

Competition, confidence and gender: shifting the focus from the overconfident to the realistic *

Tünde Lénárd §, Dániel Horn,¶, Hubert János Kiss||

2023 November

Abstract

The gender gap in competitiveness is argued to explain gender differences in later life outcomes, including career choices and the gender wage gap. In experimental settings, a prevalent explanation attributes this gap to males being more (over)confident than females (we call this the compositional channel). While our lab-in-the-field study using data from students in 53 classrooms ($N > 1000$) reproduces this finding, it also uncovers a second, potentially more impactful channel of confidence contributing to the gender gap in competitiveness (the preference channel). To disentangle the two channels, we propose a more precise measure of confidence based on whether the subjects' believed performance rank exceeds, coincides with or falls short of their actual performance in a real-effort task. We label categories of this Guessed - Actual Performance (GAP) difference as overconfident, realistic or underconfident, respectively. Surprisingly, there is no gender difference in competitiveness within the over- and underconfident subgroups, while a significant gender gap exists among the realistic. So, even if both genders had the same level of confidence, a persistent gender gap in preference (or taste) for competition would remain in the realistic group. This finding is robust across all specifications, challenging previous theories about the overconfidence of men being the sole driver of the relationship between confidence and the gender gap in competition.

JEL codes: C9, D91, J16

Keywords: adolescents, competitiveness, confidence, gender, experiment

*The experimental project received funding from the National Research and Development Office of Hungary (project no. 143415).

§SOFI, Stockholm University (Stockholm, SE-106 91, Sweden) and KRTK KTI. E-mail: tunde.lenard@sofi.su.se. Financial support from Forte (grant number: 2016-07099) and VR (grant number: 2022-02036) is gratefully acknowledged.

¶Corresponding author. HUN-REN Centre for Economic and Regional Studies - Institute of Economics and Corvinus University of Budapest. E-mail: horn.daniel@krtk.hun-ren.hu. Financial support from the Bolyai János research scholarship and the NKFIH-K 143415 grant is gratefully acknowledged.

||HUN-REN Centre for Economic and Regional Studies - Institute of Economics and Corvinus University of Budapest. E-mail: kiss.hubert@krtk.hun-ren.hu. Financial support from the Hungarian Academy of Sciences is gratefully acknowledged (Momentum Grant No. LP2021-2).

1 Introduction

There is a growing body of research showing that men tend to be more competitive than women (Markowsky and Beblo, 2022; Niederle, 2016). This phenomenon has received considerable attention as it provides a potential explanation for gender differences in career choices (Buser et al., 2014) and labor market outcomes (Goldin et al., 2006; Bertrand et al., 2010), beyond existing explanations like discrimination, differences in preferences over jobs, or differences in ability (Polachek, 1981; Cain, 1986; Goldin and Rouse, 2000). The presence of this gender gap results in a societal loss, because high-performing women compete too little and less well-performing men compete too much, and consequently women are less likely to pursue promotions (Babcock and Laschever, 2009) or enter competitive jobs (Bertrand and Hallock, 2001; Flory et al., 2015). Therefore, it is natural to investigate the mechanisms that contribute to the disparity in competitive preferences.

In experimental settings, it is a well-established result that performance, confidence, and risk preferences contribute to the gender gap in competitiveness. Most experimental studies estimate that, on average, about a third of the gender gap is attributable to these factors (Van Veldhuizen, 2022; Markowsky and Beblo, 2022), but the variance in the estimates is considerable. Even though there is no agreement regarding the exact extent of the contribution of each factor to the gender gap, the nonzero contribution of confidence as a mediator has been undisputed to date.¹ Yet, we know little about the mechanisms behind the relationship between confidence and gender differences in the willingness to compete.

The lack of knowledge regarding the underlying mechanisms is not surprising, given the uncertainty surrounding how to measure confidence in experimental studies. This uncertainty can be traced back to the seminal paper by Niederle and Vesterlund (2007), who were the first to establish that gender differences in confidence significantly contribute to the gender gap in competitiveness. They used participants' believed performance rank in a real-effort task as a measure of confidence, and as a mediator of the gender gap. Numerous subsequent experimental studies, building upon their design, have largely adopted the same approach: using a measure of believed performance as a proxy for confidence, often interchangeably referred to as confidence or overconfidence.² These studies mostly conclude that higher confidence (usually a characteristic of men) is associated with higher levels of competitiveness (Buser et al., 2014; Dreber et al., 2014; Niederle and Vesterlund, 2007), speaking to the popular belief that overconfident men drive the gender gap in competition.

While the approach of using believed performance as a measure of confidence has its merits, it can be misleading according to the definitions in the psychological literature, as believed performance alone does not necessarily indicate overconfidence without an

¹For instance, Van Veldhuizen (2022) emphasizes the relevance of risk preferences, Lozano and Reuben (2022) focus more on confidence, and Gillen et al. (2019) consider both factors important.

²See summaries and meta-analyses in Horn et al. (2022a); Markowsky and Beblo (2022); Van Veldhuizen (2022).

accuracy benchmark against which guesses can be compared (Moore and Schatz, 2017). Studies that actually compare beliefs with actual performance distinguish multiple forms of overconfidence. Among these forms, the relevant one that can be applied within the Niederle-Vesterlund design framework is overplacement: the exaggerated belief that one performs better than others (Moore and Dev, 2020; Moore and Healy, 2008). According to Moore and Healy (2008), overplacement is optimally measured by calculating the difference between one’s believed relative performance rank and the actual relative performance rank in a specific task.

In this study, we aim to make two contributions. To date, there are no studies we know of that investigate the mechanisms behind the relationship between confidence and the gender gap in competition while accurately measuring overplacement in the Niederle-Vesterlund framework. To bridge this knowledge gap, our first contribution is to propose a more precise measure for over-, underconfident, or realistic assessments of one’s own performance in the form of overplacement. More precisely, we differentiate participants as overconfident, underconfident, or realistic based on whether their believed performance rank exceeds, falls short of or coincides with their actual performance rank. We refer to this measure as the ‘Guessed-Actual Performance Difference,’ or, in short, GAP. This distinction is key for our second aim which is to disentangle two channels through which confidence might contribute to the gender gap formation in competitiveness: a compositional channel and a preference-for-competition channel (conditional on confidence). In this analysis, we do not only explore how confidence mediates the gender gap in competition (as it has been customary in the literature) but we also analyze potential moderation. This is crucial since any results from a mediation analysis only have limited implications regarding the gender gap formation in competition. They only isolate part of the mechanisms running through the compositional channel, showing that if males are on average more (over)confident, they compete more than females. But what if the distribution of males and females according to confidence were similar? Would we see similar levels of competitiveness too? We use moderation techniques to answer these questions. More precisely, we explore whether any gender gap in competitiveness arises among the over- /underconfident or realistic individuals by comparing females and males in the same categories of GAP.

We utilize experimental data collected in a series of lab-in-the-field experimental sessions in Hungarian high schools, and linked to administrative data on students’ school performance and family background (Horn et al., 2022b). We rely on fixed effect models in our mediation and moderation analyses, and use matching techniques to check the robustness of our moderation results. We also check if our findings hold in a dataset different from ours.

We find a significant gender gap of 11 percentage points in competitiveness, in favor of male students. Consistent with previous findings, this difference is mediated by confidence using both the traditional measure (believed performance) and our categorical measure of GAP (both of which are higher for males in our sample). That is, participants with a higher believed performance, and participants whose believed performance

exceeds their actual one, are more likely to enter competition. However, our moderation analysis reveals an interesting result: there is no gender difference in competitiveness among the overconfident and underconfident groups. In contrast, we find a substantial gender gap among individuals who accurately evaluate their performance. Specifically, female students in the realistic group are 14 percentage points less likely to compete compared to their male counterparts. This finding is robust across all specifications and indicates that - contrary to previous results - the gender gap in competitiveness is not entirely driven by the compositional effect imposed by more overconfident men, but there is another channel, namely a considerable difference in "taste" or preference for competition among females and males who have an accurate understanding of their performance relative to others.

2 Literature review

The computer-based laboratory experiment designed by Niederle and Vesterlund (2007) aims to test gender differences in competitive preferences and allows to eliminate the potential mediating effects of gender differences in beliefs, risk aversion or other-regarding preferences. The Niederle-Vesterlund (henceforth, NV) setup has been widely used in many studies, most of them replicating a robust gender difference in attitudes toward competition, with females generally exhibiting lower competitiveness (Niederle, 2016; Van Veldhuizen, 2022).

The NV setup consists of 3 rounds. In each round, participants have to carry out a real-effort task for a predetermined period of time, and based on their performance, they earn a certain payout. The first round involves a piece-rate payout scheme, where the payout is determined by the number of tasks correctly solved in the underlying real-effort task. The second round employs a competitive payment method, where only the top-performing quartile of participants receive any payout. In the third round, participants are given the option to choose between the two payment schemes, which indicates their competitiveness.

Since our subject pool consists of high school students, here we review briefly what we know about the gender differences in competitiveness within this group. The overwhelming majority of experimental studies consistently report a significant gender gap in competitive attitudes (Booth and Nolen, 2012; Sutter et al., 2016). The difference is substantial, as highlighted in Buser et al. (2014) where 15 year-old male students exhibit a 15.8 percentage points higher probability to enter competition than females, even after controlling for actual performance in the real-effort task. Similar differences have been reported by Sutter and Glätzle-Rützler (2015), Almås et al. (2016), and Sutter et al. (2016). There is no clear finding about *when* the gender difference emerges (Sutter et al., 2019), but Sutter and Glätzle-Rützler (2015) suggest that they may even grow in adolescence. Additionally, socioeconomic status (SES) potentially has a role as Almås et al. (2016) find that: i) low-SES male students are less willing to compete than their counterparts from better-off families, ii) there is a gender gap in competitiveness in high-SES

families, but not in low-SES ones. Dreber et al. (2014) provide suggestive evidence that the real-effort task employed may also influence competitiveness as females are as likely to compete as males in tasks where they perform equally well. Other studies hint at the possibility that there may be cultural factors behind the gender gap in competitive preferences (Zhang, 2011; Cárdenas et al., 2012; Khachatryan et al., 2015).

As previously mentioned, several factors typically measured in experiments influence competition entry, such as actual performance, confidence (or believed performance), and risk-taking. Most studies analyzing gender differences in competitive preferences examine how these three factors mediate the gender gap. The analytical strategy is usually built on a regression design, where potential mediators are gradually entered into a model containing a gender dummy (Van Veldhuizen, 2022). The residual gender gap, after controlling for all included variables, is interpreted as the net gender difference in competitive preferences. It is usually estimated that the residual gap is around 70% of the raw gender gap without controls (meaning that performance, beliefs, and risk-taking explain around 30% of the raw gap, see the meta-analyses of Van Veldhuizen (2022) and Markowsky and Beblo (2022)). Markowsky and Beblo (2022) finds that the explained part in studies closely following the NV setup is slightly higher, on average 38%

Studies typically use the actual and believed performance from the tournament round (round 2) in the mediation analysis because that round provides participants with information on how they perform under competitive circumstances. In the NV-setup, participants usually lack direct knowledge about their actual performance relative to others (they only know their absolute performance without comparison), but they may form beliefs about it, which can further influence their competition choice. The vast majority of studies use their measure of believed performance in the tournament round as a proxy for confidence (often referred to as guessed rank, or guessed chance to win the tournament round, see (Markowsky and Beblo, 2022)).³ Only three studies employ the difference between guessed and actual performance as a confidence measure, but solely as a continuous control without further analysis into any mechanisms (Almås et al., 2016; Gillen et al., 2019; Zhang, 2019).⁴ These studies find a zero or positive association between confidence and competitiveness. Unless otherwise stated, we refer to (actual and believed) performance and confidence as variables measured in the tournament round when talking about how they influence competitiveness.

It is generally found that both higher actual (Booth and Nolen, 2012; Buser et al., 2014; Dreber et al., 2014; Niederle and Vesterlund, 2007; Sutter and Glätzle-Rützler, 2015) and higher believed (Buser et al., 2014; Dreber et al., 2014; Niederle and Vesterlund, 2007) performance are associated with higher competition entry. The relationship between performance and competitiveness seems to be stronger for men compared to women. In the case of women, better actual performance is associated with a less significant increase in the likelihood of entering a tournament, and the largest gender gap

³Buser et al. (2021) explain that guessed rank and guessed chance to win can be translated into each other, which makes the two measures practically identical.

⁴Gillen et al. (2019) only use this variable to control for a potential measurement error in the guessed rank.

is often observed among the best-performing female participants (Sutter and Glätzle-Rützler, 2015). The positive relationship between believed performance and tournament entry is also more pronounced for men than for women (Niederle and Vesterlund, 2007; Almås et al., 2016). Furthermore, it has also been documented that males tend to have significantly higher believed performance compared to females (Buser et al., 2014; Dreber et al., 2014; Sutter and Glätzle-Rützler, 2015).⁵

In this paper, we take an extra step from barely using believed performance as a measure of confidence and define confidence based on whether the believed relative performance rank is higher, equal or lower than the actual performance rank (see a detailed description in the *Data* section) and explore if the gender gap in competition is mediated and moderated by this factor.

Lastly, risk aversion is another important factor that influences competition entry. The literature indicates that females tend to be more risk averse than males (Buser et al., 2014; Dreber et al., 2014; Khachatryan et al., 2015), which may contribute to the gender gap to compete. But even after controlling for the gender differences in performance, beliefs and risk aversion, a significant residual gender gap in competition usually persists (Buser et al., 2014; Sutter and Glätzle-Rützler, 2015; Almås et al., 2016; Van Veldhuizen, 2022). However, some studies find non-significant residual gaps (Gillen et al., 2019; Dreber et al., 2014; Zhang, 2019).

3 Data

Between March 2019 and March 2020, we conducted a series of incentivized lab-in-the-field experiments to measure the time, risk, social and competitive preferences of 1108 Hungarian high school students. These students were enrolled in 53 school classes, which are groups of students who study most subjects together and spend a significant amount of time at school together. We visited a total of 9 schools and collected data from entire classrooms. Our data was then linked to the database of the National Assessment of Basic Competences (NABC), which provided us with background information on the students' standardized test scores, school grades, and socioeconomic status (Balázs and Ostorics, 2020).

3.1 Experimental procedures

Before starting the project, we contacted all education providers in Hungary that maintained at least one secondary school to request permission to run the experiment in their institutions. The schools that ultimately participated either volunteered after

⁵There are instances when the piece-rate performance is associated with competitive preferences. Almås et al. (2016) show that this association is positive, while Niederle and Vesterlund (2007) find that higher actual performance in the first round means higher competitiveness only for males. Using the believed performance in the piece-rate round, Almås et al. (2016) find a less consistent association, with only males showing a positive and linear relationship between beliefs and competitiveness. Additionally, Niederle and Vesterlund (2007) report a less pronounced gender gap in these beliefs compared to believed tournament performance.

learning about our request from their maintainer or were referred to us by the educational provider. We sent a data protection statement to the students and their parents in these schools, assuring them that the experiment was voluntary.

We provided laptops that we unpacked in a designated classroom on experiment day. The experiments were conducted using the z-Tree software (Fischbacher, 2007). Since we measured whole classes, with participants in each session being familiar with each other, we adjusted the beginning of the experiments to the beginning of the school lessons. This allowed the groups to take turns approximately every hour. In Hungarian schools, there are 45-minute lessons and 10-15-minute breaks, which were our only time constraints.

When entering the room, students were free to sit wherever they wanted. First, one researcher explained the rules of the experiment. Students could also read these on the first screen of the experiment. Participants were asked to engage in 8 tasks, some of them involving multiple decisions or even some interaction with their peers in the classroom. The experiment was incentivized. Everyone who completed a session received meal vouchers that could be used in the school cafeteria as cash. Once all participants finished the tasks, the program randomly selected one task (specifically, one round if the task involved multiple rounds of decisions), and we distributed vouchers based on the decisions made in this particular task/round. We emphasized that the payoff-relevant task would be the same for everyone in the classroom. The experiment was designed so that the expected value of payoffs was 1000 HUF (approximately 3 EUR), equivalent to the price of a full meal in an average school cafeteria at that time. No additional show-up fee was provided.

We measured time and risk preferences using individual tasks, where the payoffs solely depended on individual decisions. However, for assessing social preferences, we employed tasks that involved strategic interactions, where the payoffs were determined by the decisions of two participants. We used z-Tree (Fischbacher, 2007) to randomly create student pairs at the end of the experiment, after collecting information about each student’s decision in each situation.⁶ This procedure was also explained at the beginning of each session.

3.2 Experimental tasks

All participants made decisions in 8 tasks in a given order. Tasks 1 and 6 were designed to measure time preferences and involved 5 interdependent choices between earlier and later amounts of money. Tasks 2 and 3 measured altruism using the dictator game, where participants made decisions towards a classmate in task 2 and a random schoolmate in task 3. Risk preferences were assessed in task 4 using the bomb risk elicitation task (Crosetto and Filippin, 2013). Task 5 utilized a two-person variant of the public goods

⁶For example, in the two-person public goods game, each participant had to make a choice regarding how much they wanted to contribute to the common project. If this decision was selected for payment, at the end of the experiment, z-Tree randomly paired participants together, and the payoffs were calculated based on their decisions.

game to measure cooperation. Trust and trustworthiness were measured in task 7 using the investment game by Berg et al. (1995). Finally, task 8 was designed to assess the competitiveness of the participants.⁷ Before the last task, we did not provide feedback to participants after any of the tasks to prevent the possibility that the result of one task might affect their decision in subsequent tasks.

To measure competitiveness, we used the standard NV-setup, but instead of adding up numbers we chose a different real-effort task.⁸ Participants were presented with 5x5 matrices containing zeros and ones, and had to count the zeros in them.⁹ In each stage of the game, they had 1 minute to complete as many matrices as they could (see the zTree instructions in Appendix A).

The task started with the piece-rate stage in which participants were paid 100 HUF (~ 0.3 EUR/USD at the time of the experiment) for each correctly completed matrix. Stage 2 involved a tournament in which only the best-performing quartile of the classroom was rewarded for the counting exercise.¹⁰ However, participants in the best-performing quartile earned 4 times more per matrix solved compared to stage 1. After both stage 1 and stage 2, we provided participants with feedback regarding their absolute performance (i.e., the number of correctly completed matrices) and their (potential) earnings. However, they were unaware of their performance relative to others. In stage 3, participants had the option to either receive compensation based on the piece-rate scheme (as in stage 1) or to enter the tournament and receive payment as in stage 2.

As is customary in experiments that use the NV-setup, we informed participants that if they chose the tournament-based payment scheme, their performance would be compared to that of others in round 2 (that is, in the tournament stage). This practice prevents unwanted strategic considerations, such as participants choosing the tournament based on their beliefs about others' choices.¹¹ By allowing participants to compete against predetermined tournament scores, we ensure that competitive preferences are not influenced by the choices of others but rather by participants' beliefs about their relative performance (informed by their performance in round 2) and their preference for competition.

After stage 3, we asked participants to rank themselves according to their believed performance in stage 1 and 2 relative to their classmates. The ranking options were: quartile 1, 2, 3 or 4, where quartile 1 meant being among the best performing students. To obtain considered answers, this belief was incentivized by paying 300 HUF (cca 1 EUR) to those who guessed their performance correctly (but only if the final payout was based on one of the first two rounds of the competition game).

⁷See a more detailed description of the tasks in Horn et al. (2022b), including our reasons to choose the order of tasks.

⁸We are very grateful to Lise Vesterlund who shared their z-Tree code with us.

⁹Among others, Abeler et al. (2011) use the same task.

¹⁰Niederle and Vesterlund (2007) used groups of four during the competition, and only the best performer was paid.

¹¹For instance, someone would (not) choose the tournament because she thinks she has a good chance of outperforming those who decide to compete in the third round. Conversely, someone might opt out of the competition because she believes that if she participates and wins, others receive lower rewards.

3.3 Measure of confidence: categories of the GuesSED-Actual Performance difference (GAP)

To obtain our categorical measure of confidence, we take the difference between the believed (or guessed) and the actual performance ranks in the tournament stage. We classify students as overconfident, realistic or underconfident based on whether this difference is positive (believed $>$ actual performance), zero (believed = actual performance) or negative (believed $<$ actual performance), respectively. To facilitate easier description, we will refer to this classification as categories of GAP. GAP measures confidence in the form of overplacement of own performance relative to others by taking the difference between two relative performance ranks.

3.4 Control variables

We have data on age, gender, family background, and academic performance of the participants from the NABC database. The variables related to family background include parents' education and the father's employment status. Regarding academic performance, we have information i) on the standardized mathematics and reading test scores measured in grade 6 (around age 12), and ii) on teacher-given grades (including GPA) from grade 6.

Information on age and gender is nearly complete. However, for approximately 16% of the participants, family background information is missing, and for around 24% of the participants, school grades are missing, because these were self-reported in the NABC questionnaire.

Categorical family background variables have been converted into dummy variables, with a separate category for missing values. As to grades, missing values are imputed with the sample mean, and a separate missing dummy variable has been included to account for these missing values.

4 Results

4.1 Descriptive analysis and balance tests

First, we conduct a balance test by gender, which includes participants' family background, test scores, grades, and the experimental variables, such as performance in the the piece-rate and tournament games, believed performance (guessed rank), categories of GAP, and preferences towards risk and competition. Table 7 in Appendix B contains the descriptive statistics and results of the balance test.¹²

¹²We test randomization of these variables by gender using within-class balance tests. For each variable listed in the first column of Table 7, we perform separate regressions with the variable as the dependent variable and the female dummy variable as the independent variable. We control for classroom fixed effects throughout the analysis. The coefficients of the female dummy variable from these regressions are reported in the balance test column, with statistically significant differences implying imbalance.

The balance tests indicate that, conditional on class fixed effects, female participants' parents are more likely to have a medium-level education compared to males' parents, but not less or more educated. Additionally, fathers of female participants are more likely to be self-employed than fathers of male participants. Even though the average test scores in mathematics and reading are higher for males, the within-class differences from the balance tests show a slightly different picture. Male participants have higher mathematics test scores, but female participants have higher GPA within the class, which is likely driven by their higher grades in Hungarian literature and grammar. Regarding school performance, the within-class gender differences resemble the patterns observed in the total population.¹³ The inclusion of class fixed effects makes a difference because there is considerable sorting between classrooms in Hungarian schools. Thus, we use class fixed effects and control for family background variables throughout our analyses (but also present robustness checks by dropping them). Using class fixed effects also accounts for the fact that participants were competing against their classmates.

Gender differences are also evident in the experimental measures (see all measures mentioned in this paragraph in Table 7 in Appendix B). Males outperform females in both rounds of the competition game and also have a higher believed performance (measured by the guessed quartile, where 1 is the best) as more males than females believe to be in the upper quartiles. Table 1 provides a detailed breakdown of the number of participants and share of females by actual and believed performance rank.¹⁴ Males are also more risk-taking than females. Roughly 56% of females and 66% of males choose to compete in round 3, so we observe an 11 percentage points gender gap in competitiveness within classes.

Table 1: Share of females (expressed on a 0-1 scale) and number of participants within actual-believed performance cells

Actual rank (quartiles)	Guessed rank				Total
	1	2	3	4	
1	0.32	0.60	0.78	0.83	0.50
	142	124	41	6	313
2	0.30	0.61	0.64	0.61	0.53
	81	136	72	18	307
3	0.39	0.61	0.69	0.58	0.60
	31	120	72	24	247
4	0.53	0.61	0.61	0.77	0.64
	15	72	90	44	221
Total	0.33	0.61	0.67	0.70	0.56
	269	452	275	92	1088

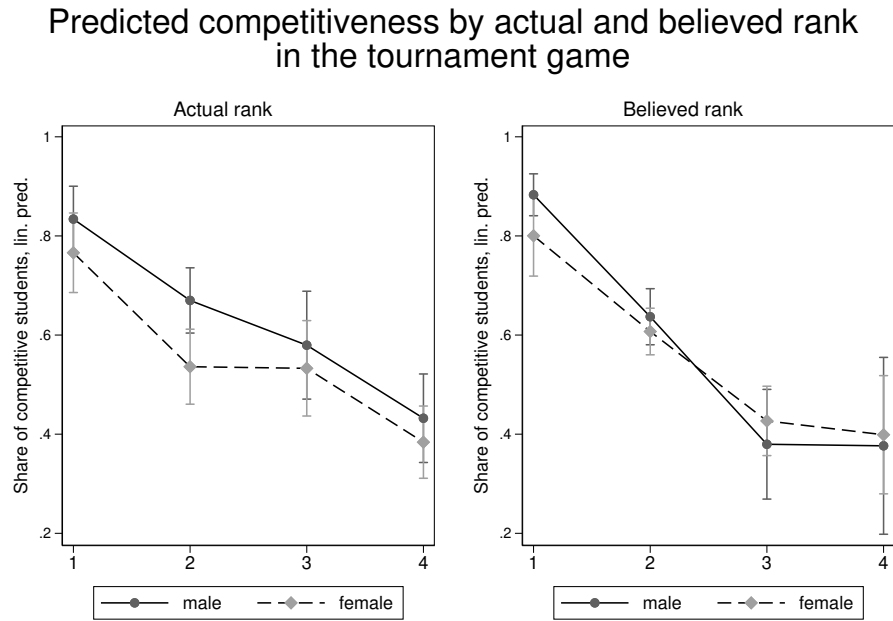
Note: share of females in each cell shows the proportion of students in the cell that is female (in cell [1,1] 32% of the 142 students is female). Cells in the *Total* columns show the proportion of females out of those students who were ranked 1st/2nd/3rd/4th according to either their actual or guessed performance.

¹³Female participants in Hungary perform better than males in grammar and literature, and males perform better in mathematics. For details, see Hajdu et al. (2022)

¹⁴Based on the breakdown in Table 8 in Appendix B, 38% / 37% / 19% / 6% of males believed to be in quartile 1 / 2 / 3 / 4 respectively, compared to a corresponding 15% / 45% / 30% / 10% of females. The Kolmogorov-Smirnov test indicates that the distribution of the believed performance differs significantly by gender (p-value<0.0001).

Males not only have better average performance but also outperform females in every performance quartile, both in terms of actual and believed performance (see Table 8 in Appendix B). However, while a slightly larger proportion of males compete in each quartile based on actual performance, based on believed performance, males only compete more in the top two quartiles. These within-quartile gender differences in competition are not statistically significant, as indicated in Figure 1 which plots females' and males' predicted competitiveness by actual and believed performance quartiles, conditional on classroom fixed effects. However, both higher actual and believed performance are clearly associated with a greater willingness to compete in both gender groups.

Figure 1: Competitiveness by gender and by actual and believed performance in the tournament game



Our descriptive results are mostly in line with the literature. Males perform better, have higher believed performance and compete more in the competition task, and there is a positive association between actual / believed performance and competitiveness for both genders. Although there is a slightly larger gender difference in competitiveness among the better performing students, we cannot observe any significant gender gap in the willingness to compete based on performance quartiles.

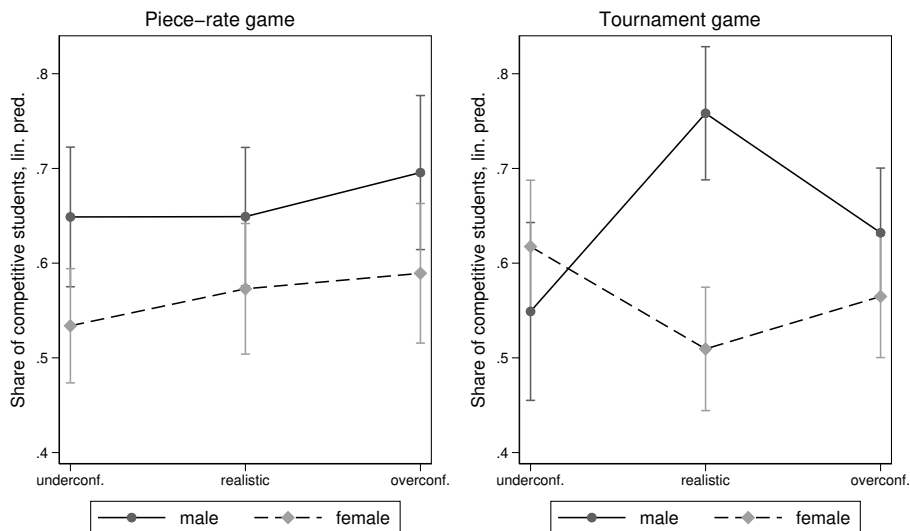
Moving on to statistics on categories of GAP: our sample, on average, exhibits slight overconfidence. Among all participants, 26% are underconfident, 36% are realistic, and 37% are overconfident. A larger proportion of females are underconfident compared to males (30% vs 21%, respectively), and a larger proportion of males are overconfident relative to females (40% vs 35%, respectively, see Table 7 in Appendix B). Consequently,

on average, males are more overconfident than females, as they tend to overestimate their relative performance to a greater extent.

Note that while the relationship between actual (believed) performance and competition entry is intuitive (higher actual (believed) performance should go hand in hand with higher tournament entry), and this relationship is supported by the data, it is not clear how our confidence classification is related to the willingness to compete. An overconfident participant has a high believed performance (which should encourage tournament entry) relative to a lower actual performance (which discourages competition entry), resulting in unclear behavior. The opposite is true for underconfident participants. On the other hand, realistic participants have their believed and actual performances aligned, so both performance measures work in the same direction. Testing the moderating role of confidence (as categories of GAP) might shed some light on the underlying mechanisms.

In Figure 2, we present a plot of predicted competitiveness by categories of GAP using LPM models that describe competition entry as a function of gender and confidence while controlling for class fixed effects. The dependent variable in these models is a binary variable where 1 corresponds to students choosing to compete in the 3rd round, and 0 corresponds to students opting not to compete. Our analysis reveals that the only statistically significant gender gap in competition is between males and females who evaluate their tournament performance realistically. These findings confirm that, when examining the factors driving the gender gap in competition, it is crucial to focus on categories of GAP, rather than relying solely on actual and believed performance.

Figure 2: Competitiveness by categories of GAP and gender
Predicted competitiveness by categories of GAP



Note: predictions from a linear probability model, 95% CI. On the first graph, categories of GAP are measured based on the actual and believed ranks in the piece-rate game. This is only for comparison purposes, GAP is measured based on the tournament stage everywhere else in the paper.

4.2 Mediation

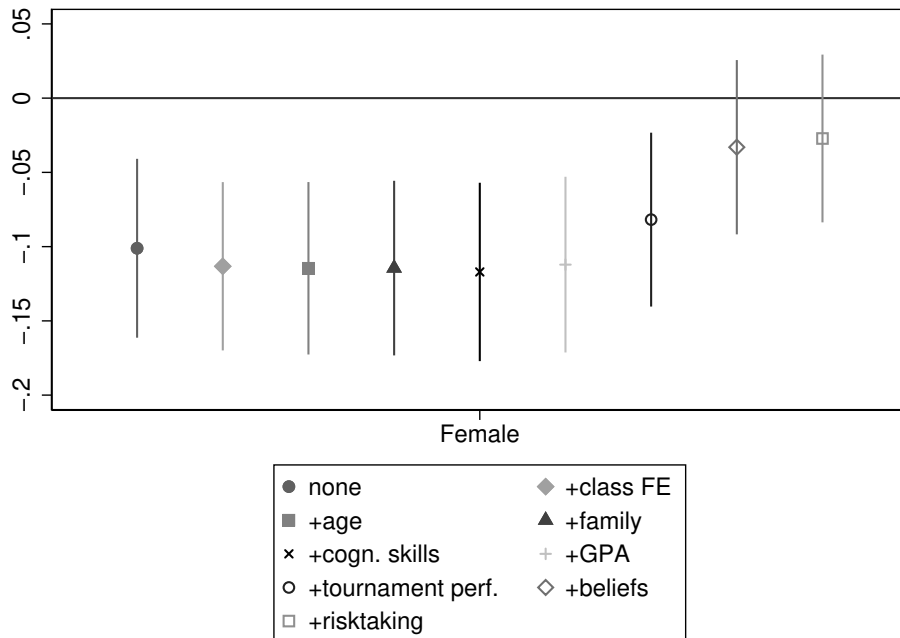
First, we test if confidence mediates the competitiveness gap by using believed performance as a proxy for confidence, then we proceed by testing categories of GAP as a mediator.

The analysis builds on regression models and follows the analytical strategy widely used in the literature. We run a linear probability model to explain the competitiveness dummy in subsequent specifications. In each step, we introduce an additional control to see how the control impacts the gender gap in competition. Figure 3 displays the coefficient of the female dummy across different specifications, facilitating easy comparison. We use believed performance as a mediator in this analysis.

The first coefficient comes from a bivariate regression, where the only independent variable is the female dummy. This coefficient reveals the raw gender gap in competition, which amounts to 10 percentage points. In the second model, class fixed effects are included, resulting in a gap of 11 percentage points. These fixed effects are retained in all subsequent models. Legends indicate the extra controls added at each step (in addition to those incorporated in previous steps).

After adding classroom fixed effects, the gender gap remains largely unchanged, even when accounting for age, family background (parental education), cognitive skills (standardized national test scores in math and reading), and GPA.

Figure 3: Female dummy coefficients from the mediation analysis

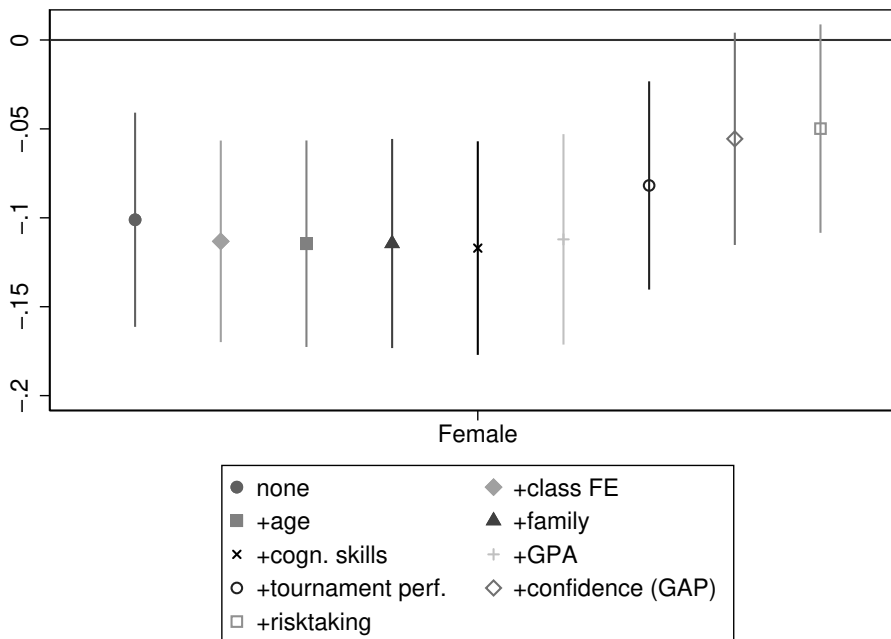


However, when controlling for actual performance in the tournament game, the gen-

der gap decreases from 11 percentage points to 8.2 percentage points. Furthermore, when believed performance is added as a control variable, the gender difference becomes statistically non-significant, with a point estimate of 3.3 percentage points (see Table 9 in Appendix B for detailed regression results). Risk preferences do not change the gap much after beliefs have been controlled for. These results are robust to dropping controls for family background, cognitive skills and GPA from the analysis (see Figure 13 in Appendix B).

We repeat the mediation analysis using categories of GAP as a mediator instead of believed performance. This analysis leads to very similar results, as depicted in Figure 4 (and the regression output in Table 10 of Appendix B presents further information.) After controlling for categories of GAP, the gender gap in competition becomes statistically non-significant at the 5% significance level, with a point estimate of 5.6 percentage points. Similarly to the results from the first mediation analysis, these findings are also robust to dropping controls for family background, cognitive skills and GPA from the analysis (see Figure 14 in Appendix B).

Figure 4: Female dummy coefficients from the second version of the mediation analysis using categories of GAP as a mediator



Our analysis confirms that confidence, in both its forms (used as pure beliefs or as categories of GAP), serves as a significant mediator of the gender gap in competition, in line with the findings of the literature. After controlling for actual performance, beliefs/categories of GAP and risk aversion, the gender gap decreases to 25% and 45% of the initial within-class gap in our two specifications. These residual gap sizes are in

the lower half of the residual gap distribution based on comparable studies (as depicted in Figure 1 of Van Veldhuizen (2022)), and they are close to the result of the original study by Niederle and Vesterlund (2007).

As shown in Table 10 of Appendix B, higher level of GAP have a statistically significant positive effect on competitiveness. Realistic / overconfident participants are around 11 / 26 percentage points more likely to enter competition compared to underconfident participants, *ceteris paribus*. Thus, if males are more overconfident than females (which they are), it contributes to widening the gender gap in competition. This represents the compositional channel.

4.3 Moderation analysis

4.3.1 Regression-based approach

Our moderation analysis essentially compares female and male participants who are in the same category of GAP. The purpose of this analysis is to provide further insight into the mechanisms underlying the relationship between confidence and the competitiveness gap while controlling for any variations in GAP between gender groups.

We use two different models, the first of which is an interaction model. We regress competitiveness on the interaction of the female dummy and the categories of GAP, controlling for the same variables as in the last specification of the mediation analysis (class FE, age, family background, cognitive skills and GPA, tournament performance and risk aversion). In the first column of Table 2, we report the marginal effects of being female by categories of GAP, conditional on performance and risk aversion.

As a comparison, we run separate regressions within subgroups based on different categories of GAP. In these regressions, we control for the same variables as before (class fixed effects, age, family background, cognitive skills, GPA, tournament performance, and risk aversion). We then report the female dummy coefficients from these three models in columns 2-4.¹⁵

Both the interaction model and the subgroup analysis yields the same result: among participants who evaluate their performance realistically, we observe a gender gap of 13.9 percentage points in competitiveness. However, in the underconfident and overconfident groups, we do not find a significant gap in competitiveness. These findings are not sensitive to excluding any controls from the models (see Figures 15, 16 and 17 in Appendix B illustrating how the gender gaps change within categories of GAP when (not) including different sets of controls in the subgroup-based models. The interaction-based results are also robust to dropping controls, output is available upon request).

Our moderation analysis suggests the presence of another important channel through which confidence might widen the gender gap in competition: the preference-for-competition channel, more precisely the gender difference in the willingness to compete among realistic participants.

¹⁵Running regressions on a split sample allows the confidence variable to interact with not only the female dummy variable but also with every control variable included in the analysis.

Table 2: Results from the regression-based moderation analysis

	Interaction model	Underconf. subgroup	Realistic subgroup	Overconf. subgroup
<i>Coefficient:</i>				
female		0.093 (0.085)	-0.139*** (0.043)	-0.028 (0.066)
<i>Margins of female:</i>				
Underconf.	0.087 (0.070)			
Realistic	-0.139*** (0.042)			
Overconf.	-0.056 (0.057)			
<i>N</i>	1073	283	388	402
<i>R</i> ²		0.245	0.393	0.245
Robust standard errors in parentheses				
*** p<0.001, ** p<0.01, * p<0.05				

Note: *Interaction model*: the female dummy is interacted with categories of GAP. Marginal female effects are reported by confidence group. *Subgroup models*: a separate regression is run for each confidence group. Female dummy coefficients are reported. In all four columns classroom FE, age, family background, test scores, GPA, real performance in the tournament, and risk preferences are controlled for.

4.3.2 Robustness checks using matching

In this section, we use matching to create sufficiently similar gender groups in order to test the effect of gender on competitiveness and perform moderation analysis. Compared to regression models, matching provides a more flexible way of controlling for various factors without requiring specific assumptions about functional forms.

We use kernel-based propensity score (PS) matching with regression adjustment, making our estimations doubly robust. Doubly robust estimators essentially combine two different models: one for estimating the exposure to treatment (PS model to estimate propensity scores or weights), and another for estimating the outcome of interest (competitiveness in our case) using the propensity scores obtained from the PS model. Although both outcome regression and PS matching without regression adjustment can be used separately to estimate causal effects (assuming unconfoundedness), they are only unbiased if the single statistical model is correctly specified. When using a doubly robust estimator that combines two models, it is sufficient to correctly specify one of them to obtain an unbiased estimate (Funk et al., 2011).

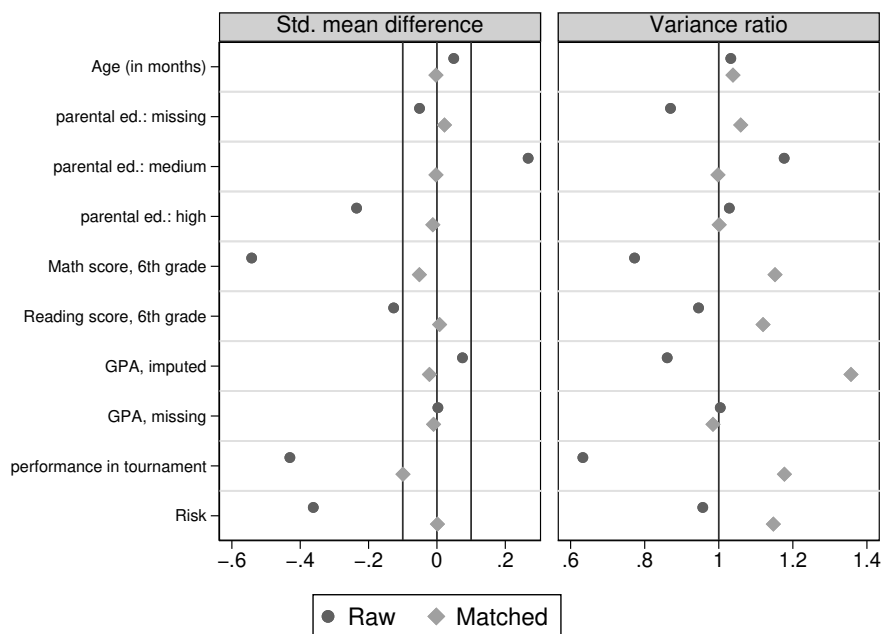
We match female and male students based on the control variables used in the mediation analysis (age, parental education, test scores, GPA, tournament performance in the experiment, risk taking and categories of GAP). We also use the same variables in the regression adjustment. To account for the multilevel nature of the data (participants nested in classrooms) cluster fixed effects can be applied to either the PS model or the outcome model. Su and Cortina (2009) and Li et al. (2013) found that applying them in both models reduces bias the most. Therefore, we control for class fixed effects in both the PS model and the regression adjustment.

Similar to the analysis in the previous section, we use an interaction-based and a

subgroup-based approach to test for moderation. As described by Green and Stuart (2014), both approaches are suitable for successfully balancing the subgroups when using matching.

In the interaction-based approach, we perform matching using the whole sample while requiring exact matching on categories of GAP, to ensure that matched pairs are from the same subgroup. We estimate three gender gaps by including an interaction term between gender and categories of GAP in the outcome regression, and then predict the marginal effects of gender at different levels of GAP. This technique is particularly effective when using small samples (such as ours), where matching in separate subgroups might be done with greater uncertainty and data loss (Wang et al., 2018).

Figure 5: Covariate balance between males and females before and after matching in the whole sample



Note: Raw values mean balance between genders before matching. Matched values mean balance after matching.

Covariate balance after matching is reported in Figure 5. None of the covariates exhibit a standardized mean difference greater than 0.1 between the matched gender groups, which is a commonly used cutoff in the literature to assess dissimilarity (Stuart et al., 2013; Zhang et al., 2019).

The marginal effects of gender from our interaction-based outcome model yield consistent results with the regression-based moderation analysis. No gender gap is observed in the underconfident and overconfident groups. However, we observe a 14-15 percentage point gap among individuals who evaluate their relative performance realistically (see column 1 in Table 3).

These findings are corroborated by the subgroup-based approach, where we perform matching and estimate the effect of gender in each GAP category separately (see Figure 18 in Appendix B for covariate balance in each subgroup). Female participants who assess their relative performance realistically are 15 percentage points less likely to enter competition compared to their male counterparts, but no difference is observed in the other two subgroups (see column 2 in Table 3). This indicates that the results are robust across all moderation approaches, despite the fact that within-subgroup matching comes with a greater share of observations where no match could be found (see Tables 12 and 13 in Appendix B for matching statistics).

Table 3: Results from the matching-based moderation analysis

	Interaction		Subgroups	
	Margins of female	Female gap 1	Female gap 2	Female gap 3
Underconf.	0.078 (0.067)	-0.007 (0.114)		
Realistic	-0.143** (0.046)		-0.153** (0.048)	
Overconf.	-0.006 (0.051)			-0.047 (0.069)
<i>N</i>	1064	1073		
Robust standard errors in parentheses				
*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$				

Note: *Interaction*: matching performed using the joint sample by requiring exact matching on categories of GAP. Three gender effects estimated by including an interaction of gender and categories of GAP in the outcome model, and predicting the marginal effect of gender by GAP levels. *Subgroup*: matching performed and gender effect estimated separately in each GAP group.

4.3.3 Robustness checks using another dataset

To further test the robustness of our arguments regarding moderation, we use data from a similarly designed experimental task of Sutter et al. (2016).¹⁶ Whilst their setup differs in many aspects from ours, the most important features of the design are similar. They also conduct the Niederle and Vesterlund (2007) experiment within small groups of teenagers, and importantly, both the believed and actual performance of the subjects are available so that categories of GAP can be replicated.¹⁷ The most important differences between their study and ours are the following:

- their sample of students comes from grade 5, 8 and 11
- their sample size is much smaller ($N=246$ as we have only kept the control group from their sample)
- their experiment was run in 4 schools and everyone from the same grade and same treatment group sat together (which essentially means that students from the same

¹⁶We are very grateful to Daniela Glätzle-Rützler for providing the necessary data in anonymized form for this robustness test.

¹⁷We were unable to get data for any other Niederle and Vesterlund (2007) experiments with teenagers where the GAP measure can be replicated.

grade sat together in the treatment group we use)

- students were grouped into groups of 6 (and not groups of 4) during the competition
- students did not get any feedback on their absolute performance until the very end of the experiment (so their beliefs regarding relative performance were elicited before getting feedback on absolute performance)
- they find no gender difference in actual tournament performance
- there is no classroom or group ID in the data (hence we cannot control either for group fixed-effects or cluster standard errors on groups level)
- there is no data on social background or school performance or any other preference (i.e. risk).

Nevertheless, since the moderation results using our original data were robust to different specifications and to using regression (see Table 2) or matching (see Table 3) we believe that a comparison of the descriptive results using our data and that of Sutter et al. (2016) is meaningful.

Figure 6: Competitiveness by categories of GAP and gender - data from Sutter et al. (2016)

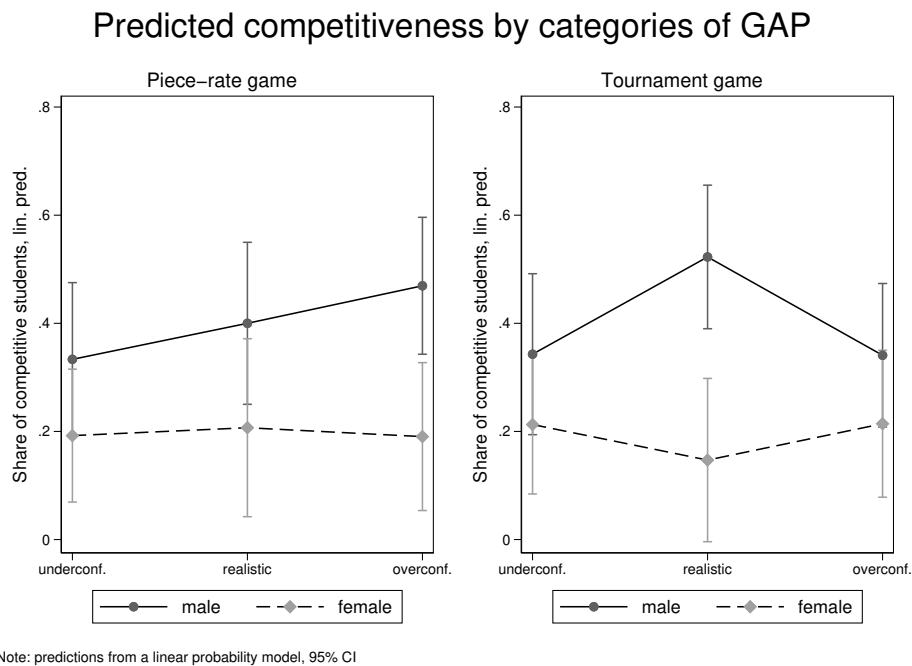


Figure 6 is a direct reproduction of Figure 2 using the data from Sutter et al. (2016). There are minor differences between the figures (e.g. albeit insignificantly, underconfident males in the Sutter et al. (2016) sample are more competitive than females even on

Table 4: Results from the regression-based moderation analysis - Sutter et al. (2016) data

	Interaction model	Underconf. subgroup	Realistic subgroup	Overconf subgroup
<i>Coefficient:</i>				
female		-0.154 (0.100)	-0.275** (0.094)	-0.106 (0.092)
<i>Margins of female:</i>				
Underconf.	-0.186* (0.094)			
Realistic	-0.278** (0.097)			
Overconf.	-0.105 (0.090)			
<i>N</i>	246	82	78	86
<i>R</i> ²		0.121	0.405	0.194
Robust standard errors in parentheses *** p<0.001, ** p<0.01, * p<0.05				

Note: *Interaction model*: the female dummy is interacted with categories of GAP. Marginal female effects are reported by categories of GAP. *Subgroup models*: a separate regression is run for each category of GAP. Female dummy coefficients are reported. In all four columns school-grade and real performance in the tournament are controlled for.

the right-side graph, while in our sample it is the other way around. The main takeaway, however, is the same: realistic males are more competitive than females, and the gender gap is the most pronounced in that group.

This message remains if we replicate our regressions from Table 2. Table 4 shows that the gender gap is the largest among the realistic.¹⁸

4.4 Breaking down the gender gap in competition among the realistic

Our original results point towards a more nuanced explanation of the gender gap in competition than simply males being more overconfident and hence competitive than females. There is a substantial gender gap in competition among participants who evaluate their performance realistically. To further examine these differences, Table 5 focuses on the realistic group and presents a breakdown of gender differences in competitiveness based on performance rank. Note that in all quartiles except the second one, females compete more than males. In the second best-performing quartile, 72% of male participants enter the competition, while the corresponding share of females is only 48%.

This indicates that realistic males are substantially more likely to compete when they perform above the median compared to when they are below it (so comparing the second quartile (Q2) to the third quartile (Q3)). In contrast, realistic females refrain from participating in competition unless they are in the top quartile. This finding cannot be driven by the higher performance of males, as the gender difference in performance is

¹⁸Note that the significant gender differences in Table 4 column (1) among the underconfident might be due to the lack of proper controls, such as group IDs, family background or test scores.

Table 5: Means (M) and Standard Deviations (SD) of the competitiveness dummy by gender and performance rank in the realistic group (number of students in each cell is in parenthesis)

		Rank in tournament game (quartiles)				
		1	2	3	4	Total
male competitiveness	M	0.95	0.72	0.27	0.20	0.76
	SD	0.22	0.45	0.46	0.42	0.43
	N	(97)	(53)	(22)	(10)	(182)
female competitiveness	M	0.96	0.48	0.32	0.26	0.51
	SD	0.21	0.50	0.47	0.45	0.50
	N	(45)	(83)	(50)	(34)	(212)
Total	M	0.95	0.57	0.31	0.25	0.62
	SD	0.22	0.50	0.46	0.44	0.48
	N	(142)	(136)	(72)	(44)	(394)

a bit smaller in the second than in the third quartile in the realistic group (see Table 14 in Appendix B). However, females do not compete less than males in any of the quartiles except for the second one.

Thus, the gender gap observed in the realistic group can be interpreted as the difference in taste or preference for competition between realistic females and males which represents a distinct channel from the compositional one (more males being overconfident). Realistic males prefer to compete if they are in the better-performing half of the class, whereas realistic females exhibit a preference for competition only when they are among the highest-performing participants ¹⁹.

5 Decomposing the total gender gap in competitiveness

In this section, we carry out a decomposition exercise to separate the previously mentioned channels in explaining the total raw gender gap in competitiveness: the gender difference in confidence/levels of GAP (composition channel) and the gender difference in the willingness to compete (preference or taste channel).

Let s_i^j denote the share of $j = m, f$ (m denoting males, and f females) in the group $i = o, r, u$ (o denoting overconfident, r realistic, and u underconfident). Hence, s_r^f represents the share of females that are in the realistic group. Note that $s_o^f + s_r^f + s_u^f = 1$ and $s_o^m + s_r^m + s_u^m = 1$, so any female (male) is in a group.

Let c_i^j denote the proportion of $j = m, f$ (m denoting males, and f females) that competes in the group $i = o, r, u$ (o denoting overconfident, r realistic, and u underconfident). Hence, c_o^f indicates the proportion of females that compete in the overconfident group.

The overall share of females and males competing can be written as $(s_o^f \times c_o^f) + (s_r^f \times$

¹⁹Note: there was no feedback on performance levels between the rounds

$c_r^f) + (s_u^f \times c_u^f)$ and $(s_o^m \times c_o^m) + (s_r^m \times c_r^m) + (s_u^m \times c_u^m)$, respectively. That is, for both genders we weigh the willingness to enter competition in a given group by the share of participants in that group, and aggregate across groups. As a consequence, the gender gap in competitiveness can be written as

$$\begin{aligned} & \overbrace{[(s_o^m \times c_o^m) + (s_r^m \times c_r^m) + (s_u^m \times c_u^m)]}^{\text{competitiveness of males}} - \overbrace{[(s_o^f \times c_o^f) + (s_r^f \times c_r^f) + (s_u^f \times c_u^f)]}^{\text{competitiveness of females}} = \\ & \underbrace{[(s_o^m \times c_o^m) - (s_o^f \times c_o^f)]}_{\text{gender gap in comp. in the overconf.,}} + \underbrace{[(s_r^m \times c_r^m) - (s_r^f \times c_r^f)]}_{\text{...realistic,}} + \underbrace{[(s_u^m \times c_u^m) - (s_u^f \times c_u^f)]}_{\text{and underconf. groups}}. \end{aligned} \quad (1)$$

The first line in equation 1 is just the gender difference between males and females, while the second line displays the gender gap by categories of GAP. Note that the gender gap has two sources: gender difference in the shares of being in a given category of GAP, and gender gap in the willingness to compete in a given category of GAP.

The second line in (1) can be rewritten as

$$\begin{aligned} & c_o^m \times (s_o^m - s_o^f) + s_o^f \times (c_o^m - c_o^f) + \\ & c_r^m \times (s_r^m - s_r^f) + s_r^f \times (c_r^m - c_r^f) + \\ & c_u^m \times (s_u^m - s_u^f) + s_u^f \times (c_u^m - c_u^f). \end{aligned} \quad (2)$$

The first term in the first line of expression (2) represents the portion of the overall gender gap in competitiveness due to the gender difference in the share of participants in the overconfident group ($s_o^m - s_o^f$), assuming that both males and females in this group are equally likely to enter competition (a likelihood equated to that of males, c_o^m). The second term in the first line captures the part of the overall gender gap in competitiveness that can be attributed to the difference between males and females in the overconfident group in their willingness to compete ($c_o^m - c_o^f$), assuming that the same share of males and females (equal to the share of females, s_o^f) is in the overconfident group.

The second and third lines follow the same logic, but apply to the realistic and underconfident groups, respectively. Therefore, the first and second terms in the second and third lines capture the gender differences in the share of subjects and the willingness to compete within the realistic and underconfident groups, respectively.

The existing literature highlights the importance of the first term in decomposition (2). It captures the idea that there are more overconfident males than females ($s_o^m > s_o^f$).²⁰ The second term adds that in the overconfident group, potentially males and females are not equally likely to enter competition. Our main finding points out that the second term in the second line of decomposition (2) is also relevant, reflecting the gender difference in the willingness to compete in the group of realistic participants. Additionally, decomposition (2) illuminates that gender differences in the share of sub-

²⁰For instance Niederle and Vesterlund (2007) claim in the abstract of their seminal paper that "the tournament-entry gap is driven by men being more overconfident".

jects and the willingness to compete in the underconfident group may also contribute to the overall gender gap in competitiveness.

Table 6: Decomposition of the total gender gap in competitiveness

Group	diff. in gender composition ($c_j^m \times (s_j^m - s_j^f)$)	gender diff. in competitiveness ($s_j^f \times (c_j^m - c_j^f)$)	Sum
Overconfident	$0.632 \times (0.405-0.354)$ 0.032 (31.8%)	$0.354 \times (0.632-0.565)$ 0.024 (23.4%)	0.056 (55.2%)
Realistic	$0.758 \times (0.382-0.347)$ 0.027 (26.2%)	$0.347 \times (0.758-0.509)$ 0.086 (85.3%)	0.113 (111.5%)
Underconfident	$0.549 \times (0.214-0.300)$ -0.047 (-46.6%)	$0.300 \times (0.549-0.617)$ -0.020 (-20.1%)	-0.068 (-66.8%)

Table 6 contains the actual numbers from our data by categories of GAP. It also shows the percentage that the components represent in the total gender gap in competitiveness.²¹ Note that the difference in the competitiveness in the realistic group accounts by far for the largest share of the overall gender gap in competitiveness. Its relevance (85.3%) is more than double of the importance of males being more likely to exhibit overconfidence (31.8%).

It is also noteworthy that in the underconfident group females are more likely to compete than males, and relatively there are more females among the underconfident. These two phenomena in the underconfident group have a mitigating effect on the overall gender gap in competitiveness.

6 Conclusion

In our study, we aimed to explore how confidence mediates and moderates the gender gap in competitiveness. To this end, we proposed a categorical confidence variable measuring the categories of the Gussed-Actual Performance difference - GAP (underconfident, realistic, overconfident). These new confidence categories enabled us to examine mechanisms between confidence and competitiveness from a gender perspective in greater detail than previous studies.

In our experimental setting, we measured preferences towards competition using the widely applied design of Niederle and Vesterlund (2007), while also assessing beliefs about students' own relative performance and risk preferences in an incentivized way. Our analysis shows that even after controlling for classroom fixed effects and background characteristics, there is a substantial gender gap of about 10 percentage points between male and female participants. This gap persists after accounting for performance differences between the gender groups in the experimental real-effort task.

The mediation and moderation analyses outline two channels of explanation. One is about the composition of gender groups according to the categories of GAP. We find

²¹By adding up all numbers in columns (2) and (3), we obtain the total gender difference in competitiveness, 0.101, that is about 10% in the willingness to compete. The percentages in parentheses, representing the relative portions, add up to 100%. The last column indicates the horizontal sums.

a positive association between higher confidence and competitiveness, and since males in our sample are more (over)confident than females, the gender gap in competition is partly driven by "overconfident men" (compositional effect). However, this is only part of the full story. Using moderation techniques we uncovered that there is no significant gender gap in competitiveness among the under- or overconfident participants. But in the realistic group (where participants' believed performance rank coincides with the actual one) we observe a 14 percentage points gender gap in favor of males. Our results suggest that the gender gap among the realistic reflects differences in taste for competition: males compete if they have above median performance, but females need to be in the best quartile to prefer competition.

Based on our findings, if we want to decrease gender differences in competitiveness, it is not enough to apply interventions aiming to raise the level of confidence of females. Even if the composition of females in terms of confidence was similar to that of males, gender differences in competition would persist among those who assess their performance correctly. Future research should address the question of how to handle this second channel effectively.

References

- Abeler, J., Falk, A., Goette, L., Huffman, D., 2011. Reference points and effort provision. *American Economic Review* 101, 470–92.
- Almås, I., Cappelen, A.W., Salvanes, K.G., Sørensen, E.Ø., Tungodden, B., 2016. Willingness to compete: Family matters. *Management Science* 62, 2149–2162.
- Babcock, L., Laschever, S., 2009. *Women don't ask: Negotiation and the gender divide*. Princeton University Press.
- Balázs, I., Ostorics, L., 2020. *The Hungarian Educational Assessment System*. Springer International Publishing, Cham. pp. 157–169. URL: https://doi.org/10.1007/978-3-030-38969-7_13, doi:10.1007/978-3-030-38969-7_13.
- Berg, J., Dickhaut, J., McCabe, K., 1995. Trust, reciprocity, and social history. *Games and economic behavior* 10, 122–142.
- Bertrand, M., Goldin, C., Katz, L.F., 2010. Dynamics of the gender gap for young professionals in the financial and corporate sectors. *American economic journal: applied economics* 2, 228–55.
- Bertrand, M., Hallock, K.F., 2001. The gender gap in top corporate jobs. *ILR Review* 55, 3–21.
- Booth, A., Nolen, P., 2012. Choosing to compete: How different are girls and boys? *Journal of Economic Behavior & Organization* 81, 542–555.

- Buser, T., Niederle, M., Oosterbeek, H., 2014. Gender, competitiveness, and career choices. *The Quarterly Journal of Economics* 129, 1409–1447.
- Buser, T., Ranehill, E., Van Veldhuizen, R., 2021. Gender differences in willingness to compete: The role of public observability. *Journal of Economic Psychology* 83, 102366.
- Cain, G.G., 1986. The economic analysis of labor market discrimination: A survey. *Handbook of labor economics* 1, 693–785.
- Cárdenas, J.C., Dreber, A., Von Essen, E., Ranehill, E., 2012. Gender differences in competitiveness and risk taking: Comparing children in colombia and sweden. *Journal of Economic Behavior & Organization* 83, 11–23.
- Crosetto, P., Filippin, A., 2013. The “bomb” risk elicitation task. *Journal of Risk and Uncertainty* 47, 31–65.
- Dreber, A., von Essen, E., Ranehill, E., 2014. Gender and competition in adolescence: task matters. *Experimental Economics* 17, 154–172.
- Fischbacher, U., 2007. z-tree: Zurich toolbox for ready-made economic experiments. *Experimental economics* 10, 171–178.
- Flory, J.A., Leibbrandt, A., List, J.A., 2015. Do competitive workplaces deter female workers? a large-scale natural field experiment on job entry decisions. *The Review of Economic Studies* 82, 122–155.
- Funk, M.J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M.A., Davidian, M., 2011. Doubly Robust Estimation of Causal Effects. *American Journal of Epidemiology* 173, 761–767. URL: <https://doi.org/10.1093/aje/kwq439>, doi:10.1093/aje/kwq439, arXiv:<https://academic.oup.com/aje/article-pdf/173/7/761/17338964/kwq439.pdf>.
- Gillen, B., Snowberg, E., Yariv, L., 2019. Experimenting with measurement error: Techniques with applications to the caltech cohort study. *Journal of Political Economy* 127, 1826–1863.
- Goldin, C., Katz, L.F., Kuziemko, I., 2006. The homecoming of american college women: The reversal of the college gender gap. *Journal of Economic perspectives* 20, 133–156.
- Goldin, C., Rouse, C., 2000. Orchestrating impartiality: The impact of” blind” auditions on female musicians. *American economic review* 90, 715–741.
- Green, K.M., Stuart, E.A., 2014. Examining moderation analyses in propensity score methods: Application to depression and substance use. *Journal of Consulting and Clinical Psychology* 82, 773–783. URL: <http://doi.apa.org/getdoi.cfm?doi=10.1037/a0036515>, doi:10.1037/a0036515.

- Hajdu, T., Hermann, Z., Horn, D., Hönich, H., Varga, J., 2022. A közoktatás indikátorrendszere 2021. Technical Report. Közgazdaság- és Regionális Tudományi Kutatóközpont Közgazdaság-tudományi Intézet. URL: https://kti.krtk.hu/wp-content/uploads/2022/02/A_kozoktatás_indikátorrendszere_2021.pdf.
- Horn, D., Kiss, H.J., Lénárd, T., 2022a. Gender differences in preferences of adolescents: evidence from a large-scale classroom experiment. *Journal of Economic Behavior & Organization* 194, 478–522.
- Horn, D., Kiss, H.J., Lénárd, T., 2022b. Preferences of adolescents—a dataset containing linked experimental task measures and register data. *Data in Brief* 42, 108088.
- Khachatryan, K., Dreber, A., Von Essen, E., Ranehill, E., 2015. Gender and preferences at a young age: Evidence from armenia. *Journal of Economic Behavior & Organization* 118, 318–332.
- Li, F., Zaslavsky, A.M., Landrum, M.B., 2013. Propensity score weighting with multi-level data. *Statistics in medicine* 32, 3373–3387.
- Lozano, L., Reuben, E., 2022. Measuring Preferences for Competition. Technical Report. New York University Abu Dhabi, Department of Social Science.
- Markowsky, E., Beblo, M., 2022. When do we observe a gender gap in competition entry? a meta-analysis of the experimental literature. *Journal of Economic Behavior & Organization* 198, 139–163.
- Moore, D.A., Dev, A.S., 2020. Overconfidence. Springer International Publishing, Cham. pp. 3382–3386. URL: https://doi.org/10.1007/978-3-319-24612-3_1157, doi:10.1007/978-3-319-24612-3_1157.
- Moore, D.A., Healy, P.J., 2008. The trouble with overconfidence. *Psychological review* 115, 502.
- Moore, D.A., Schatz, D., 2017. The three faces of overconfidence. *Social and Personality Psychology Compass* 11, e12331.
- Niederle, M., 2016. Gender, in: *Handbook of Experimental Economics*. second edition ed.. Princeton University Press, pp. 481–553.
- Niederle, M., Vesterlund, L., 2007. Do women shy away from competition? do men compete too much? *The quarterly journal of economics* 122, 1067–1101.
- Polachek, S.W., 1981. Occupational self-selection: A human capital approach to sex differences in occupational structure. *The review of Economics and Statistics* , 60–69.
- Stuart, E.A., Lee, B.K., Leacy, F.P., 2013. Prognostic score-based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *Journal of Clinical Epidemiology* 66, S84–S90.e1. URL: <https://doi.org/10.1016/j.jclinepi.2012.08.011>.

//linkinghub.elsevier.com/retrieve/pii/S0895435613001625, doi:10.1016/j.jclinepi.2013.01.013.

- Su, Y.S., Cortina, J., 2009. What do We Gain? Combining Propensity Score Methods and Multilevel Modeling. URL: <https://ssrn.com/abstract=1450058>.
- Sutter, M., Glätzle-Rützler, D., 2015. Gender differences in the willingness to compete emerge early in life and persist. *Management Science* 61, 2339–2354.
- Sutter, M., Glätzle-Rützler, D., Balafoutas, L., Czermak, S., 2016. Cancelling out early age gender differences in competition: an analysis of policy interventions. *Experimental Economics* 19, 412–432.
- Sutter, M., Zoller, C., Glätzle-Rützler, D., 2019. Economic behavior of children and adolescents—a first survey of experimental economics results. *European Economic Review* 111, 98–121.
- Van Veldhuizen, R., 2022. Gender differences in tournament choices: Risk preferences, overconfidence, or competitiveness? *Journal of the European Economic Association* 20, 1595–1618.
- Wang, S.V., Jin, Y., Fireman, B., Gruber, S., He, M., Wyss, R., Shin, H., Ma, Y., Keeton, S., Karami, S., Major, J.M., Schneeweiss, S., Gagne, J.J., 2018. Relative Performance of Propensity Score Matching Strategies for Subgroup Analyses. *American Journal of Epidemiology* 187, 1799–1807. URL: <https://academic.oup.com/aje/article/187/8/1799/4937537>, doi:10.1093/aje/kwy049.
- Zhang, J., 2011. Do girls in china compete just as much as boys? evidence from an experiment that predicts educational choice, in: Technical Report. Citeseer.
- Zhang, Y.J., 2019. Culture, institutions and the gender gap in competitive inclination: Evidence from the communist experiment in china. *The Economic Journal* 129, 509–552.
- Zhang, Z., Kim, H.J., Lonjon, G., Zhu, Y., 2019. Balance diagnostics after propensity score matching. *Annals of Translational Medicine* 7, 16–16. URL: <http://atm.amegroups.com/article/view/22865/22385>, doi:10.21037/atm.2018.12.10.

7 APPENDICES

7.1 Appendix A: Experimental instructions as shown in zTree

Figure 7: Introductory instructions as shown on the first zTree screen

Dear Participant!

First of all, we would like to thank you for participating in this experiment. We intend to use the data obtained during the session for research purposes at the Centre for Economic and Regional Studies.

Participation is VOLUNTARY. You can quit the experiment at any time without having to give any justification, or you may deny answering questions.

The experiment will proceed as follows:

- During the experiment, you will participate in 8 small games in which you will have to make different choices.
- There is no objectively right answer in any of the decisions, just answer honestly!
- Before each decision, we will describe the situation in detail and we will explain what the choice is about. If the description or the explanation is not clear, please raise your hand, and the experimenter will answer your question.
- Depending on your choices, you may earn canteen vouchers at the end of the experiment. More precisely, at the end of the experiment, we will pick one of the games randomly, and your decision in that game determines your earning. We will round the earnings to 100 HUF-s and will pay the vouchers accordingly.
- Note that in some situations, the earnings do not depend on your decision only, but also on the decision of another participant. We will describe in the presentation of each situation how the earnings would be determined if that game were picked for payment.
- The payment will take place after the experiment. You will receive the vouchers here in the room.

Participation in the research is ANONYMOUS. We will treat all information that we collect during the research confidentially.

Please, remain silent during the experiment and do not disturb each other. It is forbidden to talk! Should you have a question, turn to the experimenter.

Please, silence your mobile phone! Those who misbehave will be excluded from the experiment. In an extreme case, we may exclude the whole group, and nobody will earn anything. We will send feedback about the results of the experiment to the school. You can ask your teacher about this in a few weeks.

To keep the research anonymous, we will ask for your NABC ID.

Thank you for your cooperation!
If you agree to participate, press OK!

OK

Figure 8: General introduction to the competition game

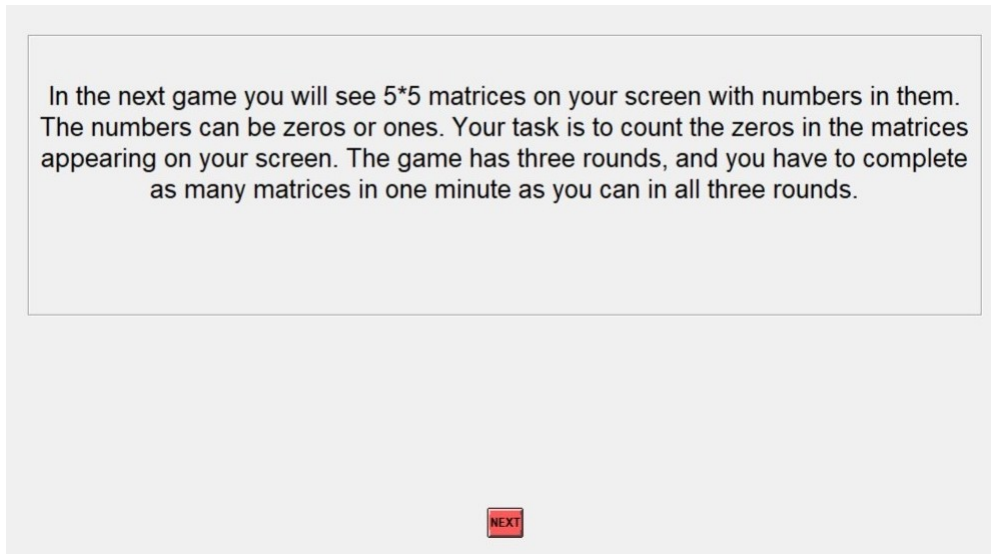


Figure 9: Instructions to the first round in the competition game (piece-rate game)

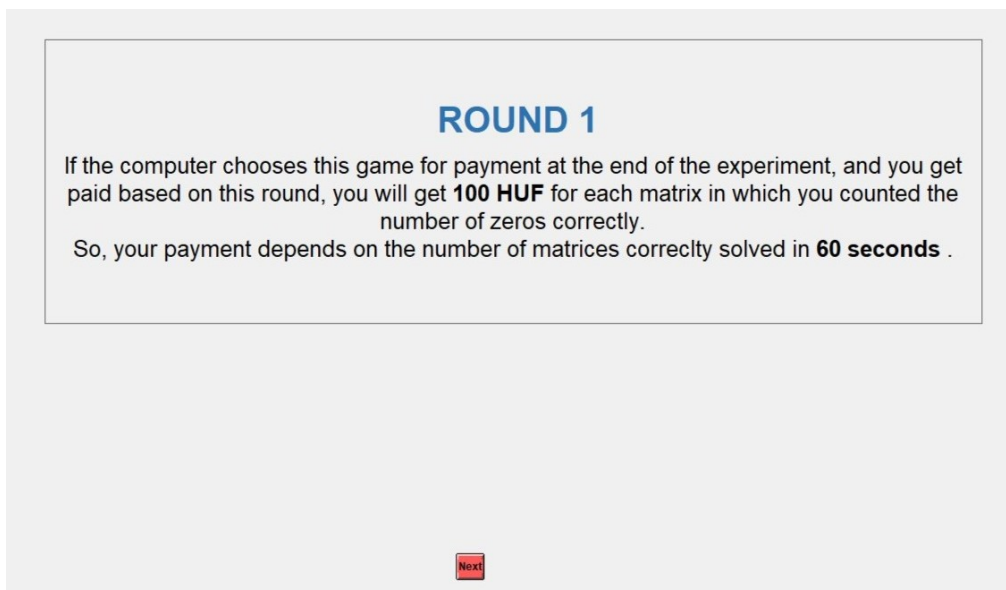
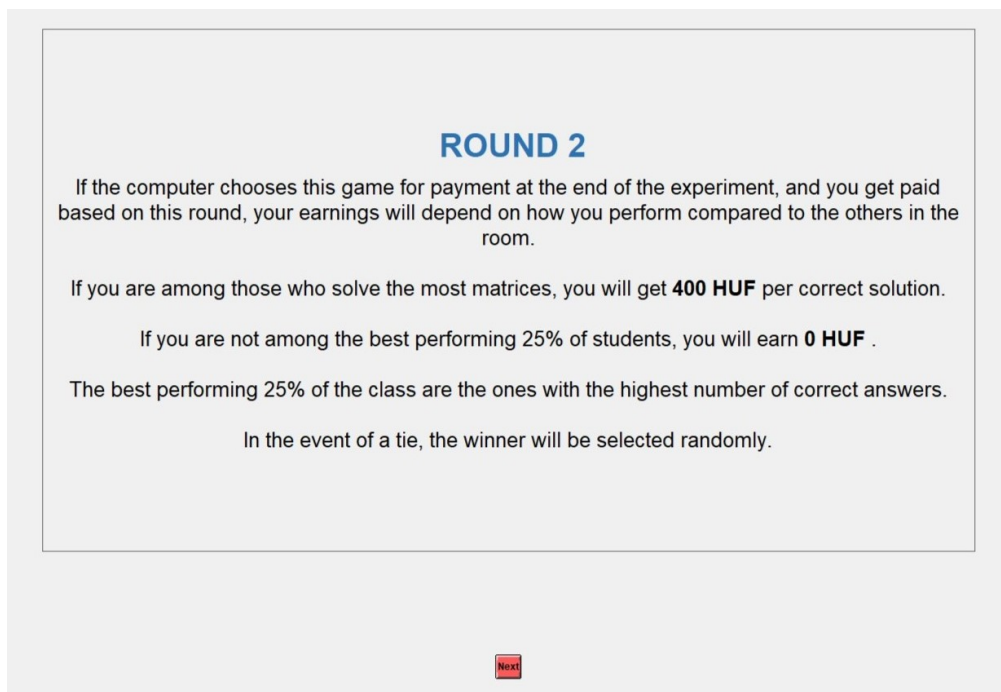


Figure 10: Instructions to the second round in the competition game (tournament game)



ROUND 2

If the computer chooses this game for payment at the end of the experiment, and you get paid based on this round, your earnings will depend on how you perform compared to the others in the room.

If you are among those who solve the most matrices, you will get **400 HUF** per correct solution.

If you are not among the best performing 25% of students, you will earn **0 HUF**.

The best performing 25% of the class are the ones with the highest number of correct answers.

In the event of a tie, the winner will be selected randomly.




Figure 11: Instructions to the third round in the competition game (choice between piece-rate and tournament)

ROUND 3

In this round, you can choose if you want to get paid based on the first or the second round's payment scheme. If the computer chooses this game for payment at the end of the experiment, and you get paid based on this round, your winnings will be calculated based on which option you choose now!

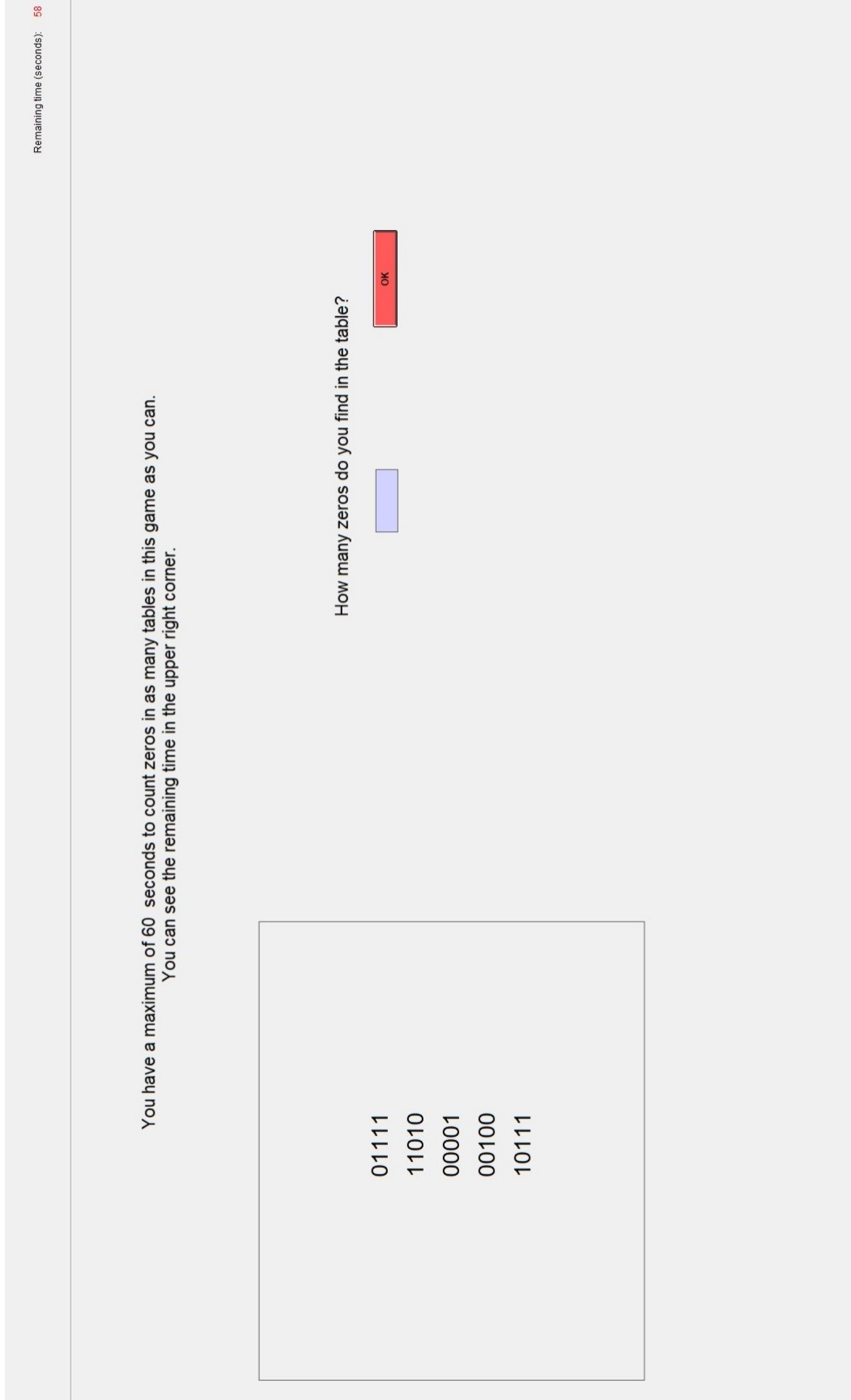
<div style="text-align: center;"><h4>1ST OPTION: THE 1ST ROUND'S PAYMENT SCHEME</h4></div> <p>If you choose the piece rate scheme used in the first round, you will get 100 HUF per correctly solved matrix.</p>	<div style="text-align: center;"><h4>2ND OPTION: THE 2ND ROUND'S PAYMENT SCHEME</h4></div> <p>If you choose the tournament scheme used in the second round, you will get 400 HUF per correctly solved matrix, but only if you are among the best performing students. We will compare your performance in this round to others' performance in the second round. If you are not among the best performing 25% of students, you will earn 0 HUF. In the event of a tie, the winner will be selected randomly.</p>
---	---

Please choose a payment scheme!

Payment according to the 1st round

Payment according to the 2nd round

Figure 12: Real effort task screen (the same was shown in all three rounds)



7.2 Appendix B: Tables and figures

Table 7: Descriptives (means) by gender and balance tests

Variable	Males	SD	Females	SD	Balance test	SE
parental ed.: low	0.01	(0.12)	0.02	(0.14)	-0.00	(0.01)
parental ed.: medium	0.29	(0.45)	0.42	(0.49)	0.08***	(0.03)
parental ed.: high	0.58	(0.49)	0.47	(0.50)	-0.04	(0.03)
parental ed.: missing	0.11	(0.32)	0.09	(0.29)	-0.04*	(0.02)
father: employed	0.68	(0.47)	0.64	(0.48)	-0.01	(0.03)
father: self-employed	0.12	(0.32)	0.17	(0.38)	0.05**	(0.02)
father: regular work	0.02	(0.14)	0.02	(0.15)	0.00	(0.01)
father: occasional work	0.02	(0.14)	0.01	(0.11)	-0.01	(0.01)
father: childcare	0.02	(0.13)	0.01	(0.11)	-0.01	(0.01)
father: retired	0.00	(0.06)	0.01	(0.11)	0.01	(0.01)
father: unemployed	0.00	(0.05)	0.01	(0.09)	0.01	(0.00)
father: disabled	0.00	(0.06)	0.01	(0.10)	0.00	(0.01)
father: missing	0.14	(0.35)	0.12	(0.32)	-0.05**	(0.02)
Math score, 6th grade	0.29	(1.03)	-0.23	(0.91)	-0.22***	(0.05)
Reading score, 6th grade	0.07	(1.01)	-0.05	(0.99)	0.10*	(0.05)
GPA, imputed	4.51	(0.45)	4.54	(0.42)	0.11***	(0.02)
GPA, missing	0.20	(0.40)	0.19	(0.39)	-0.03	(0.03)
Math grade, imputed	4.27	(0.85)	4.17	(0.84)	0.05	(0.05)
Hungarian grade, imputed	4.31	(0.75)	4.39	(0.69)	0.21***	(0.04)
Literature grade, imputed	4.48	(0.65)	4.56	(0.64)	0.17***	(0.04)
Math grade, missing	0.16	(0.37)	0.13	(0.34)	-0.05**	(0.02)
Hungarian grade, missing	0.16	(0.37)	0.13	(0.34)	-0.05**	(0.02)
Literature grade, missing	0.17	(0.37)	0.14	(0.34)	-0.05**	(0.02)
performance in piece-rate game	7.47	(3.75)	6.16	(2.86)	-0.45***	(0.16)
performance in tournament	8.78	(3.67)	7.37	(2.94)	-0.64***	(0.17)
guessed rank in tournament	1.93	(0.90)	2.36	(0.86)	0.46***	(0.06)
Confidence/overplacement, tourn.	1.19	(0.76)	1.05	(0.81)	-0.11**	(0.05)
Underconfident	0.21	(0.41)	0.30	(0.46)	0.08***	(0.03)
Realistic	0.38	(0.49)	0.35	(0.48)	-0.04	(0.03)
Overconfident	0.40	(0.49)	0.35	(0.48)	-0.03	(0.03)
Risk	37.27	(18.45)	30.80	(18.18)	-4.90***	(1.21)
Competition	0.66	(0.47)	0.56	(0.50)	-0.11***	(0.03)
Observations	477		611		1,108	

*** p<0.001, ** p<0.01, * p<0.05

Note 1: Performance in the piece-rate and tournament games means the number of matrices correctly solved; competitiveness means the share of students willing to compete in stage 3; ranks represent performance quartiles (Q1 meaning best performance), confidence is a categorical variable (0=underconf., 1=realistic, 2=overconf.), the three confidence categories show the share of students in each category by gender, risk-taking means the number of boxes taken out in the bomb risk elicitation task, competition means the share of students choosing the tournament in round 3 by gender.

Note 2: In the balance test column, every figure is a regression coefficient. Separate regressions are run, always using the variable from the first column as the dependent and the female dummy as an independent variable. Classroom fixed effects are controlled for throughout. Coefficients of the female dummy are reported.

Table 8: Performance and competitiveness by gender and by breakdown following the distribution of tournament performance ranks

	<i>female</i>		<i>male</i>	
	Mean	Std.Dev.	Mean	Std.Dev.
<i>In Q1 based on actual perf.</i>				
Performance in tournament	10.06	2.80	11.56	3.54
Competitiveness	0.76	0.43	0.83	0.37
<i>In Q2 based on actual perf.</i>				
Performance in tournament	7.95	2.01	8.67	2.68
Competitiveness	0.55	0.50	0.68	0.47
<i>In Q3 based on actual perf.</i>				
Performance in tournament	6.66	1.71	7.45	2.58
Competitiveness	0.53	0.50	0.55	0.50
<i>In Q4 based on actual perf.</i>				
Performance in tournament	4.48	1.89	5.21	2.29
Competitiveness	0.39	0.49	0.44	0.50
<i>In Q1 based on believed perf.</i>				
Share of students believing to be in Q1	0.15	0.35	0.38	0.49
Performance in tournament	9.08	3.55	10.16	3.67
Competitiveness	0.81	0.40	0.88	0.32
<i>In Q2 based on believed perf.</i>				
Share of students believing to be in Q2	0.45	0.50	0.37	0.48
Performance in tournament	7.68	2.79	8.51	3.40
Competitiveness	0.60	0.49	0.64	0.48
<i>In Q3 based on believed perf.</i>				
Share of students believing to be in Q3	0.30	0.46	0.19	0.39
Performance in tournament	6.63	2.36	7.24	3.31
Competitiveness	0.44	0.50	0.36	0.48
<i>In Q4 based on believed perf.</i>				
Share of students believing to be in Q4	0.10	0.31	0.06	0.24
Performance in tournament	5.83	2.71	6.75	3.25
Competitiveness	0.41	0.50	0.39	0.50
Observations	611		477	

Note: Q1-Q4 represent ranks or performance quartiles (Q1 meaning best performance), performance means the number of matrices correctly solved; competitiveness means the share of students willing to compete in stage 3.

Table 9: Results from the mediation analysis, using the believed performance as a mediator

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Competition	Competition	Competition	Competition	Competition	Competition	Competition	Competition	Competition
Female	-0.101*** (0.030)	-0.113*** (0.028)	-0.115*** (0.029)	-0.114*** (0.029)	-0.117*** (0.030)	-0.112*** (0.029)	-0.082*** (0.029)	-0.033 (0.029)	-0.027 (0.028)
performance in tournament							0.055*** (0.006)	0.035*** (0.006)	0.035*** (0.006)
Guessed rank								-0.147*** (0.016)	-0.145*** (0.016)
Risk								0.002***	0.002***
Constant	0.662*** (0.028)	0.669*** (0.016)	1.034 (0.631)	1.051 (0.668)	1.117* (0.662)	1.171 (0.700)	0.867 (0.754)	1.486* (0.752)	1.419* (0.776)
Observations	1088	1088	1088	1088	1081	1081	1081	1081	1073
R^2	0.011	0.079	0.079	0.080	0.084	0.091	0.163	0.217	0.223
additional controls	none	+class FE	+age	+family	+cogn. skills	+GPA	+tournament perf.	+beliefs	+risktaking

Robust standard errors in parentheses
 *** p<0.001, ** p<0.01, * p<0.05

Table 10: Results from the mediation analysis, using categories of GAP as a mediator

	(1) Competition	(2) Competition	(3) Competition
Female	-0.082*** (0.029)	-0.056* (0.030)	-0.050* (0.029)
performance in tournament	0.055*** (0.006)	0.079*** (0.007)	0.078*** (0.007)
realistic		0.102*** (0.034)	0.106*** (0.033)
overconfident		0.265*** (0.042)	0.259*** (0.042)
Risk			0.002*** (0.001)
Constant	0.867 (0.754)	0.552 (0.747)	0.493 (0.772)
Observations	1081	1081	1073
R^2	0.163	0.194	0.200
additional controls	+tournament perf.	+GAP	+risktaking

Robust standard errors in parentheses
*** p<0.001, ** p<0.01, * p<0.05

Note: Models (1) - (6) from Table 9 produce the same results in this mediation analysis, thus those are not reported again. Realistic and overconfident mean levels of GAP compared to the baseline category of underconfident.

Figure 13: Female dummy coefficients from the robustness checks to the first mediation analysis (using beliefs as a mediator)

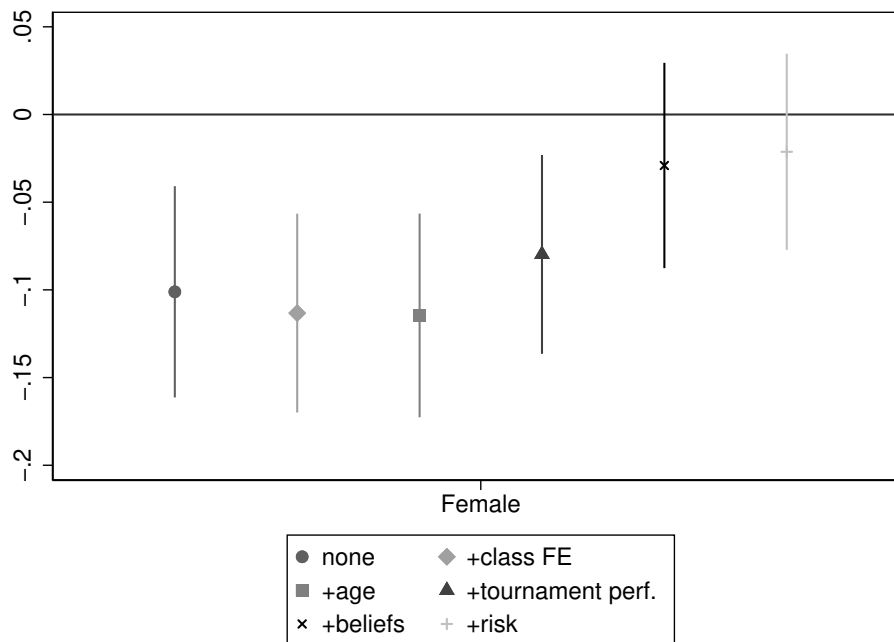


Figure 14: Female dummy coefficients from the robustness checks to the second version of the mediation analysis using the categories of GAP as a mediator

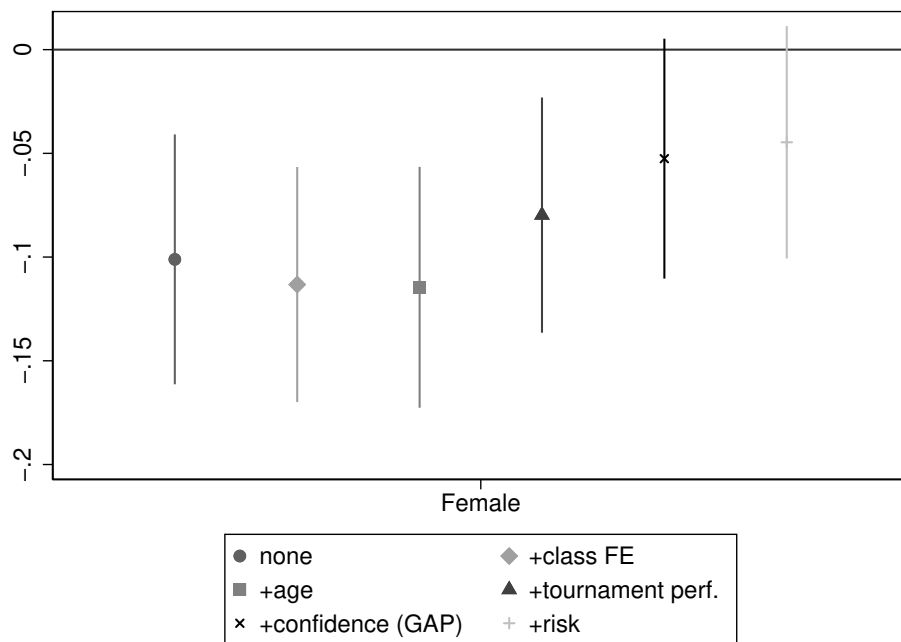


Table 11: Results from the regression-based moderation analysis

	Interaction model	Underconf. subgroup	Realistic subgroup	Overconf subgroup
Female	-0.139*** (0.042)	0.093 (0.085)	-0.139*** (0.043)	-0.028 (0.066)
underconfident	-0.239*** (0.058)			
overconfident	0.103* (0.059)			
female × underconfident	0.226*** (0.079)			
female × overconfident	0.083 (0.078)			
Observations	1073	283	388	402
R^2	0.207	0.245	0.393	0.245

Robust standard errors in parentheses
 *** p<0.001, ** p<0.01, * p<0.05

Note: all models control for class FE, age, family background (level of parental education), cognitive skills (standardized national test scores in math and reading) and GPA, tournament performance and risk aversion.

Figure 15: Female dummy coefficients (gender differences in competitiveness) among the underconfident - robustness checks to the moderation analysis

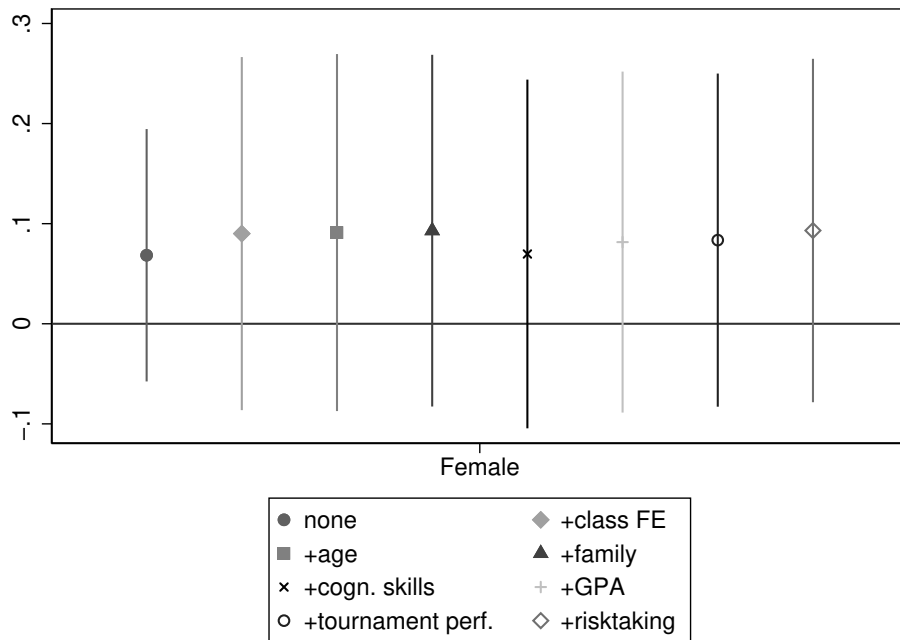


Figure 16: Female dummy coefficients (gender differences in competitiveness) among the realistic - robustness checks to the moderation analysis

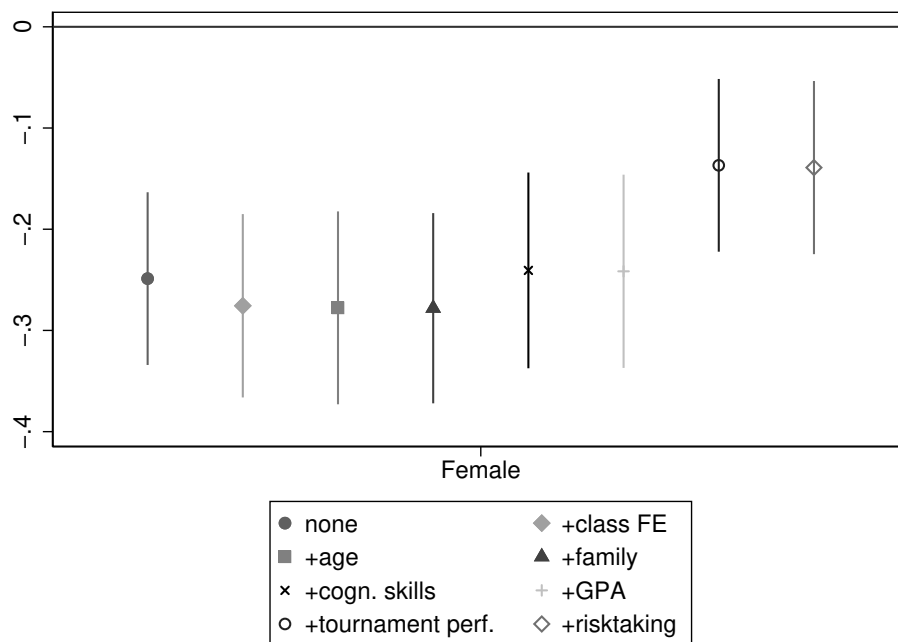


Figure 17: Female dummy coefficients (gender differences in competitiveness) among the overconfident - robustness checks to the moderation analysis

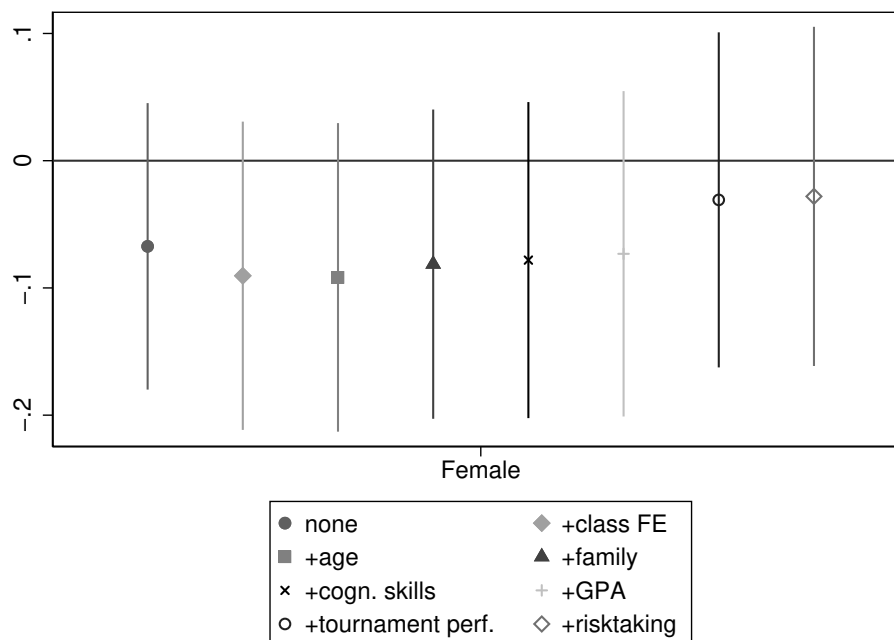
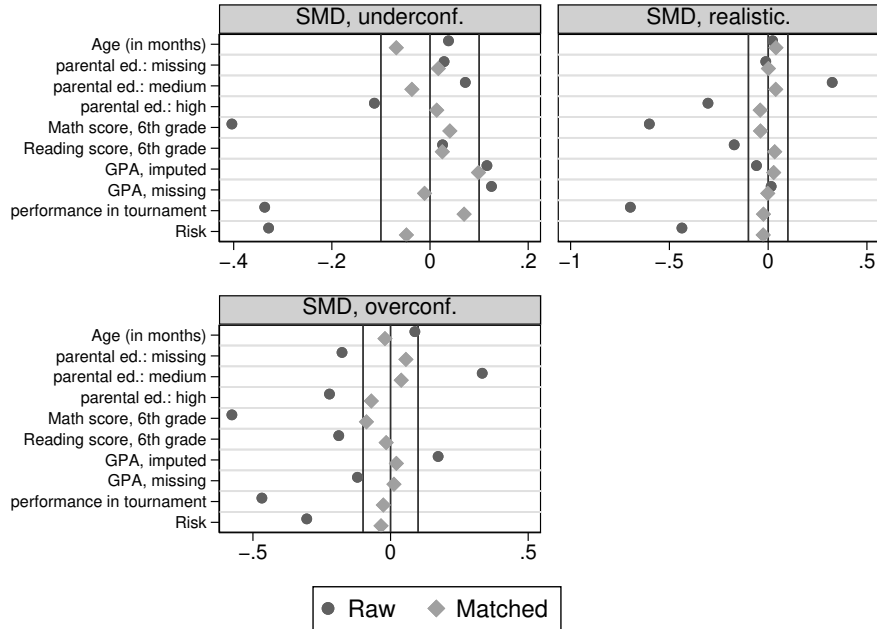


Figure 18: Covariate balance between males and females (only standardized mean differences) before and after matching in subgroups according to categories of GAP



Note: Raw values mean balance between genders before matching. Matched values mean balance after matching.

Table 12: Matching statistics after matching in the whole sample

	Treated	Untreated	Combined
Matched			
Yes	595	469	1064
No	9	0	9
Total	604	469	1073
Controls			
Used	469	595	1064
Unused	0	9	9
Total	469	604	1073

Table 13: Matching statistics after matching in subsamples according to level of confidence

	Matched:Yes	Matched:No	Matched:Total	Controls:Used	Controls:Unused	Controls:Total
<hr/>						
Underconf.						
Treated	125	56	181	85	17	102
Untreated	85	17	102	125	56	181
Combined	210	73	283	210	73	283
<hr/>						
Realistic						
Treated	178	32	210	162	16	178
Untreated	162	16	178	178	32	210
Combined	340	48	388	340	48	388
<hr/>						
Overconf.						
Treated	188	25	213	189	0	189
Untreated	189	0	189	188	25	213
Combined	377	25	402	377	25	402

Table 14: Means and Standard Deviations of tournament performance by gender and performance rank in the realistic group (number of students in each cell is in parenthesis)

		Rank in tournament game (quartiles)				Total
		1	2	3	4	
male performance	M	11.62	8.51	7.55	5.10	9.86
	SD	3.61	2.30	3.25	1.60	3.72
	N	(97)	(53)	(22)	(10)	(182)
female performance	M	10.67	7.86	6.46	4.21	7.54
	SD	3.24	1.87	1.55	2.10	3.00
	N	(45)	(83)	(50)	(34)	(212)
Total	M	11.32	8.11	6.79	4.41	8.61
	SD	3.51	2.06	2.25	2.02	3.54
	N	(142)	(136)	(72)	(44)	(394)