers and one without such penalties. Which questions carry such a penalty and which do not varies according to the exam permutations, which are randomly assigned across students. By design, it is therefore possible to recover test score distributions for both test designs, separated by gender. One can directly observe whether the representation of women and men changes at different parts of the distribution under the two testing systems. In addition to recovering women's and men's test score distributions, we shed light on the underlying mechanisms driving relative performance in the multiple-choice sections with and without penalties.

As mentioned above, risk aversion may be one important factor explaining gender differences in test performance. There is considerable empirical evidence suggesting that women are more risk averse than men (Croson & Gneezy, 2009; Eckel & Grossman, 2008). In the presence of negative marking, women may therefore answer fewer questions than optimal (in terms of maximizing their expected grade). Another factor potentially affecting answering behavior is confidence. Typically, men are found to be more confident than women (Barber & Odean, 2001). With respect to exam performance, Beyer (1999) found that students over-estimate their exam scores, and men do so more than women. If over-confidence induces males to mark too many questions, penalties for wrong answers may actually disfavor men but not women.

Ability to perform on exams ("ability" for short)[3] may be another important factor affecting relative performance under negative marking (compared to a testing system with no grade penalties for wrong answers). In particular, good students may benefit from having penalties, as they can better differentiate themselves from worse students, who benefit from lucky guesses in the part of the exam with no penalties (see Espinosa & Gardeazabal, 2010). As Fortin, Oreopoulos and Phipps (2015) show,

[3]Our definition of ability here encompasses intelligence and effort devoted to studying.

there are increasingly more women among the best students. If females are more risk averse, but also of higher ability, it is ex ante unclear whether they are harmed by penalties for wrong answers.

To analyze the effect of the above-mentioned factors on exam performance, we need to get good measures of the students' underlying characteristics. To recover students' risk aversion, we run incentivized laboratory experiments, where students play hypothetical lottery games (see Dohmen et al., 2012). Additionally, we collect data on a credible and exogenous measure of ability: the university entry grade. The university entry grade is composed of two parts: high school grades, which account for 60% of the entry grade, and a nation-wide university entry exam ("Selectividad" exam), which counts for the other 40%. The "Selectividad" exam is a national essay-type exam taken by all students who would like to study at a Spanish university after high school. The higher the score, the more likely a student is to be accepted by the best universities. As such, young people in Spain are very serious about the Selectividad exam and the scores they receive. Equipped with good measures of exogenous ability and risk aversion, we are able to investigate how these two factors shape outcomes of tests with and without penalties for wrong answers.

Our main findings are the following. First, female students do not suffer from penalties, on average. Women outperform their male colleagues on both parts of the test and do so to a greater extent on the part that carries penalties for wrong answers. Second, female students differ from male students not only in terms of risk aversion (females are relatively more risk averse) but also in their ability. Using the university entry grade as a proxy for ability, we note that females are of higher ability on average. We find that, of the two factors, only ability matters for relative performance with penalties. High-ability students perform better on both parts of the test and do so to a greater extent more on the part with negative marking. Intuitively, luck plays a larger role when there are no penalties, and low-ability students may benefit from luck to a greater extent. In contrast to ability, risk aversion has a zero effect on scores on both parts of the exam. However, this does not mean that risk aversion has no effect on answering behavior. Risk-averse students skip more questions, but also make fewer mistakes.

It seems that these two effects cancel one another out. Third, although negative marking does not harm female students on average, there are important non-linear effects. Whereas female students at the top of the ability distribution (top tercile of entry grades) clearly benefit from having penalties, female students at the bottom tend to lose out from having them. As such, whether penalties disfavor women depends crucially on their position in the ability distribution.

Our paper relates to some earlier contributions that reveal gender differences in the number of unanswered questions on math tests (women skip more than men) and suggest that this pattern is due to gender differences in risk aversion (Ramos & Lambating, 1996, Atkins et al., 1991). What is missing from these earlier studies is a valid counterfactual for performance without penalties. In a more recent contribution, Baldiga (2014) studies gender differences in the willingness to guess in a laboratory experiment. Using practice questions from SAT tests, Baldiga experimentally varies the penalty attached to wrong answers and compares the answering behavior of men and women. Baldiga finds that women are more likely to skip questions than men, and this adversely affects their test results. We also find weak evidence of gender differences in skipping under negative marking but do not replicate the negative performance effect. Differences in the gender composition of ability between the lab and the field could be one explanation for this. Students in the field experiment are a selected sample of good students, and women are among the better students. Moreover, students could prepare well for the Microeconomics exam in our field experiment. This different level of preparation could also play a role, as the subjects in laboratory experiments did not prepare for the questions in advance.[4] While closest in spirit to Baldiga (2014), our paper also relates to a recent study by Akyol et al. (2016). These authors use students' grades on the Turkish university entrance exam (a multiple-choice test with penalties for wrong answers) to study how different student characteristics (including gender) affect exam performance. They find that although women are more risk averse than

[4]See Duckworth & Seligman for evidence on gender differences in self-discipline. Moreover, monetary incentives used in labs may motivate women and men differently compared to achieving high test scores in schools (Harrison & List, 2004; Gneezy & List, 2006).

men, this has a limited impact on their final outcomes.[5] Our paper contributes to this literature by adding evidence from a field experiment in which students' answering behavior is observed when there are penalties for wrong answers and when there are no such penalties. The combination of observing student performance in a real test situation and having two different testing mechanisms within a single exam is novel. Furthermore, our measures of risk aversion and ability allow us to convincingly study how students' characteristics drive relative performance on each part of the test.

The remainder of this article is structured as follows. Section 2 provides a detailed description of the methodology used in the field experiment. In addition, we explain how we ran the incentivized experiments to uncover students' risk aversion. Section 3 presents the results, and Section 4 concludes.

2 Methodology

We ran a randomized field experiment in the undergraduate Microeconomics I class of a prestigious Spanish university. Microeconomics I is a compulsory course for students in the Economics, Business, International Business, and Law and Economics majors, and they normally take it in their freshman year. The course is offered in the third quarter (spring term). We collected the data in 2014, when nearly 600 students took the final exam. Almost one year later (winter 2015), the same group of students participated in an incentivized experiment, which was intended to gather students' risk preferences. The experiment took place as an activity in the Introduction to Game Theory course, another compulsory course for these students.

2.1 Field experiment on test performance with and without negative marking

Let us first describe the main field experiment. In the Microeconomics I course, students typically have to take two exams: the midterm exam (which takes place in May), and the final exam (which

[5] The authors do not observe individual questions, only final exam scores, and hence, do not observe question skipping behavior. Instead, they develop a structural model to recover risk aversion and the decision to guess. In a related earlier paper, Tannenbaum (2012) estimates the effect of gender differences in risk aversion on the gender gap in SAT scores, backing out students' risk aversion from answering questions with different implied risks.

takes place at the end of July). The midterm exam counts for only 20 percent of the final grade; therefore, the main analysis focuses on the final exam, which counts for the remaining 80% of the grade. Microeconomics I is taught in two large classes of Business and Economics majors (two groups of approximately 200 students each, with Business and Economics students mixed together) and two smaller classes of International Business and Law and Economics majors (approximately 100 students each, separated by major).

When designing our field experiment, we sought to maintain the same exam structure as in previous years, including the number of questions on the exam, the type of questions, and the distribution of difficult, medium and easy questions. The only difference we introduced was that half of the questions were newly exempted from grade deductions for wrong answers. As such, the final exam in 2104 had 20 multiple-choice questions, as was the case in previous years. Note that this exam is characterized by relatively little time pressure. Usually, students had enough time to finish the exam.

Key to this field experiment is the following feature: half of the questions (10 in total) had grade penalties for wrong answers, while the other half (another 10 questions) did not. Each half had an equal mix of easy, medium and difficult questions. Every question had 5 possible answers. The part with penalties was graded in the following way: 1 point for each correct answer, -0.25 points for each wrong answer, and 0 points for questions left blank. Since each question had 5 possible answers, a -0.25 grade penalty for wrong answers meant that the expected grade of a random guess was 0. The part without penalties was graded by adding 1 point for each correct answer and zero points for incorrect or blank answers. The exams clearly stated each part's grading mechanism, so that students were perfectly informed.

We prepared 4 permutations of the exam with the same 20 questions. The permutations varied in terms of which part had or did not have penalties and in the ordering of the questions, as shown in Table 1. In Permutation 1, the exam started with Questions 1 to 10, which lacked grade penalties, followed by Questions 11 to 20, which had penalties for wrong answers. Permutation 3 had the same

ordering as Permutation 1, but questions 1 to 10 had penalties, and Questions 11 to 20 did not have penalties. Permutation 2 starts with Questions 11 to 20 without penalties, followed by Questions 1 to 10 with penalties (thus, the questions with and without penalties are the same as in Permutation 3, but the order of the questions is different). Permutation 4 starts with questions 11 to 20 with penalties, followed by questions 1 to 10 without penalties (thus, the questions with and without penalties are the same as in Permutation 1, but the order of the questions is different).

— insert Table 1 about here —

We added Permutations 2 and 4 to control for the ordering of the questions. It is possible that the gender gap in performance depends on whether negative marking occurs in the first or the second part of the exam (for example, if women are slower than men, the penalties could affect them more if placed at the end of the exam). By having four permutations, we can check whether the gender gap between Permutations 1 and 3 (e.g., part I of the exam) is the same as in Permutations 2 and 4 (part II of the exam).

We sat students in alphabetical order and distributed exams such that the first students received Permutation 1, the second received Permutation 2, the third received Permutation 3, the fourth received Permutation 4, and the fifth received Permutation 1 again and so forth. To confirm that the randomization was effective, we compare students in the four groups according to their observable characteristics (gender, ability, and risk aversion). As can be seen in Appendix Table 1, students in the four groups are comparable (which is also the case if we consider females and males separately).

2.2 Experiment to elicit risk preferences

To measure risk aversion, one year later (spring 2015) we ran an incentivized laboratory experiment with the same students that participated in the field experiment. We conducted the incentivized

experiments in the seminar classes (small groups of 25-35 students) of another compulsory course, Introduction to Game Theory.[6] The experiment was organized as an activity related to the introductory chapter of the course ("Decision under Risk"). We explained to the students that they would participate in an experiment, which is a regular activity in the Introduction to Game Theory class, but that this time they could actually win real money. Each student was given a sheet of paper, describing a lottery game, which follows Dohmen et al (2012).[7] The text of the lottery game was the following:

> Imagine you won 100 euros in the Christmas lottery. Almost immediately after you collect, you receive the following financial offer from a reputable bank:
>
> You can invest a fraction of the 100 euros. With probability 0.5 you may double your investment, and with the same probability (0.5) you may lose half of the money invested.
>
> Which amount of 100 euros would you invest?

Students then had to mark whether they preferred to invest 0, 20, 40, 60, 80, or 100 euros in the lottery and turn in their sheet of paper. We also presented the lottery game on the classroom screen, read it out loud to the students and gave examples of how much they could win/loose for each amount they decided to invest. Once all sheets of paper in that seminar group were collected and placed inside a large, opaque plastic bag, the seminar teacher was asked to blindly draw a sheet of paper from the bag. The student to whom the drawn sheet belonged was asked to roll a dice and was paid out according to the result of the dice and to the amount he/she decided to invest, as marked on the sheet of paper. If the dice roll was even, the student lost half of the money invested, and if the dice roll was odd, the student received twice the invested amount.

[6] The Introduction to Game Theory classes, as is the case of most compulsory courses at this university, are divided into 4 weekly hours of lectures with the whole group and 1 and one-half weekly hours of seminars in smaller groups of approximately 25 to 35 students.

[7] Dohmen et al. (2012) mention that the lottery question has been validated in laboratory experiments and found to predict risky behavior.

3 Results

Let us begin by describing the variables. As can be seen from the summary statistics in Table 2, 547 students took the final exam. As expected, the average grade on the part with no penalties for wrong answers (6.67) is higher than in the part with penalties (5.66). Students also left more blanks in the part with penalties, and the share of correct answers is higher (73% in the part with penalties versus 67% in part without penalties). For 459 of the 547 students, we obtained their university entry grade. For a subsample of the students (390), we elicited risk preferences as part of the game theory course (we performed the lottery game in all seminar groups except the group with Law & Economics students).[8] As such, for 390 students, we have complete information on exam performance, risk preferences, and exogenous ability as proxied by the university entry grade.

— insert Table 2 about here —

Are there gender differences in exam performance? We begin with a graphical representation of the students' performance on the two parts of the final exam (with and without penalties). As shown above, grades are higher in the part without grade penalties for wrong answers. Therefore, and given that grading on a curve is common, it is also informative to examine students' positions in the distribution of scores (rank). To do so, we divide the score distribution into deciles (separately for each of the two parts) and create a variable "Difference Rank", which measures the rank on the part with penalties minus the rank on the part without penalties. Students who relatively gained from penalties have a positive value, whereas students who lost have a negative value.

The upper part of Figure 1 plots the difference in ranks for female and male students, and the lower part plots the difference in absolute scores. As can be seen in the graphs, females typically

[8] Note that in the year 2014, 90 students were in the Law & Economics dual degree. Therefore, attrition between first and second year is another reason that we lost some observations relative to the initial sample of 547 students.

improve their ranks more than males when there are penalties (the exception are a couple of men, who increased their ranks by a fairly large amount).

— insert Figure 1 about here —

Table 3 presents a formal analysis of the ideas expressed in the previous graph. Regardless of whether we consider percentile ranks or overall scores, female students perform better on both testing sections (with and without negative marking) but do so to a greater extent on the part with negative marking. The differences in ranks and scores between the exam sections with penalties and without penalties are larger for women than for men (see columns 3 and 7). If men were more prevalent at the bottom end of the score distribution in the no-penalty section, then it might be easier for them to improve their scores with penalties in a mechanical way.[9] As can be seen from Appendix Figure 1, it is indeed the case that men are more likely to be at the bottom of the distribution. Therefore, a sensible extension is to weight observations (separately for each score on the part without penalties), such that the proportion of female and male students is held constant.[10] When we do this, the estimated coefficients for the gender dummies increase (see columns 4 and 8), as is to be expected.

— insert Table 3 about here —

We mention two factors that could plausibly shape differences in test outcomes between the sections with and without penalties: risk aversion and ability. Whereas high-ability students may perform relatively better under negative marking (because it allows them to better distinguish themselves from the less-able students), risk aversion may lead to worse relative performance, especially if it leads to excessive question skipping. How different are women and men in these two dimensions?

[9] Assume that in the extreme case, a man has a 0 score on the no-penalty section, whereas a woman has a 10. Clearly, men can only improve their scores when penalties are imposed, whereas women can only be harmed.

[10] If, for example, 17 female students and 10 males received a grade of 10 on the no-penalty part, we set the weight for females equal to 1 and the weight for males equal to 1.7 (=17/10).

As can be seen from Table 4, females have a higher university entry grade than males and invest less in the hypothetical lottery game.

— insert Table 4 about here —

Given that female and male students differ in both ability and risk aversion, we now investigate the role of each of these factors in shaping test outcomes in the parts with and without negative marking. Table 5, columns 1, 4 and 7 replicate the results of Table 3, showing that female students perform better on both test sections but do so to a greater extent on the part with penalties for wrong answers.

Can this result be explained by differences in ability? Let us first investigate the relationship between ability and testing outcomes per se. Clearly, being in higher ability terciles (the omitted group is the bottom group) increases performance in general but does so to a greater extent under negative marking (as can be seen in column 8, this difference is statistically significant). As females have higher university entry grades, controlling for this exogenous measure of ability should decrease the estimated effect of the female dummy, which indeed happens. In fact, it roughly halves the estimated gender coefficient.[11]

— insert Table 5 about here —

Given that ability matters for relative performance under negative marking, we would like to investigate whether this effect is similar for women and men. Figure 2 shows the difference between the rank on the exam sections with and without penalties for wrong answers (a positive value implies a higher rank on the section with penalties) for both sexes, depending on their entry grade. Interestingly, the relationship between entry grade and relative performance under negative marking is stronger for women than for men.

[11] Note that this result does not depend on the fact that we divide the entry grade into terciles. If we use quintiles instead, the results are similar (see Appendix Table 2).

— insert Figure 2 about here —

Table 6 confirms the graphical results. The first column shows that students in the top tercile of entry scores perform better in the test section with grade penalties for wrong answers and do so to a greater extent if they are female. Since this gender difference among students with top entry grades is absent in the no-penalty section (see column 2), it comes as no surprise that women in the top tercile of ability improve their rank relative to men when there are grade penalties (see column 3). For students in the second tercile, gender does not matter for the performance difference between the questions with and without penalties. For students in the lowest tercile, women lose rank relative to men when exposed to negative marking, but the effect is not statistically significant. As such, whether women are harmed by grade penalties depends on their position in the ability distribution. High-ability female students actually benefit from penalties for wrong answers, whereas low-ability women tend to be harmed by them.

— insert Table 6 about here —

What about risk aversion? We find that it does not have a significant effect on overall exam performance. As can be seen in Table 7, the effect of risk aversion (or risk lovingness) on the rank in the tests is zero in the no-penalty section (which is to be expected), but the effect is also zero and in the section with penalties. This last result is surprising at first, as risk aversion has been found to affect the number of questions left unanswered, and excessive skipping of questions could impair test performance. To better understand why risk aversion does not affect scores in the section with negative marking, we examine answering behavior in greater detail. In particular, we investigate whether there is a relationship between the hypothetical lottery investment decisions and the number of unanswered questions (blanks), the number of correct answers and the number of incorrect answers.

— insert Table 7 about here —

Figure 3 shows that there is a relationship between investment decisions (taken one year after the exam) and answering behavior on the exam. For example, in the upper graph, it can be seen that the probability of leaving a low number of blanks (e.g. 2) is typically highest for the most risk-loving subjects (who chose to invest 100), followed by students with intermediate investment levels, and lowest for the students who chose to invest 0 in the hypothetical lottery. Exactly the opposite is true for the number of incorrect answers. Here, more risk-averse individuals have a higher probability of having fewer than a certain number of incorrect answers, and risk-loving individuals have the lowest probability of making few mistakes. For the number of correct answers, we do not see such a clear pattern (see Appendix Figure 2).

— insert Figure 3 about here —

Table 8 presents the estimates of the effect of risk lovingness on the number of answers left blank, the number of correct answers and the number of incorrect answers. The estimates show that risk-loving students (i.e., those who invest more in the lottery game) leave fewer blank answers (see column 1) but make more mistakes (see column 9). This may explain why risk aversion has no overall effect on performance under negative marking. Since women are more risk averse than men, do they differ in their answering behavior? Columns 2, 6, and 10 show that women have a higher share of correct answers and a lower share of incorrect answers. We find no gender effect on the number of questions left blank. This last result is surprising but could be driven by the confounding factor ability. Indeed, when controlling for ability, women tend to skip more (although the coefficient estimate is not statistically significant) and have a lower share of incorrect answers. When controlling for risk aversion and ability, gender differences in skipping entirely disappear (column 4). However, female students tend to have a higher share of correct answers (column 12).

— insert Table 8 about here —

In summary, risk aversion does not affect exam performance (in either part, with or without penalties). Although risk-averse individuals tend to skip more, they also make fewer mistakes. Therefore, even though women are more risk averse, they are not necessarily disadvantaged if the test carries a grade penalty for wrong answers. In contrast to risk aversion, ability matters significantly for relative performance under negative marking and more for women than for men. Top female students improve their ranks if there are grade penalties for wrong answers, but the same is not true for female students in the bottom of the university entry score distribution. Therefore, whether women or men suffer more from penalties for wrong answers in multiple-choice tests depends on their position in the distribution of ability.

4 Discussion

This paper documents how the interaction between gender and ability affects relative performance under negative marking. How general are our results? First, on the exam side, note that the Microeconomics final exam is an important exam for the students, as it has always been one of the most selective exams. We would therefore describe this exam as a "high-stakes" exam. Nevertheless, the stakes of a university entry exam are possibly higher. As an interesting avenue for research, we suggest exploring gender differences in the reaction to penalties in tests with different stakes (see Azmat, Calsamiglia and Iriberri, 2016, for gender differences in response to high stakes). It is possible that risk aversion matters more if the stakes are very high, and this may affect women's performance more (positively or negatively). On the student side, it is clear that the sample of students is a very selected one. The Spanish university we considered attracts very good students in Spain. Female students may therefore be unusually self-confident and less threatened by the existence of penalties. Again, it seems worthwhile to replicate our experiment with other student samples to obtain greater variance

in ability levels. To do so, one would need an exogenous measure of ability similar to ours. Finally, students could prepare quite well for the exam used in our field experiment. For instance, we posted practice questions on the web, which allowed students to have a good grasp of the type of questions that could appear on the exam. Whether penalties affect females differently in exams that are more difficult to prepare for (and for which more thinking on the spot is required) must be left for future research.

References

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

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

[3] Azmat, G., C. Calsamiglia, and N. Iriberri (2016). Gender Differences in Response to Big Stakes. Journal of the European Economic Association, forthcoming.

[4] Baldiga, K. (2014). Gender Differences in the Willingness to Guess. Management Science, 60(2): 434-448.

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

[6] Bertrand M., Hallock K. (2001): The gender gap in top corporate jobs. Industrial Labor Relations Review 55:321.

[7] Beyer, S. (1999). Gender differences in the accuracy of grade expectancies and evaluations. Sex Roles, Vol. 41, No. 314, pp. 279 .296.

[8] Croson R. Gneezy U. (2009). Gender differences in preferences. Journal of Economic Literature, 47(2):448474.

[9] Dohmen, T., Falk. A., Huffman, D. Sunde, U. (2012). The Intergenerational Transmission of Risk and Trust Attitudes, Review of Economic Studies, 79(2): 645677.

[10] Duckworth, A.L. Seligman, M.E. (2006). Self-Discipline Gives Girls the Edge: Gender in Self-Discipline, Grades, and Achievement Test Scores, Journal of Educational Psychology, 98(1): 198208.

[11] Eckel C. Grossman P. (2008). Forecasting risk attitudes: An experimental study using actual and forecast gamble choices. Journal of Economic Behavior and Organization, 68(1):117.

[12] Espinosa M.P. Gardeazabal J. (2010). Optimal correction for guessing in multiple-choice tests. Journal of Mathematical Psychology, 54(4):41525.

[13] Ferber, M.A., B.G. Birnbaum, and C.A. Green (1983). Gender Differences in Economic Knowledge: A Re-evaluation of the Evidence. Journal of Economic Education, 14, pp. 24 - 37.

[14] Fortin, N.M., Oreopoulos P. and S. Phipps (2015). Leaving Boys Behind. Gender Disparities in High Academic Achievement. Journal of Human Resources, 50(3), pp. 549 - 79.

[15] Gneezy, U. List J. (2006). Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments, Econometrica, 74(5): 1365-1384.

[16] Harrison, G. List J. (2004). Field Experiments, Journal of Economic Literature, XLII, 1013-1059.

[17] Leonard D.K. Jiang J. (1999): Gender Bias and the college predictions of the SATS: A cry of despair. Research in higher education, 40(4): 375-407.

[18] List, J. (2006). The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions, Journal of Political Economy, 114(1): 1-37.

[19] ONeill J (2003): The gender gap in wages, circa 2000. American Economic Review, 93(2):309314.

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

[21] 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.

[22] Walstad, W. and D. Robson. (1997). Differential Item Functioning and Male-Female Differences on Multiple-Choice Tests in Economics. The Journal of Economic Education, 28, 2, pp. 155-171.

TABLE 1
Design Experiment

Permutation 1	Permutation 2	Permutation 3	Permutation 4
Q1-10: No Penalties	Q11-20: No Penalties	Q1-10: Penalties	Q11-20: Penalties
Q11-20: Penalties	Q1-10: Penalties	Q11-20: No Penalties	Q1-10: No Penalties

Notes: Penalties means grade deductions for wrong answers. No Penalties implies 0 points for wrong answers. In Permutations 2 - 4, Q1-10 refer to questions 1-10 in Permutation 1, and Q11-20 refer to questions 11-20 in Permutation 1.

TABLE 2

Summary Statistics

	Obs.	Mean	Std. Dev.	Min	Max
	Exam Section "No Penalties"				
Grade	547	6,67	1,84	1	10
Rank	547	4,83	2,75	1	10
Correct Answers	547	6,67	1,84	1	10
Incorrect Answers	547	3,28	1,84	0	9
Blank	547	0,05	0,29	0	4
	Exam Section "Penalties"				
Grade	547	5,66	2,15	-2.25	10
Rank	547	5,35	2,81	1	10
Correct Answers	547	6,22	1,89	0	10
Incorrect Answers	547	2,22	1,44	0	9
Blank	547	1,56	1,42	0	7
	Student Characteristics				
Entry Grade	459	11,25	1,05	5,63	13,65
Investment	390	73,28	29,30	0	100

Notes: The table presents summary statistics for the students who take the final exam. The first two blocks measure grade, rank, the number of correct, incorrect and blank answers, separately for the section with and without penalties for wrong answers. The last two rows display summary statistics for the measure of risk-aversion (Investment), and the university entry grade.

TABLE 3

Performance in the part with Penalties (P) and No Penalties (NP)

	(1) Rank-P	(2) Rank-NP	(3) Diff Rank	(4) Diff Rank	(5) Score-P	(6) Score-NP	(7) Diff Score	(8) Diff Score
Female	0.682***	0.481**	0.201	0.433	0.548***	0.360**	0.188	0.327*
	(0.240)	(0.233)	(0.265)	(0.275)	(0.184)	(0.157)	(0.187)	(0.188)
Dummy Order Questions	YES	YES	YES	YES	YES	YES	YES	YES
Dummy Part with Penalties	YES	YES	YES	YES	YES	YES	YES	YES
Weights	NO	NO	NO	YES	NO	NO	NO	YES
Observations	547	547	547	547	547	547	547	547
R-squared	0.016	0.024	0.022	0.032	0.020	0.024	0.027	0.038

Notes: Dependent variable is the decile in the score distribution (columns 1-4) and overall score (columns 5-8). P denotes "Part with Penalties", and NP denotes "Part with No Penalties". Diff Rank is defined as Rank-Penalty Section minus Rank-No Penalty Section (and Diff Score is the respective difference for overall scores). Dummy Order Questions takes a value of 1 if the first part of the exam has no penalties for wrong answers. Dummy Part with Penalties takes a value of 1 if the Questions 1-10 (of Permutation 1) are not penalized. In columns 4 and 8, frequency weights are used to balance the proportion of males and females in each cell of the score in the No-Penalty section. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

TABLE 4

Gender Differences in Ability and Risk-Aversion

	(1) Entry Grade	(2) Invest
Female	0.540***	-20.01***
	(0.0957)	(2.770)
Constant	10.96***	83.96***
	(0.0727)	(1.866)
Observations	459	390
R-squared	0.066	0.116

Notes: Dependent Variable is the entry grade in column 1 and the investment (from 0 to 100) in column 2. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

TABLE 5

Rank and Ability-Type

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Rank Part with Penalties			Rank No Penalties			Differences Rank		
Female	0.682***		0.404	0.481**		0.298	0.201		0.107
	(0.240)		(0.256)	(0.233)		(0.255)	(0.265)		(0.291)
Entry Grade: Middle		0.904***	0.827***		0.926***	0.869***		-0.0217	-0.0420
		(0.312)	(0.316)		(0.296)	(0.298)		(0.340)	(0.347)
Entry Grade: Top		2.495***	2.393***		1.605***	1.530***		0.890***	0.863**
		(0.298)	(0.306)		(0.308)	(0.316)		(0.336)	(0.346)
Dummy Order Questions	YES	YES	YES	YES	YES	YES	YES	YES	YES
Dummy Part with Penalties	YES	YES	YES	YES	YES	YES	YES	YES	YES
Constant	5.034***	4.278***	4.127***	4.315***	3.654***	3.543***	0.719***	0.624**	0.584*
	(0.242)	(0.286)	(0.297)	(0.235)	(0.261)	(0.278)	(0.267)	(0.293)	(0.312)
Observations	547	459	459	547	459	459	547	459	459
R-squared	0.016	0.137	0.142	0.024	0.078	0.081	0.022	0.059	0.060

Notes: Dependent Variable is the Decile in the Score Distribution. Female is a gender dummy and Entry Grade Middle/Top the second highest and highest tercile in the entry grade. Dummy Order Questions takes a value of 1 if the first part of the exam has no penalties for wrong answers. Dummy Part with Penalties takes a value of 1 if the Questions 1-10 (of Permutation 1) are not penalized. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

TABLE 6

Rank and Ability, interacted with Gender

	(1) Rank P	(2) Rank NP	(3) Rank P-NP
Female	-0.0671	0.255	-0.323
	(0.444)	(0.424)	(0.470)
Entry Grade: Middle	0.752*	0.734*	0.0171
	(0.444)	(0.404)	(0.506)
Entry Grade: Top	1.784***	1.623***	0.161
	(0.449)	(0.464)	(0.522)
Female X Entry Grade Middle	0.288	0.248	0.0400
	(0.634)	(0.600)	(0.697)
Female X Entry Grade Top	1.146*	-0.131	1.277*
	(0.612)	(0.640)	(0.701)
Dummy Order Questions	YES	YES	YES
Dummy Part with Penalties	YES	YES	YES
Constant	4.322***	3.557***	0.765**
	(0.328)	(0.302)	(0.342)
Observations	459	459	459
R-squared	0.148	0.081	0.069

Notes: Dependent Variable is the Decile in the Score Distribution. Female is a gender dummy and Entry Grade Middle/Top the second highest and highest tercile in the entry grade. Dummy Order Questions takes a value of 1 if the first part of the exam has no penalties for wrong answers. Dummy Part with Penalties takes a value of 1 if the Questions 1-10 (of Permutation 1) are not penalized. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

TABLE 7

Rank and Risk-Aversion, interacted with Gender

	(1) Rank P	(2) Rank NP	(3) Rank P-NP	(4) Rank P	(5) Rank NP	(6) Rank P-NP
Female	0.730**	0.628**	0.103	1.116	0.816	0.300
	(0.298)	(0.291)	(0.323)	(0.761)	(0.783)	(0.829)
Investment	-0.00162	0.00278	-0.00440	0.00147	0.00429	-0.00282
	(0.00494)	(0.00492)	(0.00509)	(0.00713)	(0.00744)	(0.00808)
Female X Investment				-0.00506	-0.00247	-0.00259
				(0.00982)	(0.00982)	(0.0104)
Dummy Order Questions	YES	YES	YES	YES	YES	YES
Dummy Part with Penalties	YES	YES	YES	YES	YES	YES
Constant	5.082***	3.807***	1.275**	4.821***	3.679***	1.142
	(0.493)	(0.507)	(0.507)	(0.625)	(0.674)	(0.717)
Observations	390	390	390	390	390	390
R-squared	0.027	0.035	0.054	0.028	0.035	0.054

Notes: Dependent Variable is the Decile in the Score Distribution. Female is a gender dummy and Investment is the amount invested in the hypothetical lottery game (a larger investment implies more risk-lovingness). Dummy Order Questions takes a value of 1 if the first part of the exam has no penalties for wrong answers. Dummy Part with Penalties takes a value of 1 if the Questions 1-10 (of Permutation 1) are not penalized. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

TABLE 8

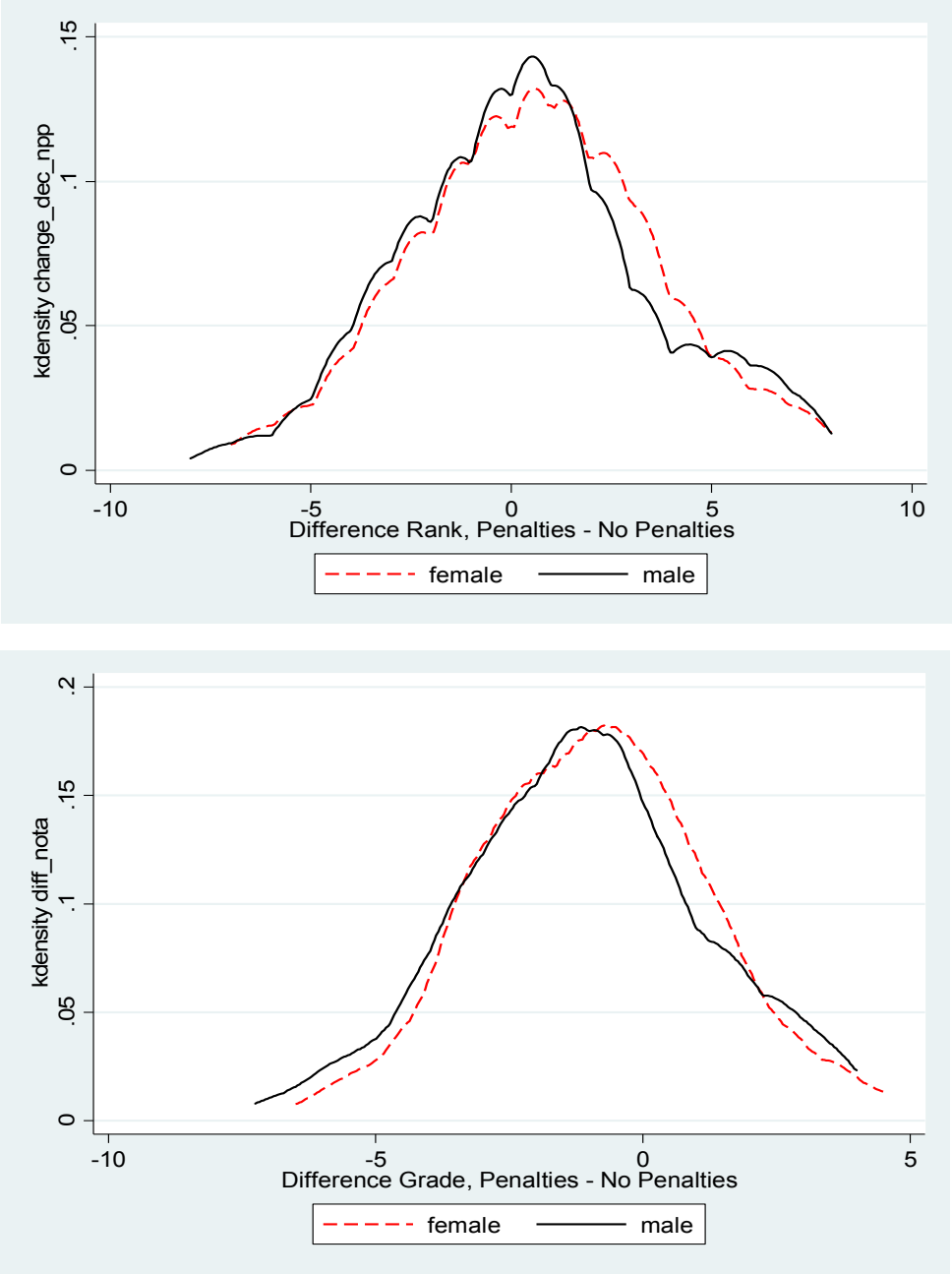
Answering Behavior and Risk Aversion

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Blanks Punishment				Correct Punishment				Incorrect Punishment			
Invest	-0.00492**			-0.00684***	-0.00164			0.00442	0.00655***			0.00242
	(0.00250)			(0.00255)	(0.00326)			(0.00325)	(0.00241)			(0.00262)
Female		-0.00346	0.0931	-0.0325		0.472**	0.204	0.285		-0.468***	-0.297**	-0.253*
		(0.150)	(0.151)	(0.155)		(0.196)	(0.193)	(0.200)		(0.145)	(0.142)	(0.152)
Entry Grade: Middle			-0.203	-0.250			0.503**	0.533**			-0.299*	-0.283
			(0.177)	(0.177)			(0.224)	(0.228)			(0.171)	(0.173)
Entry Grade: Top			-0.603***	-0.669***			1.698***	1.740***			-1.094***	-1.071***
			(0.191)	(0.190)			(0.236)	(0.239)			(0.175)	(0.181)
Dummy Order Questions	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Dummy Part with Penalties	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Constant	1.971***	1.600***	1.797***	2.417***	6.360***	5.990***	5.455***	5.054***	1.669***	2.411***	2.748***	2.529***
	(0.224)	(0.156)	(0.179)	(0.282)	(0.301)	(0.198)	(0.220)	(0.380)	(0.221)	(0.145)	(0.172)	(0.312)
Observations	390	390	390	390	390	390	390	390	390	390	390	390
R-squared	0.011	0.001	0.027	0.043	0.015	0.029	0.147	0.151	0.037	0.045	0.134	0.136

Notes: Dependent variable in (1)-(4) is the number of blank answers in the part with penalties. Dependent variable in (5)-(8) are the number of correct answers in the part with penalties, and in (9) to (12) the number of incorrect answers. See also notes to previous tables. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

FIGURE 1

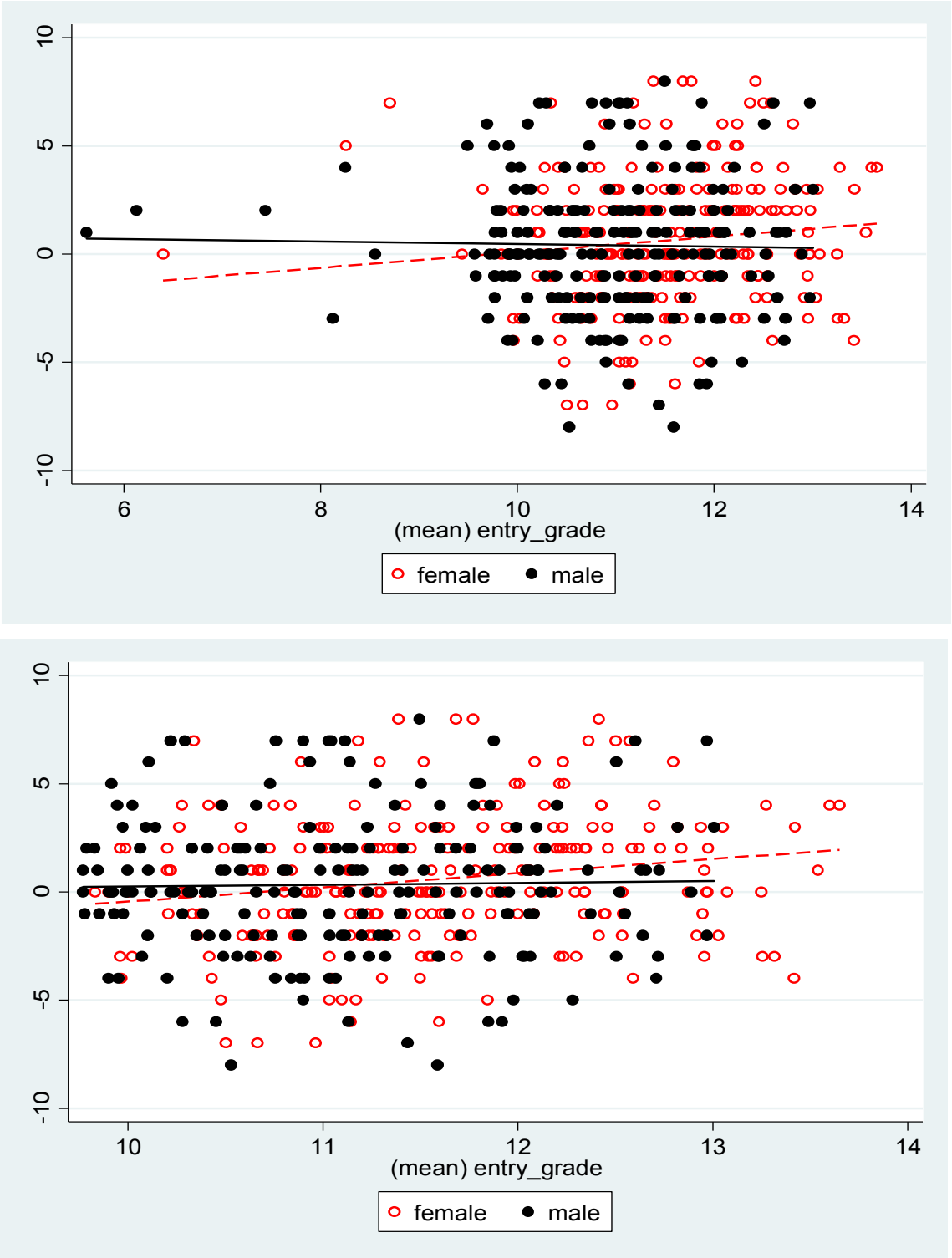
Rank Decile, Part with Penalties versus Part without Penalties



Notes: Graph 1 displays the change in the ranks (deciles scores), exam section with penalties versus exam section without penalties (upper picture), and changes in scores (penalties versus no penalties - picture at the bottom).

FIGURE 2

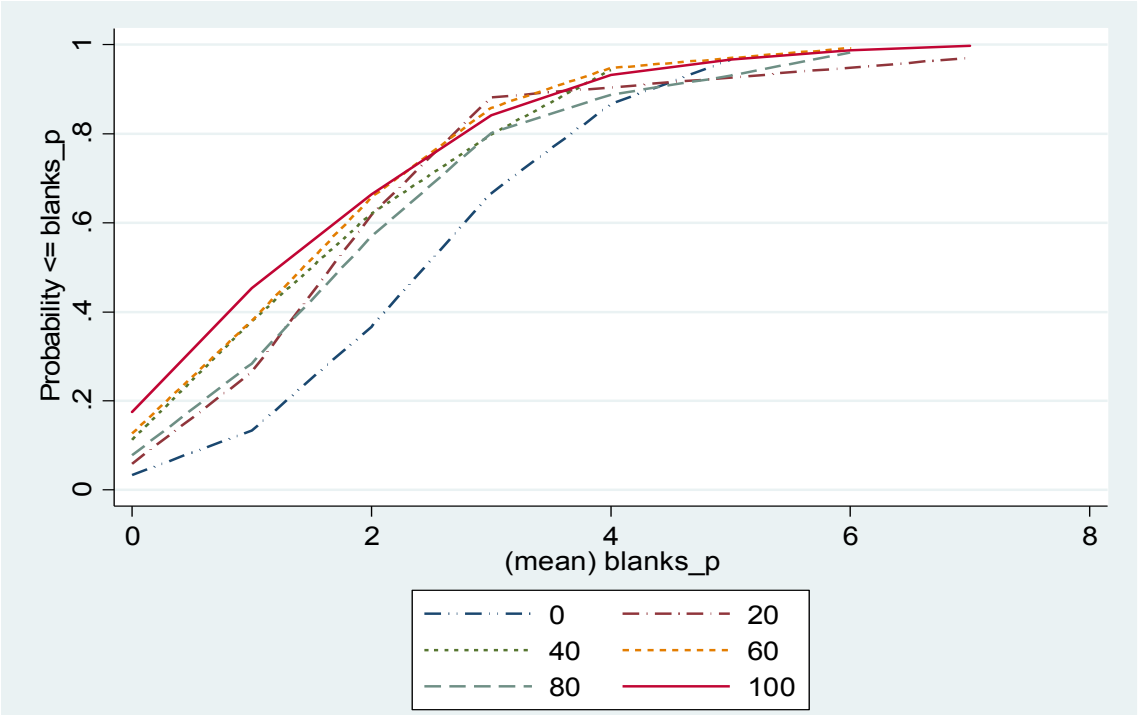
Rank Decile and Ability, Part with Penalties minus Part with No Penalties

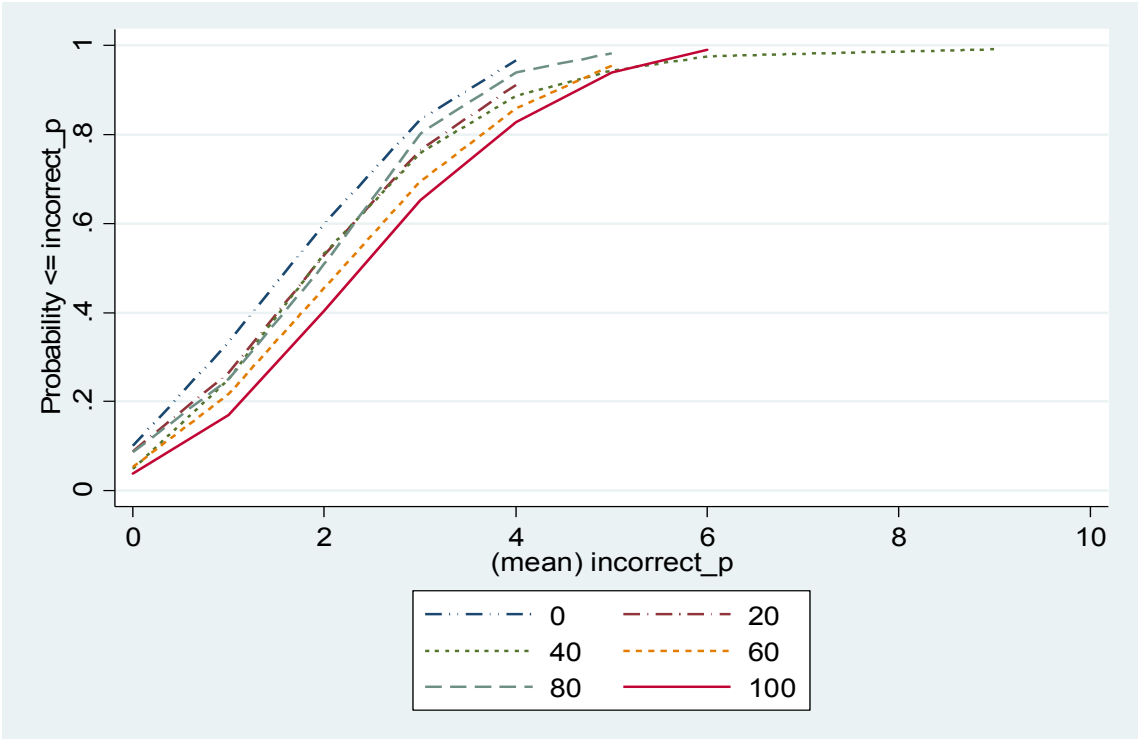


Notes: the x-axis shows the entry grade of the student, and the y-axis the difference in ranks (Penalties - No Penalties). The lower picture discards the bottom 5% entry grade students.

FIGURE 3

Investment and Blank Answers, Investment and Incorrect Answers





APPENDIX TABLE 1

Randomization Check

	(1)	(2)	(3)	(4)	(5)	(6)
	Entry Grade University			Risk Aversion		
Permutation 1	11.30***	11.55***	10.99***	70.50***	63.51***	79.55***
	(0.0993)	(0.137)	(0.134)	(3.052)	(4.372)	(3.762)
Permutation 2	11.23***	11.52***	10.93***	71.92***	57.92***	85.10***
	(0.0949)	(0.119)	(0.140)	(2.871)	(3.774)	(3.404)
Permutation 3	11.07***	11.41***	10.66***	75.45***	66.67***	86***
	(0.113)	(0.132)	(0.174)	(3.338)	(4.505)	(4.477)
Permutation 4	11.39***	11.53***	11.23***	75.49***	67.27***	85.11***
	(0.0864)	(0.113)	(0.131)	(2.658)	(3.611)	(3.467)
Student Sample	All	Female	Male	All	Female	Male
Observations	459	244	215	390	208	182
R-squared	0.991	0.993	0.991	0.863	0.828	0.919
H0: P1=P2=P3=P4 (P-Value)	0.1626	0.8781	0.0723	0.5387	0.2913	0.6147

Notes: Dependent variables are the university entry grade (columns 1-3), and a measure of risk-aversion (columns 4-6). Permutations 1-4 are dummy variables indicating the Permutation of the final exam, the student received. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

APPENDIX TABLE 2

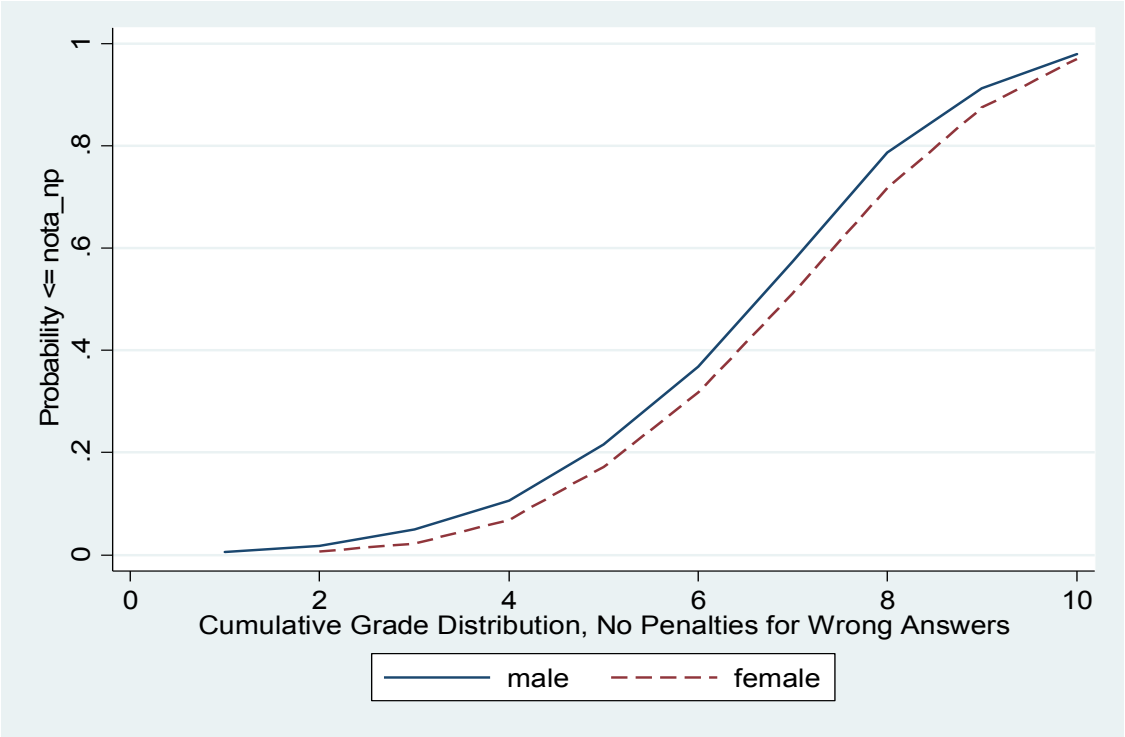
Rank and Ability-Type, Quintiles

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Rank Part with Penalties			Rank Part No-Penalties			Differences Rank		
Female	0.682***		0.326	0.481**		0.208	0.201		0.118
	(0.240)		(0.266)	(0.233)		(0.258)	(0.265)		(0.294)
Entry Grade: Second Quintile		0.0114	-0.0402		0.880**	0.847**		-0.868**	-0.887**
		(0.398)	(0.404)		(0.379)	(0.379)		(0.426)	(0.429)
Entry Grade: Third Quintile		0.811**	0.731*		1.216***	1.165***		-0.405	-0.434
		(0.410)	(0.416)		(0.384)	(0.385)		(0.424)	(0.432)
Entry Grade: Fourth Quintile		1.441***	1.372***		1.592***	1.548***		-0.151	-0.176
		(0.401)	(0.406)		(0.385)	(0.388)		(0.425)	(0.431)
Entry Grade: Top Quintile		2.643***	2.505***		2.098***	2.011***		0.544	0.494
		(0.393)	(0.417)		(0.395)	(0.410)		(0.425)	(0.444)
Dummy Order Questions	YES	YES	YES	YES	YES	YES	YES	YES	YES
Dummy Part with Penalties	YES	YES	YES	YES	YES	YES	YES	YES	YES
Constant	5.034***	4.465***	4.361***	4.315***	3.335***	3.269***	0.719***	1.130***	1.092***
	(0.242)	(0.356)	(0.360)	(0.235)	(0.319)	(0.334)	(0.267)	(0.341)	(0.355)
Observations	547	459	459	547	459	459	547	459	459
R-squared	0.016	0.127	0.130	0.024	0.087	0.088	0.022	0.063	0.064

Notes: Dependent Variable is the Decile in the Score Distribution. Female is a gender dummy and entry grade quintiles are the quintiles in entry score (omitted group: lowest quintile). Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

APPENDIX FIGURE 1

Cumulative Grade Distribution, No Penalty-Part



APPENDIX FIGURE 2

Correct Answers and Risk-Aversion