Labour Economics 64 (2020) 101815

Contents lists available at ScienceDirect

Labour Economics

journal homepage: www.elsevier.com/locate/labeco

Gender, competitiveness, and task difficulty: Evidence from the field

Britta Hoyer[a], Thomas van Huizen[b,*], Linda Keijzer[b], Sarah Rezaei[b], Stephanie Rosenkranz[b], Bastian Westbrock[c]

[a] *Faculty of Business Administration and Economics, Paderborn University, Warburger Str. 100, Paderborn 33098, Germany*
[b] *Utrecht School of Economics, Utrecht University, Kriekenpitplein 21–22, EC Utrecht 3584, the Netherlands*
[c] *Faculty of Business Administration and Economics, Fernuniversität Hagen, Universitätsstr. 47, Hagen 58097, Germany*

ARTICLE INFO

JEL classification:
C78
D47
D82

Keywords:
Gender gap
Competitiveness
Task difficulty

ABSTRACT

This study examines the gender gap in competitiveness in an educational setting and tests whether this gap depends on the difficulty of the task at hand. For this purpose, we administered a series of experiments during the final exam of a university course. We confronted three cohorts of undergraduate students with a set of bonus questions and the choice between an absolute and a tournament grading scheme for these questions. To test the moderating impact of task difficulty, we (randomly) varied the difficulty of the questions between treatment groups. We find that, on average, women are significantly less likely to select the tournament scheme. However, the results show that the gender gap in tournament entry is sizable when the questions are relatively easy, but much smaller and statistically insignificant when the questions are difficult.

1. Introduction

Despite overall convergence over the past decades, substantial differences between men and women in educational and labor market outcomes remain. For example, while the gender gap in college enrollment and graduation has reversed (Goldin et al., 2006), women are still less likely than men to enter more competitive and more rewarding MBA and STEM programs (Buser et al., 2014; Jurajda and Münich, 2011). In addition, there still is a significant gender pay gap (Blau and Kahn, 2017; Petrongolo, 2019). Women are in particular heavily underrepresented in high earning jobs at the top of the corporate ladder (Cook and Glass, 2014; Fortin et al., 2017). In US Fortune 500 companies, for example, only around one out of five board seats is held by a woman (Deloitte, 2019). Likewise, in the largest publicly listed companies in the European Union, about one quarter of the board members is female (European Commission, 2018). To narrow this gap, several countries (e.g. Norway, France, Germany, Italy) have introduced gender quotas on corporate boards (Bertrand et al., 2018).

An explanation for the pervasive gender differences in educational choice and labour market outcomes can be found in (innate) differences between men and women in their willingness to compete (Buser et al., 2014; Kamas and Preston, 2018; Reuben et al., 2017).[1] A large body of experimental evidence indeed confirms that men are more competitive than women. In their seminal paper, Niederle and Vesterlund (2007) find that men opt more than twice as often as women for a competitive tournament as their preferred compensation scheme for a series of stylized, homogeneous tasks. Their experiment has been replicated many times (Azmat and Petrongolo, 2014; Niederle and Vesterlund, 2011). However, in the recent literature a more nuanced picture emerges, where one of the organizing principles seems to be that the gender differences in competitiveness depend on the nature of the task at hand.[2] Men prefer competitive reward schemes more often than women for arithmetic, ball-tossing, and maze-solving tasks (Almås et al., 2015; Balafoutas and Sutter, 2012; Booth and Nolen, 2012; Buser et al., 2014; Healy and Pate, 2011; Niederle et al., 2013; Saccardo et al., 2017), while the gender gap is often not found, or sometimes even reversed, for verbal tasks (e.g. Dreber et al., 2014; Große and Riener, 2010; Günther et al., 2010; Shurchkov, 2012). In other words, men seem to compete more often in stereotypical male tasks, whereas women are at least equally competitive in stereotypical female tasks.

In this paper, we study the impact of two other task dimensions on the gender gap in reward scheme choices. First, we test whether the

* Corresponding author.
E-mail address: t.m.vanhuizen@uu.nl (T. van Huizen).

[1] There are of course multiple (complementary) explanations for these gender differences. For example, Bredemeier (2019) points out that the gender gap in inter-firm mobility (due to different gender roles within the household) plays an important role.

[2] In addition to the type of task, prior research suggests that culture (Booth et al., 2019; Gneezy et al., 2009; Zhang, 2019), social learning (Booth and Nolen, 2012) and the gender of the opponent (Datta Gupta et al., 2013) also play an important role explaining the gender difference in competitiveness. This evidence suggests that these gender differences are not innate. Empirical evidence from professional distance running is also inconsistent with the hypothesis that these gender differences are due to biologically evolved predispositions (Frick, 2011).

https://doi.org/10.1016/j.labeco.2020.101815
Received 31 October 2019; Received in revised form 15 February 2020; Accepted 7 March 2020
Available online 15 March 2020
0927-5371/© 2020 Elsevier B.V. All rights reserved.

gender gap extends beyond the lab by looking at a real-life educational setting. The typical lab experiment involves relatively simple gaming tasks, such as completing a basic arithmetic problem, solving a maze, or tossing a ball in a bucket.[3] It is not obvious that the finding from the lab — that men have a stronger preference for competitive reward schemes than women — can be replicated in the field with tasks that are familiar to the subjects and part of their daily lives. A growing body of research on confidence indicates that the familiarity with a task and the familiarity with the peers allow people to base their decisions on a more accurate prior of their relative standing in a group (Alicke et al., 1995; Benoît and Dubra, 2011; Lavy, 2013; Perloff and Fetzer, 1986). In the lab, where tasks and competitors are unfamiliar, subjects may be overconfident. If this applies mainly to men, the gender gap in competitiveness is likely to be smaller in the field than in the lab.

Second, we examine whether the gender gap in competitiveness depends on the difficulty of the task at hand. As a large body of confidence research shows that task difficulty influences people's perception of their own task performance and their relative performance in comparison to others in particular (Benoît et al., 2015; Kruger, 1999; Moore and Cain, 2007; Windschitl et al., 2003), task difficulty is likely to have an impact on the willingness to join a competition. Because self-assessments (How good am I?) have been shown to have a greater impact than other-assessments (How good are others?) (Kruger, 1999), people tend to overestimate their relative performance in easy tasks and underestimate their relative performance in difficult tasks (Hoelzl and Rustichini, 2005). We therefore expect that the preference for a competitive reward scheme diminishes in the difficulty of the underlying task.

However, men and women may not respond equally to variations in task difficulty. This follows from another group of confidence studies (Buser, 2016; Mobius et al., 2011; Roberts and Nolen-Hoeksema, 1989) suggesting that men and women differ in the extent to which they place weight on different self-assessment criteria. While men place more weight on their self-assessed performance in an ongoing task, women tend to be more conservative in their belief updating and rather rely on (external) assessments of their past performance in similar tasks. This leads to the prediction that men are more likely to overestimate their relative performance in an easy task and underestimate their performance in a difficult task, implying that the negative impact of an increase in task difficulty on the willingness to compete is stronger for men.

We provide evidence from three framed field experiments (Harrison and List, 2004), which we conducted during the final exams of a compulsory course in an undergraduate business and economics program between 2015 and 2017. In these experiments, we confronted students (in total 681) with a set of bonus questions and the choice to earn grade points for these questions either according to an absolute grading scheme, that is, the equivalent of a piece rate scheme, or according to a tournament scheme. To test the role of task difficulty, we varied the level of the bonus questions between treatments. In particular, in Experiment 1 (the 2015 cohort) we provided a set of relatively easy bonus questions. The difficulty of the questions was significantly increased in Experiment 2 (the 2016 cohort). Finally, in Experiment 3 (the 2017 cohort) we randomly allocated students from the same cohort to the easy and difficult questions.

Our findings largely confirm our hypotheses. Consistent with most of the literature on confidence, we find that the students are more likely to shy away from competition under the difficult-task treatments. Furthermore, in the easy-task treatments, men chose the tournament compensation scheme significantly more often than women. In contrast, in the difficult-task treatments the gender difference was small and insignificant. Hence, the result from the lab that men are more likely to opt for a competitive reward scheme has been replicated in our exam setting, but only under the condition that the task was relatively easy. Task difficulty significantly diminished the tournament-entry choices of everyone so that, on average, students were overconfident about their performance in the easy-task treatments: there is no evidence of overconfidence in the difficult-task treatments. Most importantly, men were more affected by the increased difficulty of the questions so that men and women appeared almost equally confident in the difficult-task treatments.

Our study contributes to the literature on gender differences in two important ways. First, it is one of a few experimental studies to explore gender differences in the willingness to compete in a real-life setting. There is a large body of evidence from the field that investigates a related hypothesis, namely that women perform worse than men under competitive pressures (Czibor et al., 2014; De Paola et al., 2015; Iriberri and Rey-Biel, 2019; Jurajda and Münich, 2011; Ors et al., 2013; Pekkarinen, 2015).[4] However, these studies focus on the gender gap in performance under competition rather than on the gender gap in the willingness to compete. Flory et al. (2014) and Samek (2019) are two noteworthy exceptions. They test for gender differences in the willingness to compete in a labor market setting and find that women are significantly less likely to apply for jobs with a more competitive compensation regime. In contrast, we test for gender differences in competitiveness in an exam setting.[5] While the exam setting was also used by a number of field studies in the confidence literature (Bengtsson et al., 2005; Dahlbom et al., 2011; Jakobsson, 2012; Nekby et al., 2015), the focus of these studies is only on students' confidence about their own performance and not on their confidence about their relative performance which is key to the concept of competitiveness.[6]

A second contribution of our study is that, to the best of our knowledge, it is the first to investigate whether the gender gap in the willingness to compete depends on the difficulty of a task at hand. However, a small number of studies have examined related questions. In a series of experiments, Windschitl et al. (2003) investigated the moderating impact of task difficulty (shared adversities or shared benefits in their treatments) on subjects' confidence about their relative task performance. Yet, their focus is on the self-assessed performance in an arranged competition. The impact of task difficulty on subjects' degree of confidence in their skill relative to others is the subject of the lab experiment by Hoelzl and Rustichini (2005), but the authors do not study gender differences. Finally, Gneezy et al. (2003) let the participants in one of their experiments choose the difficulty level from a menu of tasks and show that men prefer, on average, more difficult tasks in return for a higher piece-rate reward. Nevertheless, their findings do not imply that men more often than women prefer a more competitive environment when the difficulty of a task increases.[7]

[3] In fact, most studies in this field use (a variation of) the lab task developed by Niederle and Vesterlund (2007), where participants have to add up sets of two-digit numbers within a couple of minutes.

[4] Lavy (2013) is an exception, who does not find a gender-performance gap in an experiment with high-school teachers.

[5] De Paola et al. (2015) also test whether men are more likely to enter a tournament in the same context. However, their study involves a 'risk-free' tournament choice. If students decide against the tournament, they will not earn any points for their bonus questions (i.e., there is no piece-rate alternative as in our study). This might also explain why 85% of their participants opted for the tournament and why they found no difference in tournament entry between men and women.

[6] In line with the studies mentioned above, a general finding is that male students are more overconfident (or less underconfident) about their exam results than female students. Also in the context of financial trading, men appear to be more overconfident than women (Barber and Odean, 2001; Deaves et al., 2009).

[7] In another recent study, Krawczyk and Wilamowski (2019) use data on amateur runners to test whether task difficulty (as captured by the length of a race) affects overconfidence as measured by the degree of slowdown during the race and the percentage difference between the individual's expected and actual finishing time. Like the other studies on overconfidence discussed above, these measures concern, however, an individual's estimate of her own absolute performance and not her relative performance in comparison to others.

The remainder of this paper is organized as follows: Section 2 discusses our experimental design and Section presents 3 the results. Section 4 summarizes and discusses potential practical implications of our study.

2. Data

2.1. Setting

We administered three rounds of experiments between 2015 and 2017, as part of the introductory microeconomics course in the undergraduate business and economics track at Utrecht University, the Netherlands. The course is one of the four compulsory courses in the first semester and runs for nine weeks between November and January each year. We argue that business and economics students are an interesting group to study because these students are likely to end up in managerial positions and especially in these positions gender differences appear to be very persistent. For a similar reason, Reuben et al. (2015) focus on MBA students.

Format, content, and assessment method are in line with the established practices for an undergraduate economics course in the Netherlands and remained stable in the relevant time period. In particular, the course is taught in eight general lectures and sixteen tutorials held in smaller classes of around 20–30 students. The assessment is based on three written examinations: two multiple choice midterm exams (in weeks four and seven) and one open-ended final exam (in week nine), each counting towards the final course grade with the weights 15%, 15%, and 70%, respectively. Following Utrecht University policy, the multiple choice exams were machine-graded by the independent 'Test and Evaluation Service Center' of the university, whereas the open-ended exams were graded by a team of 5–6 examiners. While examiners can see the student names, exams are split up by questions and reassembled afterwards to minimize the role of subjectivity. As a standard, all exams are graded based on an absolute grading scheme. This means that each student's performance is, in principle, independent of the performance of other enrolled students. In order to pass the course, a final grade of 5.5 or higher is needed (in the Dutch system, grades run from 1 to 10 and are rounded to the nearest half-point). Students who have failed the course can do a resit exam three months later, given that their final grade is higher or equal to 4.0.

Content-wise, the course follows closely the introductory microeconomics textbook by Pindyck and Rubinfeld (2013), as well as its accompanying test bank. Students are trained in formal logical reasoning and familiarized with basic economic concepts, such as equilibrium supply and demand, consumer and producer theory, and monopoly pricing. To support this goal, the test bank contains a large pool of multiple choice questions, which are used for weekly self-assessment exercises and for our experiment.

2.2. Experimental design

We conducted the experiments as part of the final exam. The exam lasted three hours in total. During that time, students were mostly answering open-ended questions on the entire course material. The experimental part appeared at the end of the exam sheet. There, students were offered five multiple choice questions comparable in form and content to the multiple choice questions of the midterm exams. Answering these questions earned the students some bonus points in a way clearly communicated to them (see instructions in Appendix B1) and described in more detail below. The appearance of these bonus questions on the exam and their precise role were, however, not communicated neither before nor after the exam to avoid potential experimenter effects.[8] Neither were examiners made aware of the purpose of the bonus questions or of the experimental setup and differences in treatments. For our experiments also the choice of the reward scheme was recorded separately during the grading procedure, to safeguard non-interference and avoid spillovers.

Approval of the board of examiners was requested in advance and granted based on the following arguments: (i) Bonus questions in exams are not uncommon at the department where the experiments took place, (ii) the bonus questions demanded only a minor share (5–10 minutes) of the total available time for the exam and therefore only minimally interfered with the regular part of the exams, and (iii) students could avoid all uncertainty through their choice of scoring method.[9] The new element of our experiment was thus only that students needed to choose between two different scoring methods.

We conceptually followed Niederle and Vesterlund (2007) in the design of these scoring methods. Students were offered the choice between a competitive and non-competitive scoring method for their bonus questions. The default option, an absolute grading scheme, yielded one point per correct answer. The alternative option, a tournament grading scheme, yielded four points per correct answer, but only if the student's answers belonged to the best. Otherwise, he or she received 0 points for the bonus question part. We chose as the critical threshold the top 25% of all answers submitted in a cohort including also those students who opted for the absolute grading scheme. In this way, we tried to minimize the role of gambling. Because the expected number of bonus points is unconditionally the same for both options, a risk-neutral decision maker would be indifferent between the two scoring options, unless he or she expects to perform better than other students. To further lower the influence of risk, students could choose their preferred scoring method after having answered the questions. In this way, we did not burden students with the task of predicting the bonus questions that they had to solve or their own aptitude in answering them. The choice of reward scheme is therefore determined by a correct assessment of one's relative ability compared to a (familiar) peer group (and a possible genetic disposition for the thrill of competing against others).[10]

The bonus points earned in this way counted towards the grade of the second midterm exam. The maximum number of points for that exam was 45 excluding the bonus questions. This means that a student who answered all bonus questions correctly and chose the absolute grading scheme could raise her midterm grade by $5/45 * 10 = 1.11$ (on a scale from 1 to 10) or, after weighting this grade, her final course grade by $1.11 * 0.15 = 0.167$. Choosing the tournament scheme could yield a final grade gain of $20/45 * 10 * 0.15 = 0.667$. Hence, because final grades are rounded to nearest half-points, the bonus questions could translate into maximum a half point increase when the absolute grading scheme was selected and a maximum full grade difference if the tournament scheme was selected.[11]

To analyze the impact of task difficulty on the choice of scoring method, we administered three different experiments. All students of the 2015 cohort were exposed to a set of relatively easy bonus questions (Experiment 1). The difficulty was significantly increased for the students of the 2016 cohort (Experiment 2). In the 2017 cohort, we tried to alleviate concerns about cohort-specific confounders (e.g., changes in student composition, teaching staff, etc.) that might invalidate comparisons between Experiments 1 and 2. We therefore randomly exposed half

[8] We were of course not able to exclude the possibility that students of older cohorts informed their younger fellows of the questions' appearance.

[9] Students at Utrecht University are familiar with a correction for guessing formula for multiple choice exams as it is commonly applied by the 'Test and Evaluation Service Center'. However, it was made clear to our subjects that the bonus questions in our experiment did not include such a correction.

[10] This is in contrast to other confidence studies using the exam setting (e.g. Nekby et al., 2015), where the choice of reward scheme is primarily determined by students' confidence in their own skills and/or their willingness to take risk.

[11] For example, if a student received a 7.1 based on the regular exams and answered all bonus questions correctly, the final rounded course grade would be 7.5 (absolute grading scheme) or 8 (tournament grading scheme).

of the students to the easy bonus questions and the other half to the difficult questions (Experiment 3), allowing for a within-cohort analysis.[12]

All the bonus questions are from the Pindyck and Rubinfeld (2013) test bank. Questions concern the basic principles of microeconomics. For instance, students are asked about income and substitution effects resulting from a decline in the price of an inferior good. Other questions refer to graphical analysis. For example, where on the average total cost curve is a perfectly competitive firm earning negative profits? Thus, the experimental task consists of standard questions in economic reasoning. When selecting the bonus questions for the different experiments, we made use of the fact that the Pindyck and Rubinfeld (2013) test bank distinguishes between questions of three difficulty levels, varying from 1 (easy) to 3 (difficult). In particular, for Experiment 1 and the easy treatment of Experiment 3, we chose five questions with an average difficulty of 1.6. For Experiment 2 and the difficult treatment of Experiment 3, the questions had an average difficulty of 2.8.

2.3. Sample

First, as expected by the small effort costs, the participation rates in our experiments were quite high. The shares of first-year students that participated in the final exams were around 89% on average during 2015–17. Of those, 93.5% participated in our experiments. The non-participants were almost exclusively students who gave up early in the exams and earned a grade of 0. Hence, our main sample includes all first-year students who have actively participated in the study program. Students may repeat the course and therefore could participate multiple times in our experiments: in these cases, we included only data of the first participation in the analysis.[13]

More men than women enroll in the business and economics program at Utrecht University and the same gender composition can also be found in our sample (Panel A of Table 1). The shares of female first-year students that registered for the course are around 32.5%. In our experiments, these shares are only slightly lower (31% on average).

Panel B of Table 1 presents the bonus question scores of our participants. The bonus question scores split up by year and/or treatment can be found in Appendix Table B1. Clearly, the numbers confirm the validity of our treatment. The number of correct answers dropped significantly between the easy and difficult task treatments for both men and women. This pattern can be found in the aggregate (Table 1) as well as in the separate comparison between the two treatments of Experiment 3 and the comparison between Experiments 1 and 2 (Table B1).

The course performance of all students participating in our experiment is shown in Panel C of Table 1, split up by gender and experimental treatment they were exposed to. The course performance of the three separate student cohorts is presented in Appendix Table B1. As expected, overall course performance (excluding the bonus questions) does not systematically vary between cohort and experimental treatment.[14] Also, the failing rates (not presented in the table) are similar across years: 33.3% (2015), 37.9% (2016), and 33.9% (2017), which are typical percentages for this course. As expected from earlier studies, female students achieved on average a higher grade in the mid- and end-term exams than their male counterparts, which is particularly the case for the regular part of the final exams.

Table 1
Task and course performance (Experiments 1–3).

	All	Easy	Difficult	Difference Easy-Difficult
Panel A: Number of participants				
All	681	340	341	
Men	471	230	241	
Women	210	110	100	
Panel B: Experimental task performance (No. of correct bonus questions)				
All	2.6	3.1	2.1	1.0***
Men	2.6	3.0	2.1	0.9***
Women	2.6	3.3	1.9	1.4***
Diff Men-women	0.0	-0.3*	0.2**	
Panel C: Course performance (excluding bonus questions)				
Midterm 1 grade				
All	5.1	5.3	4.8	0.5***
Men	5.0	5.3	4.8	0.5***
Women	5.1	5.4	4.9	0.5***
Diff Men-women	-0.1	-0.1	-0.1	
Midterm 2 grade				
All	5.2	5.4	5.1	0.3**
Men	5.1	5.2	5.0	0.2*
Women	5.6	5.7	5.4	0.3*
Diff Men-women	-0.5***	-0.5***	-0.4**	
Final exam grade				
All	6.5	6.9	6.1	0.7***
Men	6.3	6.6	5.9	0.7***
Women	7.0	7.3	6.6	0.7***
Diff Men-women	-0.7***	-0.7***	-0.7***	

NOTES. (a) Easy refers to the pooled sample of Experiment 1 and Experiment 3 (easy treatment); Difficult refers to the pooled sample of Experiment 2 and Experiment 3 (difficult treatment); (b) all grades are from [1,10], with 10 highest; (c) Asterisks are based on two-sided Mann-Whitney U tests. * $p < .1$, ** $p < .05$, and *** $p < .01$.

3. Results

3.1. Main result

Table 2 presents the tournament entry choices by gender. On average across the three experiments, men are 13 percentage points more likely to select the tournament scheme than women. Hence, we find evidence of a gender gap in the willingness to compete in our field setting. However, a striking finding is that this result depends critically on the difficulty of the task at hand: When the task is easy, the gap is significant and sizable (around 20 percentage points), while the gap is much smaller (around 7 percentage points) and statistically insignificant when the task is difficult. This finding holds for the pooled sample (Panel A), as well as for comparisons between Experiments 1 and 2 (Panel B) and within Experiment 3 (Panel C), where we randomly varied the difficulty of the task. In the difficult-task treatment of Experiment 3, the gender gap in tournament entry narrows to 8.2 percentage points. Interestingly, conditional on gender and task difficulty, the shares of tournament-entry choices in Experiments 1 and 2 are almost identical to the ones of the easy, respectively difficult, treatment in Experiment 3. This finding suggests that confounding cohort effects, invalidating a direct comparison of Experiments 1 and 2 to test the role of task difficulty, play no substantial role.

Men and women responded differently to an increase in task difficulty, which resulted in a narrowing of the gender gap in the difficult treatment. The share of women choosing the tournament scheme is re-

[12] This means that the 2017 cohort received bonus questions of different difficulty levels. To alleviate concerns about unfairness, we organized two separate competitions, one for each of the two difficulty levels. We then raised the bonus question scores of the students with the difficult questions ex-post so their mean score was identical to the mean score in the easy-task treatment. Of course, the participating students were unaware of this issue as they did not know that the difficulty of the bonus questions varied between different versions of the exam.

[13] There were 14 students who participated twice in the experiments during 2015–2017. Including the second participation of these repeaters hardly affects our findings.

[14] There are two exceptions to this rule: the first midterm exam and the final exam of 2016, which were significantly more difficult than the first midterm and final exams in the other two years. As a result, the grades on these exams dropped considerably in 2016, which is also reflected in the average grades of the students participating in our difficult task treatments presented in Table 1. We will take account of this fact in our regression analysis.

Table 2
Tournament-entry choices by experiment and gender.

	All	Easy	Difficult	Difference Easy-Difficult
Panel A: Experiments 1–3				
All	29.9%	35.0%	24.9%	10.1%***
Men	33.8%	41.3%	26.6%	14.7%***
Women	20.9%	21.8%	20.0%	1.8%
Diff Men-women	12.8%***	19.5%***	6.5%	
No. Participants	681	340	341	
Panel B: Experiment 1–2		(Exp 1)	(Exp 2)	
All	29.8%	35.0%	24.4%	10.6%**
Men	33.7%	41.8%	26.1%	15.7%**
Women	20.7%	21.0%	20.3%	0.7%
Diff Men-women	13.0%**	20.7%***	5.7%	
No. Participants	463	234	229	
Panel C: Experiment 3				
All	30.1%	34.9%	25.7%	9.2%*
Men	33.8%	40.3%	27.6%	12.6%*
Women	21.4%	23.5%	19.4%	4.1%
Diff Men-women	12.4%**	16.7%*	8.2%	
No. Participants	218	106	112	

Notes. Asterisks are based on two-sided Mann–Whitney U tests. * $p < .1$, ** $p < .05$, and *** $p < .01$.

Table 3
Tournament-entry choices by task performance.

	All	Easy	Difficult	Difference Easy-Difficult
Panel A: Prospective winners (top 25% bonus question score)				
All	40.0%	48.9%	35.3%	13.5%**
Men	48.4%	62.2%	33.3%	28.9%***
Women	31.5%	26.5%	41.6%	-15.1%
Diff Men-women	16.9%**	35.7%***	-8.3%	
Panel B: Prospective losers (bottom 75% bonus question score)				
All	23.2%	26.3%	20.6%	5.7%
Men	26.4%	29.7%	23.5%	6.2%
Women	15.3%	18.0%	13.2%	4.8%
Diff Men-women	11.1%**	11.7%*	10.3%*	

Notes. (a) Easy refers to the pooled sample of Experiment 1 and Experiment 3 (easy treatment); Difficult refers to the pooled sample of Experiment 2 and Experiment 3 (difficult treatment); (b) prospective winners (losers) are determined based on whether a participant's bonus question score belongs to the top 25% (bottom 75%) of all submitted answers (tournament contestants and piece-rate choosers together) independent of whether a participant actually chooses the tournament; (c) Asterisks are based on two-sided Mann-Whitney U tests. * $p < .1$, ** $p < .05$, and *** $p < .01$.

markably stable across all cohorts and treatments and lies in the range of 19.4–23.5%. This difference is statistically insignificant in all three comparisons (Table 2, Panel A-C). In contrast, the tournament entry decision of men was substantially affected by task difficulty. While around 41% of them choose the tournament scheme in the easy treatments of Experiments 1 and 3, only 26–27% choose the tournament in the difficult treatments of Experiments 2 and 3. This difference is highly significant (p<.01) in the pooled sample (Panel A) as well as the comparison between Experiments 1 and 2 (Panel B), and marginally significant (p<.1) in Experiment 3 (Panel C).

Interestingly, the descriptive results in Table 2 are largely in line with modern theories of confidence. Compared to the 25%-benchmark of students who should have chosen the tournament scheme based on the higher number of expected grade points they could earn from this scheme, students in our easy-task treatments selected the tournament scheme too often: 35% on average over the three years. This suggests that our participants were, on average, overconfident in the easy-task treatments. In contrast, there is no clear indication that students were overconfident on average in the difficult task treatments.[15]

3.2. Winners, losers, and welfare implications

On average almost 30% of the students selected the tournament option (Table 2). Compared to the counterfactual situation where bonus questions are rewarded according to a piece-rate scheme (the common practice), introducing a choice between compensation schemes created a group of winners and losers: under the assumption that the introduction of the compensation choice had no impact on performance, some students are better off (they selected the tournament scheme and belonged to the top 25% performers), while others are worse off (they selected the tournament scheme but belonged to the bottom 75% performers).[16]

In Table 3 we distinguish between the 'upper quarter' of our participant pool, that is, the participants who — based on their actual bonus question scores — would have won the tournament (the prospective winners), and the 'bottom three-quarters' of students who would have lost the tournament. The impact of task difficulty on the tournament-entry choices of men can be found for both groups alike. It is, however, most pronounced among the prospective winners. As shown in Panel A of Table 3, the gender gap in our easy-task treatments is even more sizable among the prospective winners than for our entire subject pool: 62.2% of prospective male winners versus 26.5% of prospective female winners choose the tournament, yielding a gender gap in tournament-entry choices of 35.7 percentage points. In the difficult-task treatments, in contrast, the gender gap is even reversed: 33.3% of men versus 41.6% of women choose the tournament there. In other words, despite the fact that this difference is statistically insignificant, our female participants in the difficult-task treatments rightfully (ex-post) chose the tournament option even more often than their male counterparts. Moreover, as shown in Panel B of Table 3, among the prospective losers men are more likely to opt for the tournament scheme in both treatments. However, the prospective losers are less likely to enter the tournament when the task is difficult (although this difference is not statistically significant).

3.3. Can differences in (pre-)task performance explain the results?

As suggested by the descriptive results, the key mechanism behind the effect of task difficulty on the gender gap in tournament-entry choices seems to be the significant drop in the male tournament entry rate when moving from the easy to the difficult task treatments. What might have caused this drop? Below we discuss and rule out three performance-based explanations.

First, the average grades of our male (and female) participants in the mid-term exams and the regular parts of the end-term exams are statistically indistinguishable across all three experiments. The single exception to this rule is the mid-term exam of 2016 (as mentioned in footnote 12). Nevertheless, it is not clear why the male participants in the easy-

[15] In fact, the null-hypothesis that the share of tournament choices is smaller or equal to 25% in the easy-task treatments of Experiment 1 and Experiment 3 can be rejected at $p < .05$. We cannot reject the null-hypothesis that the share of tournament choices is greater or equal to 25% in the difficult treatments of Experiments 2 and 3. Note that our evidence for overconfidence in the easy-task treatments is not significantly affected by the small number of questions in our experiment and the several ties we had to break in order to determine the winners. In fact, the average bonus question score among the participants choosing the tournament for the easy tasks is no larger than 3.6, which means that many of them took a risky decision.

[16] The reward they earned through this choice — as a percentage of the total grade points a student would have earned for the course without the tournament option — are illustrated in Figure B1.

task treatments are better equipped to answer the bonus questions than their male counterparts in the difficult-task treatments.[17] Second, based on the persistent and substantial gender performance gap in the regular exam parts, one might argue that our male participants needed the bonus points more badly than our female participants to pass the exam. Men may therefore be more willing to gamble in the bonus questions part of the exam. However, as this argument applies to all three experiments alike, it is not clear why men would gamble more often in the easy treatments. Third, in the bonus question part, our male participants perform in no obvious way worse than the female participants. In fact, men even outperform women in the difficult treatment of Experiment 3. Hence, there is again no obvious performance-based explanation for why men should have selected the tournament less often when the task is difficult.

To further examine whether men's reduced willingness to compete in the difficult task treatments is in no systematic way related to their (pre-)task course performance, we regressed the tournament-entry choices of our participants on their gender, exam grades, and bonus question scores. The results from probit models presented in Table 4 show that men are more likely to select the tournament scheme (on average) and that the difference in the willingness to compete is highly dependent on the difficulty of the task. In all specifications, we find evidence of a significant gender gap in tournament entry in the easy task treatment and no significant gender gap in the difficult task treatment. The overall gender difference is clearly driven by the gender differences in the easy task treatments. This is also clear when considering the experiments separately: there is a sizable and significant gender gap in Experiment 1 (Panel C), no significant gender gap in Experiment 2 (Panel D), and a significant gender gap in Experiment 3 (Panel E), which is driven by the easy task treatment. These findings hold across all specifications. In additional analyses, we allow for interactions between the (past) task performance and gender: this hardly affect our main results (see Appendix Table B2 and Appendix Table B3). Furthermore, results from linear models are also consistent with the results from the probit models.[18] The coefficients of the interaction terms between gender and task difficulty are significant in almost all specifications, indicating that increasing task difficulty significantly reduces the gender difference in the willingness to compete.[19] It is also striking that the point estimates based on Experiment 1–2 are very similar to the point estimates based on Experiment 3, despite the fact that these models are estimated on different samples (cohorts). Based on these findings we conclude that men reduce their willingness to compete more than women when they are faced with a more difficult task.

Finally, we also included interaction terms between gender and (past) task performance in the linear models (see Appendix Table B4). These results confirm the findings from the previously discussed probit and OLS models: the changes in the estimates of the interaction between gender and task difficulty are negligible. However, the main gender effect becomes larger when we include interactions between gender and performance measures that are not directly part of the experiment (Column (5)-(7)): While the estimates hovers around 22 percentage points in models without interactions (Appendix Table A1; Column (5)-(7)), the estimates increases to 33–38 percentage points when these interactions are included. This suggests that men and women weigh past performance and assessments differently when choosing the payment scheme. We discuss this issue more extensively in Section 3.5.

3.4. Does risk aversion explain the results?

There is robust evidence that men tend to be less risk averse than women. Following the literature on gender differences in competitiveness (Buser et al., 2014; Reuben et al., 2017), we test whether heterogeneity in risk aversion could in theory explain (part) of the observed gender difference in tournament entry. Almost all students who participated in our experiments also participated in the so-called "Matching Days" before entering the undergraduate program. During this day the students participated in economics classes and in lab-in-the-field experiments. Before participation in the lab-in-the-field experiments, we asked students for their consent to use their data for research purposes.[20] This allows us to link the information from these experiments to our data. Specifically, we use two measures to capture variation in risk preferences. First, students were asked to assess their risk attitudes on a 10-point Lickert scale, as in Dohmen et al. (2011). Second, we used an incentivized measure by administering the "bomb risk elicitation task", as introduced by Crosetto and Filippin (2013).[21] The two measures of risk aversion are significantly correlated, although the correlation is relatively weak (around 0.25). While the risk aversion measure elicited through the survey questions is significantly correlated to all course performance measures, the risk aversion measure elicited through the bomb task is not significantly related to any of the course performance variables. In line with most of the literature, women are significantly more risk averse than men according to both measures.

Table 5 presents the results when introducing risk aversion according to the survey question (Panel A) or risk aversion elicited through the bomb task (Panel B) as an additional control. It appears that risk aversion is not systematically related to the decision to opt for the tournament scheme. Moreover, the main results are hardly affected by the introduction of this additional explanatory variable: on average, there is a significant gender difference in the tournament entry decision, but this effect varies depending on the difficulty of the task at hand.[22]

A related issue is whether students perceived the tournament option in the easy or in the difficult treatment relatively more risky.[23] Although it is not clear a priori under which treatment the tournament scheme will be perceived more risky, we can test this empirically: risk aversion should matter more for the tournament entry decision in the task where the tournament option is the most risky. Models that include interactions between risk aversion and task difficulty show that there is no significant difference in the risk aversion coefficients between the two treatments (see Appendix Table B6 and B7). This suggests that the perceived riskiness of the tournament does not depend on task difficulty in our experiments.

[17] Based on their average high-school grades (of the subsample of students for which we have high-school grades), the 2016 male cohort answering the difficult questions is actually slightly better than the 2015 male cohort answering the easy questions: the average grades of the cohorts are 6.65 and 6.43, respectively (statistically significant at $p < .05$). Note that we do not include high-school grades in our main analysis because we do not have this information for the majority of participants, and in particular not for the many foreign students whose grades are difficult to compare with the Dutch grading system.

[18] The results from a pooled linear probability model are presented in Appendix Table A1; Appendix Table A2 shows the results from separate regressions for Experiments 1–2 and Experiment 3.

[19] The exceptions are the models that do not control for the bonus question score (Column (3) of Appendix Table A2). Of course, one can argue that the models should control for the bonus question scores as these are directly affected by task difficulty.

[20] Data protection is guaranteed and outlined in the privacy policy as stipulated in the terms and conditions of the experimental platform GXP through which we administered the tasks. See https://gxp.world.

[21] The incentive consisted of lottery tickets for an iPad2.

[22] To further examine whether our main findings can be explained by heterogeneity in risk preferences we extend our analysis and allow for gender risk aversion interactions. It appears that this does not substantially affect our main results (Appendix Table B5).

[23] For instance, students may anticipate potential ties among students who chose the tournament scheme. In fact, on average over the three years, 33% of the students choosing the tournament were eligible for winning it. Out of these, we randomly drew 25% of winners.

Table 4
Tournament-entry choices.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Experiments 1–3 (coefficients)							
Female	-0.389***	-0.400***	-0.559***	-0.679***	-0.688***	-0.672***	-0.692***
	(0.114)	(0.115)	(0.158)	(0.164)	(0.165)	(0.166)	(0.167)
Difficult task		-0.313***	-0.407***	-0.132	-0.127	-0.123	-0.164
		(0.102)	(0.120)	(0.130)	(0.130)	(0.130)	(0.132)
Female*Difficult task			0.343	0.495**	0.492**	0.490**	0.497**
			(0.230)	(0.238)	(0.238)	(0.239)	(0.240)
Bonus score				0.319***	0.313***	0.301***	0.288***
				(0.0476)	(0.0499)	(0.0505)	(0.0509)
Final exam grade					0.0148	-0.0168	-0.0640
					(0.0334)	(0.0395)	(0.0443)
Midterm 1 grade						0.0612	0.0359
						(0.0409)	(0.0423)
Midterm 2 grade							0.105**
							(0.0437)
Panel B: Experiment 1–3 (marginal effects): Easy/difficult task pooled							
Gender difference	-0.133***	-0.135***	-0.130***	-0.136***	-0.139***	-0.135***	-0.139***
	(0.0382)	(0.0379)	(0.0354)	(0.0342)	(0.0347)	(0.0350)	(0.0348)
Gender difference at:							
Easy task			-0.195***	-0.200***	-0.202***	-0.197***	-0.204***
			(0.0510)	(0.0435)	(0.0435)	(0.0439)	(0.0441)
Difficult task			-0.0656	-0.0582	-0.0620	-0.0577	-0.0609
			(0.0491)	(0.0534)	(0.0539)	(0.0542)	(0.0534)
N	681	681	681	681	681	681	681
Panel C: Experiment 1 (marginal effects): Easy task							
Gender difference			-0.213***	-0.225***	-0.232***	-0.208***	-0.219***
	(0.0640)			(0.0592)	(0.0592)	(0.0604)	(0.0600)
N	234			234	234	234	234
Panel D: Experiment 2 (marginal effects): Difficult task							
Gender difference	-0.0592			-0.0504	-0.0497	-0.0415	-0.0436
	(0.0644)			(0.0604)	(0.0619)	(0.0628)	(0.0616)
N	229			229	229	229	229
Panel E: Experiment 3 (marginal effects): Easy/difficult task randomized							
Gender difference	-0.128*	-0.127*	-0.123**	-0.128**	-0.127**	-0.128**	-0.128**
	(0.0665)	(0.0661)	(0.0623)	(0.0576)	(0.0593)	(0.0591)	(0.0591)
Gender difference at:							
Easy task			-0.167*	-0.172**	-0.170**	-0.171**	-0.172**
			(0.0929)	(0.0696)	(0.0707)	(0.0705)	(0.0705)
Difficult task			-0.0819	-0.0651	-0.0632	-0.0648	-0.0648
			(0.0836)	(0.0890)	(0.0905)	(0.0903)	(0.0903)
N	218	218	218	218	218	218	218

NOTES: Coefficients (Panel A) and average marginal effects (Panel B-E) from probit models (standard errors in parentheses). * $p < .1$, ** $p < .05$, and *** $p < .01$

3.5. Task difficulty and tournament-entry choices: An explanation

The finding that men's (but not women's) tournament entry choices are substantially affected by task difficulty cannot be explained by differences in task performance or risk aversion. Here we turn to a potential alternative explanation for the pattern documented in this study. Interestingly, it is the response of our male participants that is in line with the most recent Bayesian theories on this topic (Benoît and Dubra, 2011; Benoît et al., 2015), despite the fact that their choices are ex-post inefficient. The reason is that performing well in a task is interpreted as a signal that one's ability is higher than expected by a Bayesian updater. Poor performance, in contrast, is interpreted as a lack of ability. When confronted with an easy task, Bayesian updaters are more confident about their performance and select a competitive scheme more often. This is consistent with the behavior of the male students. However, the female students hardly responded to variation in task difficulty; the results for women therefore do not provide support for Bayesian theories (although the direction of the point estimates is consistent with the theoretical prediction).

An explanation for the result that women responded less to task difficulty can be found in the work of Mobius et al. (2011). The authors conduct an online experiment and find that women are significantly more conservative than men in updating their beliefs upon arrival of a noisy signal of their relative performance on a computerized IQ test. Their finding is supported by early survey studies conducted by Roberts and Nolen-Hoeksema (1989), who find that women place significantly more weight on their past experience with similar tasks and external assessments by peers and supervisors in their evaluation of their performance on an ongoing task. Hence, this line of work provides an explanation for why task difficulty does not not matter much for the tournament-entry decisions of our female participants.

Even though a comprehensive test of this theory is beyond the scope of this study, we provide some indicative evidence that supports this explanation. We test this explanation using three past performance indicators: Midterm 1 and Midterm 2 grades (available for all participants), and average high-school grades (available only for a subset of Dutch students).[24] If women place more weight on past performance and assessments, one may expect that these variables are stronger predictors of the tournament entry for women than for men. In support of a more

[24] For most students we have data on nationality (Dutch/foreign). Foreign students are more likely to opt for the tournament scheme. However, controlling for student nationality does not substantially affect our main results (see Appendix Table B9).

Table 5
Tournament-entry choices: controlling for risk aversion.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Risk aversion (survey question)							
Risk aversion	0.00111	-0.00119	-0.00113	-0.00759	-0.00842	-0.00958	-0.0124
	(0.0103)	(0.0102)	(0.0102)	(0.0100)	(0.0101)	(0.0101)	(0.0101)
Gender difference	-0.154***	-0.156***	-0.156***	-0.159***	-0.167***	-0.162***	-0.167***
	(0.0382)	(0.0376)	(0.0379)	(0.0371)	(0.0373)	(0.0378)	(0.0374)
Gender difference at:							
Easy task			-0.226***	-0.222***	-0.227***	-0.222***	-0.228***
			(0.0542)	(0.0472)	(0.0469)	(0.0474)	(0.0479)
Difficult task			-0.0809	-0.0758	-0.0882	-0.0835	-0.0898
			(0.0517)	(0.0584)	(0.0588)	(0.0593)	(0.0576)
N	563	563	563	563	563	563	563
Panel B: Risk aversion (bomb task)							
Risk aversion	0.00638	0.00443	0.00477	0.00270	0.00304	0.00302	0.00207
	(0.00861)	(0.00854)	(0.00851)	(0.00834)	(0.00833)	(0.00834)	(0.00830)
Gender difference	-0.151***	-0.155***	-0.155***	-0.162***	-0.171***	-0.167***	-0.173***
	(0.0375)	(0.0369)	(0.0371)	(0.0361)	(0.0364)	(0.0368)	(0.0364)
Gender difference at:							
Easy task			-0.221***	-0.222***	-0.227***	-0.223***	-0.231***
			(0.0533)	(0.0464)	(0.0461)	(0.0464)	(0.0468)
Difficult task			-0.0822	-0.0820	-0.0955	-0.0916	-0.0986*
			(0.0513)	(0.0577)	(0.0581)	(0.0584)	(0.0569)
N	568	568	568	568	568	568	568

NOTES: Entries represent average marginal effects of probit models (standard errors in parentheses). The original bomb task measure (0–100) is divided by 10 for presentation purposes. * $p < 0.1$, ** $p < 0.05$, and *** $p < 0.01$.

task-sensitive updating strategy of men and a more conservative updating rule applied by women, we find stronger effects of the pre-task performance variables on women's tournament entry choices (in particular midterm 1 grades and high-school grades, see Appendix Table B9).

4. Conclusion and discussion

Niederle and Vesterlund (2007) and many experiments after their seminal paper demonstrated a striking gender gap in the choice of a compensation scheme for a stylized experimental task, providing support for the hypothesis that women tend to shy away from competitive environments. We administered a series of experiments around a final exam of a university course, where we varied the difficulty level of a number of exam questions. Our setting is fundamentally different from those performed typically in the lab for several reasons. First, the participants in our experiment were familiar with the potential contestants (other students) and the type of task at hand (solving exam questions) because similar tasks are actually part of the daily lives of students. Hence, compared to results from the lab, our findings are more relevant for promotion competitions in organizations or university programs with competitive rank-order grading schemes. Second, the participants in our experiment are undergraduate business and economics students. These students are likely future candidates for managerial positions and may climb to the higher end of the corporate ladder. This is therefore a relevant population for the environments that this literature ultimately wants to speak to (as in e.g. Reuben et al., 2015).

Our results show that on average across our experiments men choose more often than women a competitive rank-order tournament as their preferred reward scheme. However, we also find that the gender gap in compensation scheme choices is significantly affected by the difficulty of the underlying task: whereas there is a sizeable gender gap in our easy-task treatments, the gender gap is small and insignificant in our difficult-task treatments. The narrowing of the gender gap is, in fact, mostly due to the impact of task difficulty on the willingness to compete of men, whereas women's choices are almost entirely unaffected in our experiments. In sum, our experiments do replicate the previously found gender gap from numerous lab studies. Gender differences in the willingness to compete thus seem to play a role outside the lab as well. Nevertheless, if these gender differences in competitiveness only appear in settings involving easy tasks, one can cast doubt about whether gender differences in competitiveness can explain the under-representation of women in top positions.

More generally, our study demonstrates that it is important to replicate lab experiments in the field — and to replicate these replications. We essentially provide evidence from three field replication studies and we show that conclusions derived from replications may hinge on rather subtle differences in the experimental design. Indeed, across our experiments, there are no noteworthy differences in participants (i.e., all undergraduate students in the same program) and the experimental design is almost identical in all three experiments (i.e., we use the same type of task in all of them). However, changing the difficulty of questions had major implications in our study. Results from a single replication study (Experiment 1 or Experiment 2) would not have revealed this important pattern.

We conclude by pointing out three directions for future research. First, although we claim that we study a highly relevant sample, it is clear that our sample is not representative of the general population or business and economics students worldwide. We should therefore be careful generalizing our findings, especially since recent evidence shows that competitiveness (measured using lab tasks) seems to be dependent on culture and institutions (Booth et al., 2019; Zhang, 2019). Hence, an important avenue for future research will be to test the gender difference in competitiveness using similar real-life tasks but other samples in other contexts.

Second, our results suggest that the prospective losers in our difficult treatments take better decisions (from an ex-post perspective) than the prospective losers in our easy treatments. Interestingly, the prospective female winners choose the tournament more often in our difficult treatments. Task difficulty therefore seems to correct for the typically observed tournament 'under-entry' of women who would have won the tournament but decided not to enter and observed tournament-'over-entry' of those students who have no chance of winning the tournament (e.g., Niederle and Vesterlund, 2007). This correction of task difficulty and the welfare implications deserve more attention in future research.

Finally, our result that men's but not women's tournament entry choices are affected by task difficulty might suggest a way to further develop theories of self-confidence. A promising direction for future research might be the integration of Bayesian updating theories (Benoît and Dubra, 2011; Benoît et al., 2015) with the biased, self-serving learning theory proposed by Mobius et al. (2011). At least in our context, a combination of their main insights helped to explain why task difficulty mattered for men and not for women.

Acknowledgments

We would like to thank two anonymous referees and participants of the KVS New Paper Sessions, the ESPE Annual Conference and the EALE Annual Conferece for helpful comments and suggestions. This study makes use of data obtained during the Matching Days at Utrecht University School of Economics (U.S.E.), financial support for the experiments is gratefully acknowledged. We thank the administrative staff of U.S.E., and Margreet van Soest and Andrea Naylor in particular, for their excellent organizational support.

Appendix A. Results from OLS models

Table A1
Tournament-entry choices: Pooled linear regressions.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	-0.128***	-0.132***	-0.195***	-0.222***	-0.226***	-0.219***	-0.222***
	(0.0377)	(0.0375)	(0.0522)	(0.0506)	(0.0510)	(0.0511)	(0.0510)
Difficult task		-0.108***	-0.147***	-0.0557	-0.0537	-0.0528	-0.0634
		(0.0346)	(0.0415)	(0.0421)	(0.0423)	(0.0423)	(0.0424)
Female*Difficult task			0.129*	0.180**	0.179**	0.177**	0.176**
			(0.0749)	(0.0726)	(0.0727)	(0.0726)	(0.0724)
Bonus score				0.105***	0.102***	0.0984***	0.0936***
				(0.0148)	(0.0155)	(0.0157)	(0.0157)
Final exam grade					0.00603	-0.00416	-0.0183
					(0.0108)	(0.0126)	(0.0139)
Midterm 1 grade						0.0202	0.0127
						(0.0129)	(0.0132)
Midterm 2 grade							0.0313**
							(0.0131)
Constant	0.338***	0.393***	0.413***	0.0981*	0.0658	0.0375	0.0215
	(0.0209)	(0.0273)	(0.0297)	(0.0528)	(0.0784)	(0.0804)	(0.0804)
Observations	681	681	681	681	681	681	681
R-squared	0.017	0.031	0.035	0.102	0.102	0.105	0.113

NOTES: Estimates are coefficients from OLS regressions (standard errors in parentheses). * $p < 0.1$, ** $p < 0.05$, and *** $p < 0.01$.

Table A2
Tournament-entry choices: linear regressions by experiment.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Experiments 1–2							
Female	-0.130***	-0.136***	-0.207***	-0.222***	-0.228***	-0.211***	-0.216***
	(0.0460)	(0.0457)	(0.0629)	(0.0618)	(0.0623)	(0.0621)	(0.0617)
Difficult task		-0.112***	-0.157***	-0.0853	-0.0788	-0.0724	-0.0951*
		(0.0420)	(0.0502)	(0.0518)	(0.0524)	(0.0521)	(0.0523)
Female*Difficult task			0.150	0.177**	0.175*	0.172*	0.167*
			(0.0914)	(0.0899)	(0.0900)	(0.0893)	(0.0887)
Bonus score				0.0805***	0.0755***	0.0629***	0.0583***
				(0.0183)	(0.0192)	(0.0196)	(0.0196)
Final exam grade					0.0111	-0.0108	-0.0296*
					(0.0135)	(0.0156)	(0.0169)
Midterm 1 grade						0.0491***	0.0353*
						(0.0178)	(0.0184)
Midterm 2 grade							0.0482***
							(0.0173)
Constant	0.337***	0.395***	0.418***	0.172***	0.112	0.0332	-0.00342
	(0.0253)	(0.0330)	(0.0359)	(0.0659)	(0.0987)	(0.102)	(0.102)
Observations	463	463	463	463	463	463	463
R-squared	0.017	0.032	0.038	0.077	0.078	0.093	0.109
Experiment 3							
Female	-0.124*	-0.123*	-0.167*	-0.233***	-0.231***	-0.235***	-0.235***
	(0.0661)	(0.0659)	(0.0946)	(0.0878)	(0.0888)	(0.0891)	(0.0894)
Difficult task		-0.0990	-0.126*	0.00461	0.00524	0.00607	0.00614
		(0.0616)	(0.0748)	(0.0720)	(0.0723)	(0.0724)	(0.0726)
Female*Difficult task			0.0856	0.205*	0.206*	0.208*	0.208*
			(0.132)	(0.123)	(0.123)	(0.124)	(0.124)
Bonus score				0.157***	0.158***	0.159***	0.158***
				(0.0251)	(0.0261)	(0.0262)	(0.0266)
Final exam grade					-0.00307	0.00351	0.00207
					(0.0184)	(0.0218)	(0.0250)
Midterm 1 grade						-0.0107	-0.0111
						(0.0188)	(0.0192)
Midterm 2 grade							0.00241
							(0.0204)
Constant	0.338***	0.389***	0.403***	-0.0518	-0.0361	-0.0250	-0.0244
	(0.0375)	(0.0489)	(0.0536)	(0.0879)	(0.129)	(0.131)	(0.131)
Observations	218	218	218	218	218	218	218
R-squared	0.016	0.028	0.030	0.180	0.180	0.181	0.181

NOTES: Coefficients from OLS regressions (standard errors in parentheses). * $p < 0.1$, ** $p < 0.05$, and *** $p < 0.01$.

Appendix B. Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.labeco.2020.101815.

References

Alicke, M.D., Klotz, M.L., Breitenbecher, D.L., Yurak, T.J., Vredenburg, D.S., 1995. Personal contact, individuation, and the better-than-average effect.. J. Pers. Soc. Psychol. 68 (5), 804.

Almås, I., Cappelen, A.W., Salvanes, K.G., Sørensen, E.Ø., Tungodden, B., 2015. Willingness to compete: family matters. Manage. Sci. 62 (8), 2149–2162.

Azmat, G., Petrongolo, B., 2014. Gender and the labor market: what have we learned from field and lab experiments? Labour Econ. 30, 32–40.

Balafoutas, L., Sutter, M., 2012. Affirmative action policies promote women and do not harm efficiency in the laboratory. Science 335 (6068), 579–582.

Barber, B.M., Odean, T., 2001. Boys will be boys: gender, overconfidence, and common stock investment. Q. J. Econ. 116 (1), 261–292.

Bengtsson, C., Persson, M., Willenhag, P., 2005. Gender and overconfidence. Econ. Lett. 86 (2), 199–203.

Benoît, J.-P., Dubra, J., 2011. Apparent overconfidence. Econometrica 79 (5), 1591–1625.

Benoît, J.-P., Dubra, J., Moore, D.A., 2015. Does the better-than-average effect show that people are overconfident?: two experiments. J. Eur. Econ. Assoc. 13 (2), 293–329.

Bertrand, M., Black, S.E., Jensen, S., Lleras-Muney, A., 2018. Breaking the glass ceiling? the effect of board quotas on female labour market outcomes in norway. Rev. Econ. Stud. 86 (1), 191–239.

Blau, F.D., Kahn, L.M., 2017. The gender wage gap: extent, trends, and explanations. J. Econ. Lit. 55 (3), 789–865.

Booth, A., Fan, E., Meng, X., Zhang, D., 2019. Gender differences in willingness to compete: the role of culture and institutions. Econ. J. 129 (618), 734–764.

Booth, A., Nolen, P., 2012. Choosing to compete: how different are girls and boys? J. Econ. Behav. Org. 81 (2), 542–555.

Bredemeier, C., 2019. Gender gaps in pay and inter-firm mobility. IZA Discussion Paper No. 12785.

Buser, T., 2016. How does the gender difference in willingness to compete evolve with experience? Mimeo.

Buser, T., Niederle, M., Oosterbeek, H., 2014. Gender, competitiveness, and career choices. Q. J. Econ. 129 (3), 1409–1447.

Cook, A., Glass, C., 2014. Women and top leadership positions: towards an institutional analysis. Gender Work Org. 21 (1), 91–103.

Crosetto, P., Filippin, A., 2013. The "bomb" risk elicitation task. J. Risk Uncertain. 47 (1), 31–65.

Czibor, E., Onderstal, S., Sloof, R., Van Praag, M., 2014. Does relative grading help male students? evidence from a field experiment in the classroom. IZA Working Papers No.8429.

Dahlbom, L., Jakobsson, A., Jakobsson, N., Kotsadam, A., 2011. Gender and overconfidence: are girls really overconfident? Appl. Econ. Lett. 18 (4), 325–327.

Datta Gupta, N., Poulsen, A., Villeval, M.C., 2013. Gender matching and competitiveness: experimental evidence. Econ. Inq. 51 (1), 816–835.

De Paola, M., Gioia, F., Scoppa, V., 2015. Are females scared of competing with males? results from a field experiment. Econ. Educ. Rev. 48, 117–128.

Deaves, R., Lüders, E., Luo, G.Y., 2009. An experimental test of the impact of overconfidence and gender on trading activity. Rev. Financ. 13 (3), 555–575.

Deloitte, 2019. Missing Pieces Report: The 2018 Board Diversity Census of Women and Minorities on Fortune 500 Boards. Technical Report. Alliance for Board Diversity.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., Wagner, G.G., 2011. Individual risk attitudes: measurement, determinants, and behavioral consequences. J. Eur. Econ. Assoc. 9 (3), 522–550.

Dreber, A., von Essen, E., Ranehill, E., 2014. Gender and competition in adolescence: task matters. Exp. Econ. 17 (1), 154–172.

European Commission, 2018. Report on equality between women and men in the EU 2018. Technical Report. European Commission.

Flory, J.A., Leibbrandt, A., List, J.A., 2014. Do competitive workplaces deter female workers? a large-scale natural field experiment on job entry decisions. Rev. Econ. Stud. 82 (1), 122–155.

Fortin, N.M., Bell, B., Böhm, M., 2017. Top earnings inequality and the gender pay gap: Canada, Sweden, and the United Kingdom. Labour Econ. 47, 107–123.

Frick, B., 2011. Gender differences in competitiveness: empirical evidence from professional distance running. Labour Econ. 18 (3), 389–398.

Gneezy, U., Leonard, K.L., List, J.A., 2009. Gender differences in competition: evidence from a matrilineal and a patriarchal society. Econometrica 77 (5), 1637–1664.

Gneezy, U., Niederle, M., Rustichini, A., 2003. Performance in competitive environments: gender differences. Q. J. Econ. 118 (3), 1049–1074.

Goldin, C., Katz, L.F., Kuziemko, I., 2006. The homecoming of american college women: the reversal of the college gender gap. J. Econ. Perspect. 20 (4), 133–156.

Große, N. D., Riener, G., 2010. Explaining gender differences in competitiveness: Gender-task stereotypes. Jena economic research papers, No. 2010,017.

Günther, C., Ekinci, N.A., Schwieren, C., Strobel, M., 2010. Women cant jump? an experiment on competitive attitudes and stereotype threat. J. Econ. Behav. Org. 75 (3), 395–401.

Harrison, G.W., List, J.A., 2004. Field experiments. J. Econ. Lit. 42 (4), 1009–1055.

Healy, A., Pate, J., 2011. Can teams help to close the gender competition gap? Econ. J. 121 (555), 1192–1204.

Hoelzl, E., Rustichini, A., 2005. Overconfident: do you put your money on it? Econ. J. 115 (503), 305–318.

Iriberri, N., Rey-Biel, P., 2019. Competitive pressure widens the gender gap in performance: evidence from a two-stage competition in mathematics. Econ. J. 129 (620), 1863–1893. doi:10.1111/ecoj.12617.

Jakobsson, N., 2012. Gender and confidence: are women underconfident? Appl. Econ. Lett. 19 (11), 1057–1059.

Jurajda, Š., Münich, D., 2011. Gender gap in performance under competitive pressure: admissions to czech universities. Am. Econ. Rev. 101 (3), 514–518.

Kamas, L., Preston, A., 2018. Competing with confidence: the ticket to labor market success for college-educated women. J. Econ. Behav. Org. 155, 231–252.

Krawczyk, M., Wilamowski, M., 2019. Task difficulty and overconfidence. evidence from distance running. J. Econ. Psychol. 75, 102128.

Kruger, J., 1999. Lake wobegon be gone! the "below-average effect" and the egocentric nature of comparative ability judgments.. J. Pers. Soc. Psychol. 77 (2), 221.

Lavy, V., 2013. Gender differences in market competitiveness in a real workplace: evidence from performance-based pay tournaments among teachers. Econ. J. 123 (569), 540–573.

Mobius, M. M., Niederle, M., Niehaus, P., Rosenblat, T. S., 2011. Managing self-confidence: theory and experimental evidence. NBER Working Paper No. 17014.

Moore, D.A., Cain, D.M., 2007. Overconfidence and underconfidence: when and why people underestimate (and overestimate) the competition. Organ. Behav. Hum. Decis. Process. 103 (2), 197–213.

Nekby, L., Skogman Thoursie, P., Vahtrik, L., 2015. Gender differences in examination behavior. Econ. Inq. 53 (1), 352–364.

Niederle, M., Segal, C., Vesterlund, L., 2013. How costly is diversity? affirmative action in light of gender differences in competitiveness. Manage. Sci. 59 (1), 1–16.

Niederle, M., Vesterlund, L., 2007. Do women shy away from competition? do men compete too much? Q. J. Econ. 122 (3), 1067–1101.

Niederle, M., Vesterlund, L., 2011. Gender and competition. Annu. Rev. Econom. 3 (1), 601–630.

Ors, E., Palomino, F., Peyrache, E., 2013. Performance gender gap: does competition matter? J. Labor Econ. 31 (3), 443–499.

Pekkarinen, T., 2015. Gender differences in behaviour under competitive pressure: evidence on omission patterns in university entrance examinations. J. Econ. Behav. Org. 115, 94–110.

Perloff, L.S., Fetzer, B.K., 1986. Self–other judgments and perceived vulnerability to victimization.. J. Pers. Soc. Psychol. 50 (3), 502.

Petrongolo, B., 2019. The gender gap in employment and wages. Nat. Hum. Behav. 3 (4), 316.

Pindyck, R., Rubinfeld, D., 2013. Microeconomics, 8th edition. Prentice Hall.

Reuben, E., Sapienza, P., Zingales, L., 2015. Taste for competition and the gender gap among young business professionals. NBER Working Paper No. 21695.

Reuben, E., Wiswall, M., Zafar, B., 2017. Preferences and biases in educational choices and labour market expectations: shrinking the black box of gender. Econ. J. 127 (604), 2153–2186.

Roberts, T.-A., Nolen-Hoeksema, S., 1989. Sex differences in reactions to evaluative feedback. Sex Roles 21 (11–12), 725–747.

Saccardo, S., Pietrasz, A., Gneezy, U., 2017. On the size of the gender difference in competitiveness. Manage. Sci. 64 (4), 1541–1554.

Samek, A., 2019. Gender differences in job entry decisions: auniversity-wide field experiment. Manage. Sci. 65 (7), 3272–3281.

Shurchkov, O., 2012. Under pressure: gender differences in output quality and quantity under competition and time constraints. J. Eur. Econ. Assoc. 10 (5), 1189–1213.

Windschitl, P.D., Kruger, J., Simms, E.N., 2003. The influence of egocentrism and focalism on people's optimism in competitions: when what affects us equally affects me more.. J. Pers. Soc. Psychol. 85 (3), 389.

Zhang, Y.J., 2019. Culture, institutions and the gender gap in competitive inclination: evidence from the communist experiment in china. Econ. J. 129 (617), 509–552.