ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

van Veldhuizen, Roel

Article — Published Version

Gender Differences in Tournament Choices: Risk Preferences, Overconfidence or Competitiveness?

Journal of the European Economic Association

Provided in Cooperation with:
WZB Berlin Social Science Center

Suggested Citation: van Veldhuizen, Roel (2022) : Gender Differences in Tournament Choices: Risk Preferences, Overconfidence or Competitiveness?, Journal of the European Economic Association, ISSN 1542-4766, Oxford University Press, Oxford, Vol. 20, Iss. 4, pp. 1595-1618, https://doi.org/10.1093/jeea/jvac031

This Version is available at:
https://hdl.handle.net/10419/261094

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

https://creativecommons.org/licenses/by/4.0/

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

GENDER DIFFERENCES IN TOURNAMENT CHOICES: RISK PREFERENCES, OVERCONFIDENCE, OR COMPETITIVENESS?

Roel van Veldhuizen
Lund University, Sweden and WZB
Berlin Social Science Center, Germany

Abstract
A long line of laboratory experiments has found that women are less likely to sort into competitive environments. Although part of this effect may be explained by gender differences in risk attitudes and self-confidence, previous studies have attributed the majority of the gender gap to gender differences in a competitiveness trait. I re-examine this result using a novel experiment that allows me to separate competitiveness from alternative explanations using causal treatments. In contradiction to the main conclusion drawn in a long literature, my results imply that the entire gender gap is driven by gender differences in risk attitudes and self-confidence, which has implications for policy and research. (JEL: J16, D01, C90)

Teaching Slides
A set of Teaching Slides to accompany this article are available online as Supplementary Data.

The editor in charge of this paper was Paola Giuliano.

Acknowledgments: I thank the editor (Paola Giuliano), three anonymous reviewers, Marina Agranov, Kai Barron, Vojtech Bartos, Nina Bonge, Yves Breitmoser, Thomas Buser, David Danz, Marie-Pierre Dargnies, Thomas Dohmen, Dirk Engelmann, Hande Erkut, Jana Friedrichsen, Uri Gneezy, Rustam Hakimov, Bob Hammond, Simone Haeckl, Håkan Holm, Macartan Humphreys, Botond Köszegi, Dorothea Kübler, John List, Johanna Mollerstrom, Muriel Niederle, Eva Ranehill, Maria Recalde, Martin Schonger, Andrew Schotter, Simeon Schudy, Erik Snowberg, Joep Sonnemans, Robert Stüber, Petra Thiemann, Bertil Tungodden, Joël van der Weele, Lise Vesterlund, Melinda Vigh[†], and Erik Wengström for valuable comments. I also thank Orkun Altun, Tobias Breunig, Miro Koeiy, Svetlana Povolotskaia, Petya Prodanova, and Renke Schmacker for excellent research assistance. Financial support from the Deutsche Forschungsgemeinschaft through CRC TRR 190 is gratefully acknowledged.

E-mail: roel.van_veldhuizen@nek.lu.se

Journal of the European Economic Association 2022 20(4):1595–1618
https://doi.org/10.1093/jeea/jvac031
© The Author(s) 2022. Published by Oxford University Press on behalf of European Economic Association. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

1. Introduction

Why do men and women differ in their labor market outcomes? In a seminal contribution, Niederle and Vesterlund (2007) propose that existing gender differences in the labor market may at least partially be driven by a greater female reluctance to enter competitive environments. Using a novel experimental paradigm in which participants work on an arithmetic task, they find that women are indeed less likely to enter a tournament. Importantly, they present evidence that much of this gender gap in tournament entry can be attributed to a novel psychological trait they label as "competitiveness", a trait that is distinct from risk preferences and overconfidence. This discovery started a new literature on gender and competitiveness that has been very influential.[1]

A key part of Niederle and Vesterlund (2007)'s contribution therefore lies in the evidence they provide supporting the existence of gender differences in a competitiveness trait. This evidence comes from regression analysis that controls for confounding variables like risk preferences, overconfidence, and performance. Figure 1 applies their identification strategy to their data and the data from 30 follow-up experiments. While controlling for confounding variables eliminates 28% of the gender gap in tournament entry on average, a large and significant fraction (72%) remains unexplained. This residual gap is what Niederle and Vesterlund (2007) interpret as the effect of the competitiveness trait (e.g. Niederle 2017).

However, the role of the competitiveness trait has recently come under scrutiny. In particular, Gillen, Snowberg, and Yariv (2019) demonstrate that measurement error in laboratory measures of risk attitudes and overconfidence leads to a systematic upward bias in the estimated importance of competitiveness. Intuitively, it is well known (e.g. Hausman 2001) that measurement error in risk attitudes and overconfidence implies that the coefficients for these variables are downward biased and inconsistently estimated. Since Niederle and Vesterlund (2007) identify competitiveness as the gender gap that remains after controlling for risk attitudes and overconfidence, underestimating the importance of these variables in turn implies overestimating the importance of the competitiveness trait. Gillen, Snowberg, and Yariv (2019) are able to adjust their estimates for measurement error econometrically by including additional control variables and using instrumental variable techniques, and find that after doing so, the estimated importance of competitiveness is small and no longer statistically different from zero.

While Gillen, Snowberg, and Yariv (2019)'s arguments are persuasive, their adjustments rely on assumptions of their own (such as the availability of good instruments), while maintaining most of the assumptions of Niederle and Vesterlund

1. Gillen, Snowberg, and Yariv (2019) refer to Niederle and Vesterlund (2007) as the "most influential experimental study of the last decade". In line with this, as of May 2022, Niederle and Vesterlund (2007) had over 3700 citations in Google scholar; numerous follow-up studies using their paradigm have been published in high-profile outlets (e.g. Gneezy, Leonard, and List 2009; Balafoutas and Sutter 2012; Buser, Niederle, and Oosterbeek 2014).

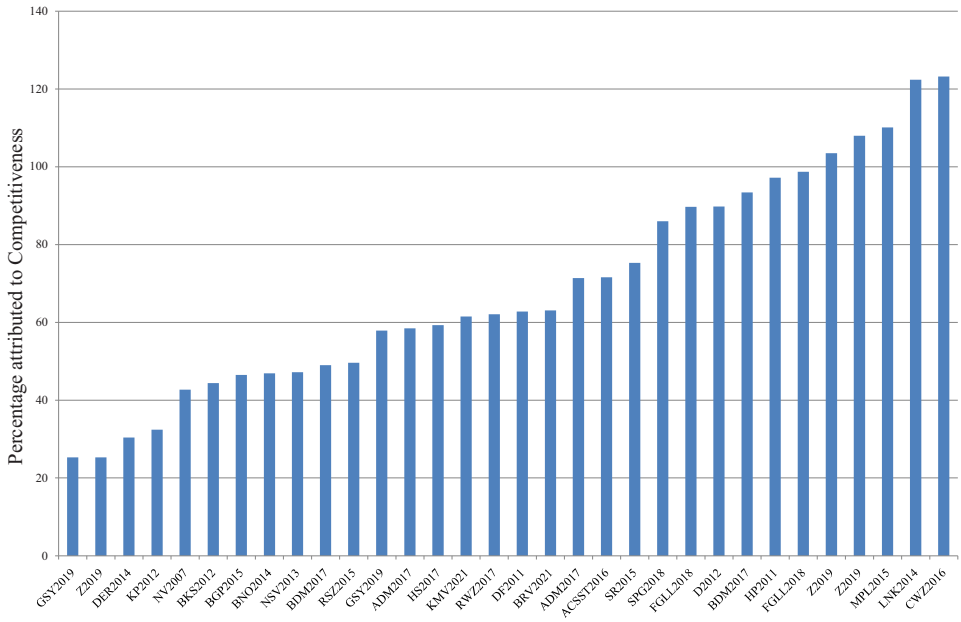

FIGURE 1. Previous estimates of the importance of competitiveness. The figure plots the percentage of the total gender difference in tournament entry that is attributed to gender differences in competitiveness using Niederle and Vesterlund's (2007) identification strategy. Each bar represents the result of a single experiment. The point estimates are based on my own calculations: each bar is the ratio between (a) the residual gender gap after controlling for risk attitudes, overconfidence, and performance and (b) the raw gender gap in tournament entry in that particular experiment. For an overview of the abbreviations and more details concerning the individual studies, see Online Appendix A6.

(2007)'s approach (such as linearity and the absence of omitted variables). Furthermore, if competitiveness is correlated with risk preferences, overconfidence or any other control variable, Gillen, Snowberg, and Yariv (2019)'s adjustments may even lead to *underestimating* the importance of competitiveness.[2] As a result, there is still considerable debate about whether the gender gap in tournament entry reflects well-known gender differences in risk aversion and confidence, or whether it can only be explained by invoking a competitiveness trait. Given the importance of this literature and the abundance of previous studies emphasizing the role of competitiveness, we therefore need further evidence that avoids both the criticisms raised by Gillen, Snowberg, and Yariv (2019) and the critiques of the solutions they propose.

2. In particular, if a control variable such as risk tolerance is positively correlated with competitiveness, then its coefficient will already capture part of the competitiveness effect. After adjusting for measurement error, the residual gender gap would then underestimate the true importance of competitiveness.

This is the purpose of this paper. Previous studies, including both Niederle and Vesterlund (2007) and Gillen, Snowberg, and Yariv (2019), identify the role of competitiveness by measuring the gender gap in a single context where it is assumed to play a role, and then partialling out the role of other factors using regression analysis (a "residual-based" approach). By contrast, I propose a novel identification strategy that identifies the role of competitiveness using causal treatments. My "treatment-based" approach relies on comparing two treatments that hold everything constant except for the feature assumed to be relevant for competitiveness to play a role: the presence of a competition. For this purpose, my competitive baseline treatment replicates Niederle and Vesterlund (2007)'s paradigm in which participants solve addition problems and choose whether to be paid according to piece rate or tournament incentives. I then compare behavior in this treatment to behavior in a non-competitive control treatment that holds the riskiness of the tournament entry decision as well as participants' subjective beliefs constant, but eliminates the role of competitiveness. In particular, participants in this second treatment choose between two non-competitive payment schemes, a fixed payment and a lottery, calibrated to match the payoff structure of the piece rate and tournament, respectively. By comparing the size of the baseline gender gap in tournament entry to the gender gap in the second decision, I can then identify the importance of the competitiveness trait without relying on Niederle and Vesterlund (2007)'s residual-based approach. More specifically, if competitiveness is a key driver of the gender gap in tournament entry, then removing its effect should lead to a significantly smaller gender gap in the non-competitive control treatment relative to the competitive baseline.

However, this is not what I find. While I am able to replicate the large gender gap in tournament entry observed in previous work, this gap remains almost identical in the control treatment where the role of competitiveness has been removed. In other words, my results imply that the gender gap in tournament entry is not driven by a gender difference in a competitiveness trait. Notably, I find similar results across two experiments with a total of 564 participants and using several robustness checks based on variations of my treatment-based identification strategy. Across all these comparisons, the point estimate for competitiveness is estimated at -8.6%, which is significantly smaller than the average found in Figure 1 (72%) and not significantly different from zero (the one-sided 95% confidence interval is $(-\infty, 16.2\%)$).

My main contribution lies in developing a causal treatment-based strategy that identifies competitiveness while avoiding the critiques raised against previous work. In particular, my identification strategy relies on a difference-in-difference test in which the two explanatory variables (gender and treatment) are perfectly measured so that measurement error in the x-variable cannot affect my results. I also present robustness checks showing that my results cannot be explained by order effects or other differences in the way my treatments are elicited. Finally, I use additional treatments to further decompose the gender gap in tournament entry, showing that approximately half can be attributed to risk attitudes, with the remainder being due to gender differences in overconfidence and performance.

The remainder of this paper proceeds as follows. The next section presents a brief theoretical framework for tournament entry decisions and introduces my identification strategy. I present the design and results of an initial experiment in Section 3 and a follow-up experiment in Section 4, respectively. In Section 5, I discuss how my identification strategy may be affected by measurement error in subjective beliefs and potential treatment differences in attitudes toward risk. This section also includes a pooled estimate of the importance of competitiveness based on multiple data sets and presents an additional treatment aimed at distinguishing the effects of confidence and risk attitudes. Section 6 concludes.

2. Theoretical Framework

Consider the decision problem faced by a participant i choosing between tournament and piece rate pay. In Niederle and Vesterlund (2007), the piece rate pays 50 cents per correct answer, whereas the tournament pays $2 per correct answer for winners and zero otherwise. Suppose participant i expects to solve x_i exercises and expects to win the tournament with some subjective probability p_i^s. Her choice will then be as follows:

Tournament entry (baseline)

Piece rate	Tournament
$0.5x_i$	p_i^s chance of getting $2x_i$

Standard expected utility theory predicts that participant i will choose the tournament if

$$p_i^s u_i(2x_i) > u_i(0.5x_i). \tag{1}$$

This requires the participant to be sufficiently confident (p_i^s large enough) and not too risk averse (as reflected by the curvature of her utility function u_i). Note that a high level of confidence can be due to either high ability or overconfidence. To incorporate competitiveness into this framework, I assume that tournament payoffs are evaluated through a different "competitive" utility function, $u_i^T()$. In this case, participant i chooses the tournament if

$$p_i^s u_i^T(2x_i) > u_i(0.5x_i). \tag{2}$$

If competitiveness is unimportant, then $u_i^T(x_i) = u_i(x_i)$, and hence equations (1) and (2) are identical. By contrast, if competitiveness is important, then they may differ. For example, for a participant who is sufficiently competitive, it is possible that $p_i^s u_i^T(2x_i) > u_i(0.5x_i)$, even when $p_i^s u_i(2x_i) < u_i(0.5x_i)$.

To identify the importance of competitiveness, I compare the baseline gender gap in tournament entry to the gender gap in a control treatment that

removes the effect of competitiveness by design but keeps the riskiness of the environment and the subjective probability p_i^s constant. In particular, this treatment (treatment NoComp for "non-competitive") presents participant i with the following choice:

Treatment NoComp

Fixed amount	**Lottery**
$0.5x_i$	p_i^s chance of getting $2x_i$

Similar to the baseline tournament entry decision, participant i chooses between obtaining $0.5x_i$ with certainty and obtaining $2x_i$ with probability p_i^s.[3] The key difference is that the second option is now a *lottery* instead of a tournament. Assuming that lotteries are not regarded as competitions, treatment NoComp therefore removes the role of competitiveness. Hence, irrespective of her competitiveness, participant i in treatment NoComp chooses the lottery if

$$p_i^s u_i(2x_i) > u_i(0.5x_i). \tag{3}$$

My focus lies in explaining gender differences. If competitiveness is unimportant, then the condition for choosing the lottery in treatment NoComp (equation (3)) is identical to the condition for choosing the tournament in the baseline (equation (2)). In this case, the gender difference in treatment NoComp and the baseline should be identical. By contrast, if competitiveness matters in the way suggested by the literature—namely that women are less competitive than men—I obtain (for a given performance x_i and with W indicating women's utility and M men's):

$$u_W^T(2x_i) - u_W(2x_i) < u_M^T(2x_i) - u_M(2x_i).$$

If competitiveness is important, then it is easy to see that the gender difference should then be smaller in NoComp. Intuitively, transforming the tournament into a non-competitive lottery makes it less attractive to competitive types and more attractive to the competition-averse. If the former group is composed primarily of men and the latter primarily of women (as suggested by Niederle and Vesterlund 2007), then more women and fewer men will choose the lottery in NoComp. Hence, provided that the subjective probability and riskiness of the environment are kept constant across the two treatments, I can identify the importance of competitiveness by comparing the gender gap across treatment NoComp and the baseline. In Online Appendix A3.4, I show that this remains true under models of non-expected utility such as prospect theory. A formal discussion of the identifying assumptions is presented in Online Appendix A7.

3. In the experiment, I elicit p_i^s using a separate belief elicitation task. The details of the belief elicitation task and the other procedures involved in constructing treatment NoComp are presented in the next section.

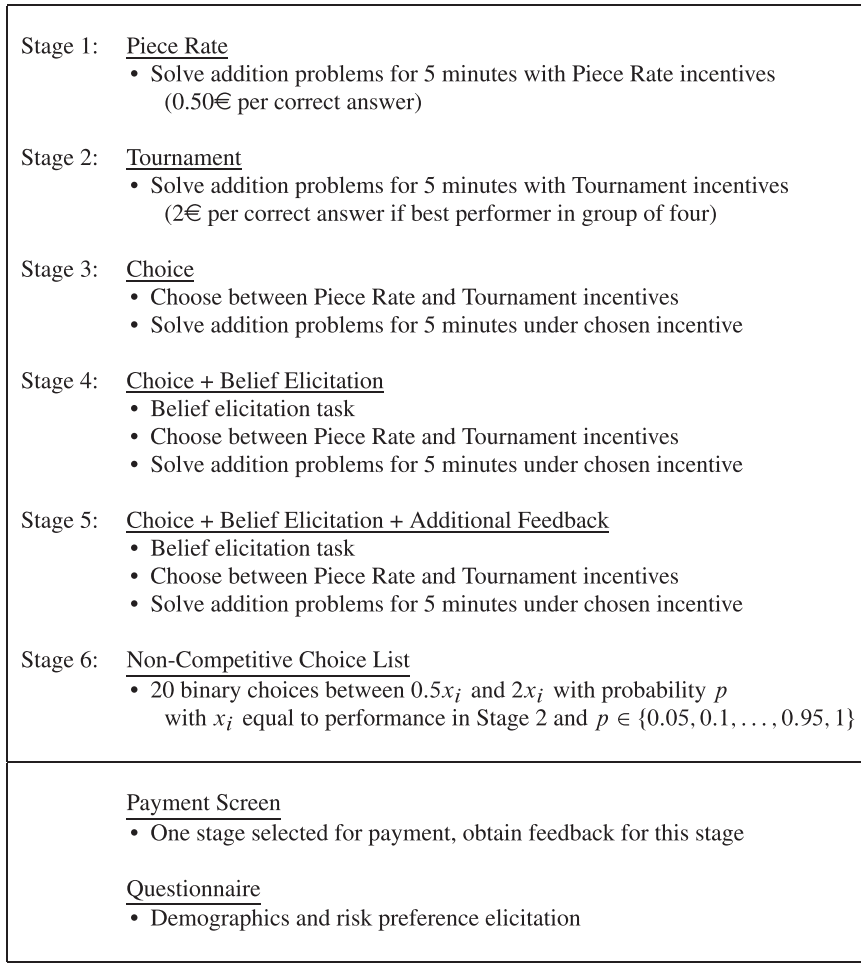

FIGURE 2. Timeline for experiment 1.

3. Experiment 1

3.1. Experimental Design

Experiment 1 consisted of six stages plus a questionnaire (see Figure 2). In the first five stages, participants solved addition problems under different incentive schemes. Stages 4 and 5 also included a belief elicitation task. Stage 6 then presented participants with 20 binary choices between a fixed sum of money and a lottery. Every participant took part in all six stages (always in the same order), and one stage was randomly selected for payment at the end of the experiment.

3.1.1. Replication of Niederle and Vesterlund (2007). To obtain a baseline measure of tournament entry, the experiment started with a replication of the first three stages

of Niederle and Vesterlund (2007). In each stage, participants had five minutes to solve addition problems consisting of five two-digit numbers. Within a given stage, each participant faced the same sequence of problems. After participants submitted their answer, they learned whether it was correct and were simultaneously presented with the next addition problem.

The three stages differed only in their incentive schemes. In *Stage 1 (Piece Rate)*, participants were paid 50 cents per correct answer. In *Stage 2 (Tournament)*, participants were matched into groups of four. In each group, the top performer was paid €2 for each correct answer. Second, third, and fourth-placed participants did not receive any payment. In case of a tie, the computer randomly drew one of the top performers as the winner.

In *Stage 3 (Choice)*, participants had to choose whether they wanted to apply piece rate or tournament incentives to their next performance. Tournament incentives were such that participants earned €2 per correct answer in case their score exceeded the score of their teammates *in Stage 2*. This feature is a key element of Niederle and Vesterlund (2007)'s design for two reasons. First, it guarantees that participants' choices in Stage 3 do not impose externalities on the earnings of other participants. Second, it removes all strategic considerations. This implies that Stage 3 can effectively be seen as an individual decision problem.

Stage 3 gives me a baseline measure of tournament entry. I follow previous work by including Stages 1 and 2 to obtain a measure of participants' ability under both types of incentives and to give participants some experience prior to making their choice in Stage 3. Participants were informed about their individual performance at the end of each stage.

3.1.2. Additional Stages. To estimate the importance of competitiveness, I added three additional stages that were not part of Niederle and Vesterlund (2007). In *Stage 4 (Choice + Belief Elicitation)*, participants first took part in a task meant to elicit the subjective probability of winning p_i^s, then made their tournament entry decision, and then solved addition problems for five minutes. I elicited beliefs immediately prior to the tournament entry decision in order to minimize the effect of belief changes over the course of the experiment (e.g. as a response to learning or performance feedback).

The belief elicitation task itself required participants to specify the reservation probability (p^r) for which they were indifferent between the following two options:

1. Receiving €2 if their Stage 4 performance exceeded the Stage 2 performance of their teammates (i.e. was good enough to win the tournament).
2. Receiving €2 with probability $p^r \in \{0, 0.01, 0.02, \ldots, 0.99, 1\}$.

If Stage 4 was selected for payment, then participants received their earnings for the addition problems depending on whether they had chosen the piece rate or tournament, in a similar fashion to Stage 3. In addition, a random value p would be drawn for each participant. If p was above the reservation probability, then the participant would be paid according to a lottery with probability p. Otherwise, the participant would be paid

TABLE 1. Stage 6 choices.

	Option A	Option B
1	€0.5x_i	100% chance to obtain €2x_i; 0% chance to obtain €0
2	€0.5x_i	95% chance to obtain €2x_i; 5% chance to obtain €0
3	€0.5x_i	90% chance to obtain €2x_i; 10% chance to obtain €0
...	...	...
19	€0.5x_i	10% chance to obtain €2x_i; 90% chance to obtain €0
20	€0.5x_i	5% chance to obtain €2x_i; 95% chance to obtain €0

Notes: x_i was equal to performance in Stage 2 (the forced tournament). In practice, the average value of option A in experiment 1 ranged from €2 to €12.50, with an average of €5.35.

€2 if her performance in Stage 4 exceeded the performance of her teammates (i.e. was good enough to win the tournament). This procedure makes it incentive-compatible for expected utility maximizing participants to report a reservation probability equal to their subjective probability of winning p_i^s (Karni 2009). Risk-neutrality is not required. The mechanism itself was carefully explained following the wording used by Mobius et al. (2014), and understanding was tested using a comprehension question.[4]

Stage 5 differed from Stage 4 in only one way: if Stage 5 was selected for payment, then participants who chose the piece rate were told whether they would have won the tournament. This may matter if—as proposed as an alternative explanation by Niederle and Vesterlund (2007)—women are more likely to choose the piece rate as a way to avoid receiving relative performance feedback. However, since gender differences in feedback aversion are not discussed in subsequent papers and I find no evidence for them, I postpone the main discussion of the results of Stage 5 to Online Appendix A1.2.

In *Stage 6*, each participant i made 20 choices between a fixed amount €0.5x_i and a lottery that paid either €2x_i or nothing, as per Table 1. Here, x_i was equal to participant i's performance in a prior part of the experiment (Stage 2). In other words, the payments faced by a given participant i depended on that particular participant's performance x_i in a prior part of the experiment (Stage 2). This procedure ensures that the stake size in Stage 6 was similar to previous tournament entry decisions.[5] The win probability for the lottery varied from 1 for the first row to 0.05 for the twentieth row. If Stage 6 was selected for payment, then one of the 20 choices was randomly picked to be implemented.

3.1.3. Treatment NOCOMP. My identification strategy relies on comparing the gender gap in tournament entry to the gender gap in a second decision that keeps

4. The comprehension question asked participants what they should report as their reservation probability if they thought they had a 44% chance of winning the tournament. All participants were required to answer this question correctly before continuing the experiment.

5. I used Stage 2 because performance in Stage 2 cannot be affected by the choice of incentives. This is a typical approach in the literature, though in any case performance across Stages 1–5 is highly correlated ($0.75 < r < 0.86$ for each individual correlation in experiment 1).

the riskiness and subjective win probability constant but removes the effect of competitiveness. To obtain the second decision, I take the subjective belief elicited in Stage 4 as a measure of the subjective probability of winning (p_i^s). In Stage 6, I then look for the single decision row for which the win probability of the lottery most closely corresponds to the elicited belief p_i^s. The binary choice made in this single decision row is what I will refer to as treatment NoComp in the remainder of the paper.

For example, consider a participant i who expects to solve 12 addition problems and expects to have a 35% chance of winning the tournament ($p_i^s = 0.35$). When deciding to enter the tournament, this participant would then implicitly be choosing between a piece rate payment of €6 and a 35% chance of obtaining a tournament payment of €24. Assuming that the participant in fact solved 12 problems in Stage 2 (so that $x_i = 12$), Stage 6 for her would then consist of 20 decisions between a fixed payment of €6 ($0.5x_i$) and a lottery paying either €24 ($2x_i$) or nothing. I would then use the binary choice in the decision row with a lottery win chance of 35% as this participant's treatment NoComp decision in my analysis.[6]

Given that treatment NoComp is a single binary decision, why did I choose to present participants with a list of 20 decisions instead of just eliciting a single one? First, using a choice list provides me with rich data that make it possible to infer choices for the full range of probabilities. Second, using a choice list prevents the belief elicited in Stage 4 from directly affecting payment in Stage 6. Third, using a choice list lowers the similarity between Stage 6 and the baseline, reducing the influence of order effects caused by preferences for consistency (Cialdini 1984; Cialdini, Trost, and Newsom 1995; Falk and Zimmermann 2011) and similar phenomena.[7]

3.1.4. Remaining Procedures. The experiment was conducted at the experimental economics laboratory of the Technical University of Berlin in June 2014. Only participants who had taken part in five or fewer previous experiments and had never failed to show up for a previous experiment were allowed to register. There were six sessions, one with 20 participants and five with 24. Each session had an equal number of men and women, for a total of 140 participants (70 men and 70 women). The experiment was programmed using PHP/MySql, and participants were recruited using ORSEE (Greiner 2015). Participants were assigned to a random computer upon entering the laboratory. They received an €8 show-up fee for the experiment, and were told that they would have to complete six separate stages, one of which would

6. For participants for whom p_i^s was not a multiple of 0.05 (e.g. when $p_i^s = 0.54$), I instead took the average of the choice made in the closest rows (e.g. 0.5 and 0.55). For 98 participants in experiment 1 (70%), p_i^s was a multiple of 0.05. Out of the remaining 42 participants, 40 made identical choices in the two closest rows. I classify the other two participants as indifferent between the lottery and the fixed amount.

7. Previous research (see e.g. Harrison et al. 2005 or Andersson et al. 2016) suggests that choice lists responses may be biased toward the middle, perhaps because the midpoint serves as a cognitive default to cognitively uncertain individuals (Enke and Graeber 2021). In Section 5.2 and Online Appendix A3, I present several robustness checks suggesting that using a choice list did not affect my results.

be randomly selected for payment. Instructions for the respective stages were only provided after the previous stage had ended. English translations and the original German instructions can be found in Online Appendix B2 and C1, respectively.

After the end of Stage 6, a random participant in each session was asked to roll a die to determine the stage selected for payment. Participants then received feedback on their selected stage, but not the other stages. Feedback included absolute performance, total earnings and—when applicable—the outcome of the tournament and belief elicitation task. After receiving feedback, participants then went through a questionnaire containing basic demographic questions as well as the Holt and Laury (2002), Eckel and Grossman (2002) and SOEP measures (Dohmen et al. 2011) of risk preferences. The first two measures were incentivized.

Each session took approximately 90 minutes. Average earnings in the experiment were €21.73 with a minimum of €8.20 and a maximum of €75.40. A total of 98.6% of participants indicated that they were students, most commonly majoring in engineering (26%), economics (15%), or dual majoring in economics and engineering or mathematics (16%). The mean and median age of participants in the experiment was 24.

3.2. Results

3.2.1. Preliminary Results. There were no gender differences in performance in the forced piece rate (men: 9.03, women: 8.80; $p = 0.723$, t-test) and the forced tournament (men: 10.90, women: 10.51; $p = 0.569$, t-test), see Figure A1 in Online Appendix A1.5.[8] Based on Stage 2 performance, 34.3% of men and 37.1% of women would have maximized their expected payoffs by competing. Nevertheless, men (58.6%) were more likely to choose the tournament than women (27.1%) in Stage 3. The gender gap is 31.4%, which is comparable to Niederle and Vesterlund (2007), and significant ($p < 0.001$, Fisher's exact test). In Stage 4, men were still significantly more likely to choose the tournament (67.1% versus 34.3%, $p < 0.001$, Fisher's exact test). Since Stage 4 is more closely connected to the elicited beliefs used in treatment NoComp, I will therefore use Stage 4 as the baseline for comparisons in the main analysis. However, my results are identical when I use Stage 3 as the baseline instead.

As a next step, Table 2 replicates Niederle and Vesterlund (2007) and Gillen, Snowberg, and Yariv (2019)'s approach to residualize competitiveness using regressions. Controlling for standard measures of risk attitudes and beliefs (column (2)), controlling for treatment NoComp choices as a proxy for these variables (column (3)), or controlling for all available measures of risk attitudes and beliefs (column (4)) eliminates between 31.4% and 50.2% of the raw gender gap in tournament entry (printed in column (1)). Following Niederle and Vesterlund (2007), these results would imply that competitiveness explains between 49.8% and 68.6% of the gender gap in tournament entry. These results are in line with the estimates in Figure 1; a more

8. Unless otherwise indicated, all p-values reported in this paper are based on two-sided tests.

TABLE 2. Tournament entry regressions for experiment 1.

	Dependent variable: tournament entry (Stage 3)			
	(1)	(2)	(3)	(4)
Female	−0.310***	−0.157*	−0.213**	−0.154*
	(0.078)	(0.081)	(0.085)	(0.087)
Elicited belief in Stage 4		0.419*		
		(0.228)		
Eckel–Grossman		0.039*		
		(0.023)		
SOEP		0.058***		
		(0.016)		
Treatment NoComp			0.269***	0.109
			(0.092)	(0.137)
All risk measures				$F = 2.99$**
				$p = 0.021$
All confidence measures				$F = 0.44$
				$p = 0.647$
Constant	0.381*	−0.084	0.270	−0.009
	(0.206)	(0.195)	(0.190)	(0.250)
Competitiveness		50.8%	68.6%	49.8%
Ability controls	Yes	Yes	Yes	Yes
Observations	140	140	140	140

Notes: OLS estimates, robust standard errors in parentheses. Dependent variable is the Stage 3 choice of compensation scheme (1-tournament, 0-piece rate). Ability controls include performance in Stage 2, the difference between performance in Stage 2 and Stage 1, and the objective probability of winning the tournament given Stage 2 performance. Elicited belief in Stage 4 is the elicited subjective probability of winning from Stage 4. Eckel–Grossman and SOEP are the Eckel and Grossman (2002) and SOEP measures of risk preferences, respectively. Treatment NoComp is the choice made in NoComp (1-lottery, 0-fixed amount). In column (4), the risk measures include the Eckel–Grossman, Holt–Laury, and SOEP measures plus the number of risky choices taken in Stage 6; the confidence measures include beliefs elicited in Stages 4 and 5. The competitiveness estimate is equal to the ratio between the female coefficient in the respective column and the female coefficient in column (1). $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

detailed discussion of these results and several robustness checks can be found in Online Appendix A1.1.

Before moving on to the treatment estimate of competitiveness, it is worthwhile to examine how well participants understood the experiment. My data allow me to flag a lack of comprehension in three ways. The first is by flagging eight participants who violated expected utility in Stage 6, by either switching multiple times, switching in the wrong direction, or by preferring $0.5x_i$ over a 100% chance of receiving $2x_i$. The second and third are by flagging participants who displayed either unrealistically high levels of risk tolerance (three participants) or overconfidence (six participants).[9] In the next section, I will refer to the sample that includes only the 123 participants without

9. Extremely risk tolerant participants are those who preferred a 5% chance of obtaining $2x_i$ over a certain payment of $0.5x_i$ in Stage 6. Extreme overconfidence requires a participant to be in the top 5% of overconfidence, which amounts to overestimating their probability of winning by at least 75 percentage points.

TABLE 3. Treatment estimate of competitiveness in experiment 1.

	Dependent variable: compensation scheme choice		
	(1)	(2)	(3)
Female	−0.329***	−0.316***	−0.318***
	(0.081)	(0.084)	(0.086)
NoComp	0.014	0.014	−0.016
	(0.072)	(0.073)	(0.079)
Female × NoComp	−0.043	−0.030	−0.016
	(0.096)	(0.100)	(0.105)
Constant	0.671***	0.681***	0.672***
	(0.057)	(0.057)	(0.060)
Competitiveness	−13.0%	−9.6%	−5.0%
p (comp ≥56.4%)	0.009***	0.020**	0.032**
EU violators	Yes	No	No
Extreme risk prefs	Yes	Yes	No
Extreme confidence	Yes	Yes	No
Observations	280	264	246
N (men)	70	69	61
N (women)	70	63	62

Notes: OLS estimates, standard errors clustered at the participant level in parentheses. The dependent variable contains two observations per individual: their tournament entry decision in Stage 4 (1-tournament, 0-piece rate) and their NoComp decision (1-lottery, 0-fixed amount). The independent variables are binary variables for gender (1-female, 0-male), treatment (1-NoComp, 0-baseline) and their interaction. "Competitiveness" is the point estimate for the importance of competitiveness, computed as the negative of the ratio between the first and third coefficient. "p (comp ≥ 56.4%)" provides the result of a one-sided Wald test investigating whether the estimated importance of competitiveness is greater than 56.4% (the average estimate across columns (2)–(4) in Table 2). The first column includes all observations; the second removes participants who violated expected utility. The third column is the preferred sample used in the main analysis in the text, which also removes participants with extreme risk preferences or extreme levels of overconfidence. More details regarding these exclusion criteria are presented at the end of Section 3.2.1. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

violations as the "preferred sample", and will use it for the analysis reported in the main text. The results for the full sample are very similar, however, and presented in Table 3.

3.2.2. Treatment Estimate of Competitiveness. I identify the importance of competitiveness by comparing the gender gap in tournament entry to the gender gap in treatment NoComp. If competitiveness is important, then the gender gap should be smaller in treatment NoComp. However, this is not what I find (see Figure 3). Instead, men in the preferred sample (67.2%) were still significantly more likely than women (33.9%) to choose the lottery ($p < 0.001$, Fisher's exact test). The size of the gender gap (33.3 percentage points) is not significantly smaller than in the baseline (31.8pp; $p = 0.560$, one-sided difference-in-difference test). If anything, it is slightly larger. The point estimate implies that competitiveness explains −5% (i.e. (31.8–33.3)/31.8)

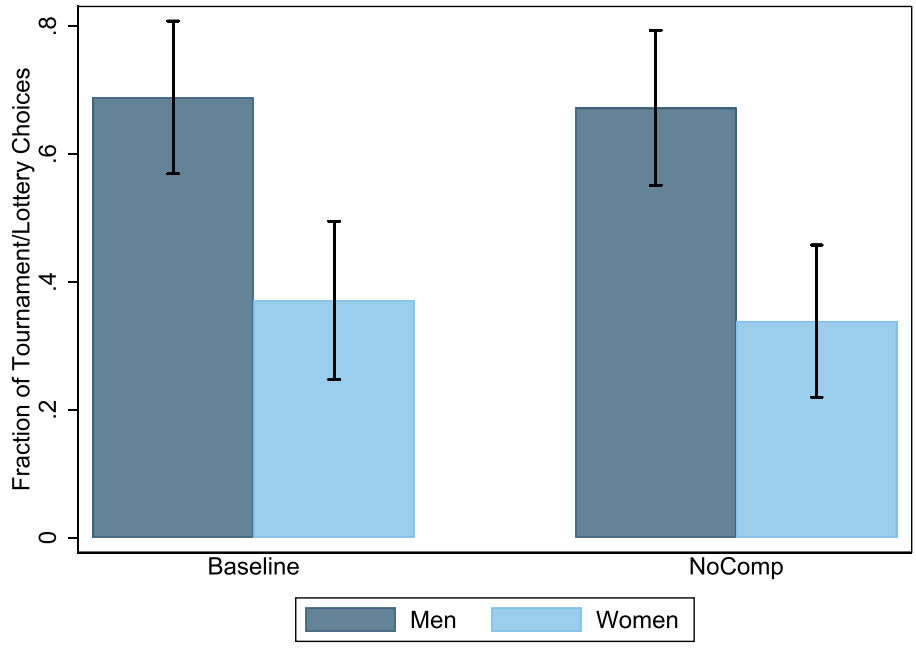

FIGURE 3. Summary of choices in experiment 1 by gender. This figure gives the fraction of participants choosing the tournament (Baseline) or lottery (NOCOMP) by gender in experiment 1, using Stage 4 as the baseline. The error bars represent 95% confidence intervals. The figure excludes 17 participants who violated expected utility or displayed extreme levels of overconfidence or risk aversion.

of the baseline gender gap in tournament entry. Keeping in mind that −5% is not significantly different from 0%, this estimate suggests that, if anything, women are more competitive and, hence, the other factors have to explain more than the whole raw gender gap to compensate for the effect of competitiveness. Table 3 shows that I obtain very similar results if I use the full sample (column (1)) or a sample that only removes participants who violate expected utility theory (column (2)).

It is worth emphasizing that these results differ considerably from the ones obtained using Niederle and Vesterlund's (2007) residual-based identification strategy (Table 2). Indeed, I reject the hypothesis that competitiveness explains 56.4% of the gender difference in tournament choices, as implied by the average estimate in Table 2 ($p = 0.032$, one-sided Wald test). Instead, my treatment estimate suggests that the gender difference in tournament entry can be explained by gender differences in risk preferences and subjective beliefs (confidence); competitiveness is unimportant. In Section 5.4 below, I use an additional control treatment to decompose the gender gap into the effects of risk aversion and confidence.

4. Experiment 2

4.1. Experimental Design

The purpose of experiment 2 was to replicate experiment 1 with a larger sample size while randomizing the order of tasks and adding additional stages to be used as robustness checks. There were four main changes compared to experiment 1. First, experiment 2 had a larger sample size (213 men and 211 women) to ensure that it could pick up even modest effect sizes of competitiveness (50% of the gender gap in tournament entry or less) with high probability (a power of up to 0.91); the power calculation is presented in Online Appendix A8. Second, experiment 2 randomized the order of Stages 3–6 in order to control for order effects. Third, experiment 2 replaced the old Stage 5 with a new Stage 5 in which participants made a binary decision between a lottery and a fixed payment. Fourth, experiment 2 added another tournament entry decision (Stage 7) for which participants were told their true win probability based on prior performance before making their entry decision. These two additional stages will be used in the robustness checks that I describe in greater detail in Section 5. A more detailed description of the design of experiment 2 can be found in Online Appendix B1.

4.2. Results

4.2.1. Preliminary Results. Men performed a little better in both the forced piece rate (men: 8.23, women: 7.58; $p = 0.083$, t-test) and the forced tournament (men: 10.37, women: 9.62; $p = 0.066$, t-test).[10] Based on Stage 2 performance, 44.6% of men and 36.5% of women would have maximized their expected payoffs by competing. The actual gender gap in tournament entry in Stage 3 is larger, however, with men (60.1%) being more likely to choose the tournament than women (33.6%). The gender gap is 26.4 percentage points ($p < 0.001$, Fisher's exact test). Table 4 replicates Niederle and Vesterlund (2007)'s approach to residualize competitiveness using regressions; the resulting point estimate for competitiveness ranges from 77.4% to 90.0%.

In Stage 4, male tournament entry was still 60.1%, but female tournament entry increased slightly to 42.7% ($p = 0.071$, Fisher's exact test). The resulting gender gap is 17.4 percentage points ($p < 0.001$, Fisher's exact test). These results tentatively suggest that encouraging women to think through their win chance before choosing their payment scheme makes them more willing to compete, though the lack of an a priori hypothesis and the relatively large p-value implies that this effect may also be due to chance. Because I use the elicited beliefs from Stage 4 to construct treatment

10. Gender differences in performance are sometimes observed in similar experiments (e.g. Niederle, Segal, and Vesterlund 2013) and have also been observed in other experiments run at the same laboratory (Buser, Ranehill, and Van Veldhuizen 2021; Kessel, Mollerstrom, and Van Veldhuizen 2021).

TABLE 4. Tournament entry regressions for experiment 2.

	Dependent variable: tournament entry (Stage 3)			
	(1)	(2)	(3)	(4)
Female	−0.239***	−0.206***	−0.215***	−0.185***
	(0.047)	(0.047)	(0.047)	(0.047)
Elicited belief in Stage 4		0.423***		
		(0.114)		
Eckel–Grossman		0.010		
		(0.014)		
SOEP		0.034***		
		(0.009)		
Treatment NoComp			0.204***	0.080
			(0.047)	(0.074)
All risk measures				$F = 3.15^{**}$
				$p = 0.014$
All confidence measures				$F = 8.52^{***}$
				$p < 0.001$
Constant	0.561***	0.179	0.493***	0.207
	(0.101)	(0.132)	(0.109)	(0.145)
Competitiveness		86.0%	90.0%	77.4%
Ability controls	Yes	Yes	Yes	Yes
Observations	424	424	424	424

Notes: OLS estimates, robust standard errors in parentheses. Dependent variable is the Stage 3 choice of compensation scheme (1-tournament, 0-piece rate). For variable definitions, see the notes to Table 2. In column (4), the confidence measures include the beliefs elicited in Stage 4 and prior to Stage 7. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

NoComp, I will use Stage 4 as the baseline for tournament entry in this section, similar to experiment 1.

In experiment 2, 11 participants (2.6%) violated expected utility in Stage 6, 20 participants (4.7%) are classified as extremely overconfident, and 32 participants (7.5%) are classified as extremely risk loving using similar criteria as in experiment 1. In addition, experiment 2 also allows me to identify 40 participants (9.4%) who appear to have had some trouble understanding the instructions.[11] To be consistent with experiment 1, the analysis in the next section will use a "preferred sample" that excludes the 60 participants (14.2%) who violated expected utility or were extremely overconfident or risk tolerant. The results for the full sample and the more restrictive

11. In defining risk-loving participants, the only difference in experiment 2 is that I define extreme risk-loving participants as those who preferred a 5% or smaller chance of obtaining $2x_i$ over a certain payment of $0.5x_i$ in *either Stage 5 or Stage 6*. When it comes to trouble with understanding, five participants were flagged by the experimenter during the session. The remaining 35 participants were in the top 5% of either the number of attempts or the total time needed to answer the comprehension question in Stage 4, variables that were only saved in the database in experiment 2.

TABLE 5. Treatment estimate of competitiveness in experiment 2.

	Dependent variable: compensation scheme choice			
	(1)	(2)	(3)	(4)
Female	−0.174***	−0.170***	−0.165***	−0.182***
	(0.048)	(0.049)	(0.052)	(0.054)
NoComp	0.026	0.041	0.030	0.038
	(0.038)	(0.038)	(0.040)	(0.042)
Female × NoComp	0.043	0.027	−0.002	−0.014
	(0.054)	(0.054)	(0.057)	(0.060)
Constant	0.601***	0.599***	0.582***	0.600***
	(0.034)	(0.034)	(0.036)	(0.038)
Competitiveness	24.6%	16.0%	−1.3%	−7.7%
p (comp ≥ 86.0%)	0.024**	0.015**	0.006**	0.002***
EU violations	Yes	No	No	No
Extreme risk prefs	Yes	Yes	No	No
Extreme confidence	Yes	Yes	No	No
Low understanding	Yes	Yes	Yes	No
Observations	848	824	728	670
N (men)	213	207	184	170
N (women)	211	205	180	165

Notes. OLS estimates, standard errors clustered at the participant level in parentheses. For variable definitions, see the notes to Table 3. "p (comp≥ 86.0%)" provides the result of a one-sided Wald test investigating whether the estimated importance of competitiveness is greater than 86.0% (the average estimate across columns (2)–(4) in Table 4). The first column includes all observations; the second removes participants who violated expected utility in Stage 5 or 6. The third column is the preferred sample used in the main analysis in the text, which also removes participants with extreme risk preferences or extreme levels of overconfidence. The fourth column also removes participants with low understanding. More details regarding these exclusion criteria are presented at the end of Section 4.2.1. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

sample that also excludes the participants with limited understanding are presented in Table 5.

4.2.2. Treatment Estimate of Competitiveness. Figure 4 displays the fraction of men and women choosing the tournament in Stage 4 and the lottery in treatment NoComp, respectively. In treatment NoComp, 61.1% of men and 44.4% of women chose the risky option ($p < 0.001$, Fisher's exact test). The resulting gender gap (16.7 percentage points) is not significantly smaller than in the baseline (16.5 percentage points in the preferred sample; $p = 0.515$, one-sided difference-in-difference test). The point estimate implies that −1.3% (i.e. (16.5–16.7)/16.5) of the gender gap in tournament choices can be explained by a competitiveness trait, which is significantly less than the average estimate of 86.0% obtained using Niederle and Vesterlund (2007)'s approach of residualizing competitiveness (Table 4, $p = 0.006$, one-sided Wald test), and very similar to the point estimate in experiment 1 (−5%). Table 5 shows that my results remain similar if I use a more or less restrictive sample.

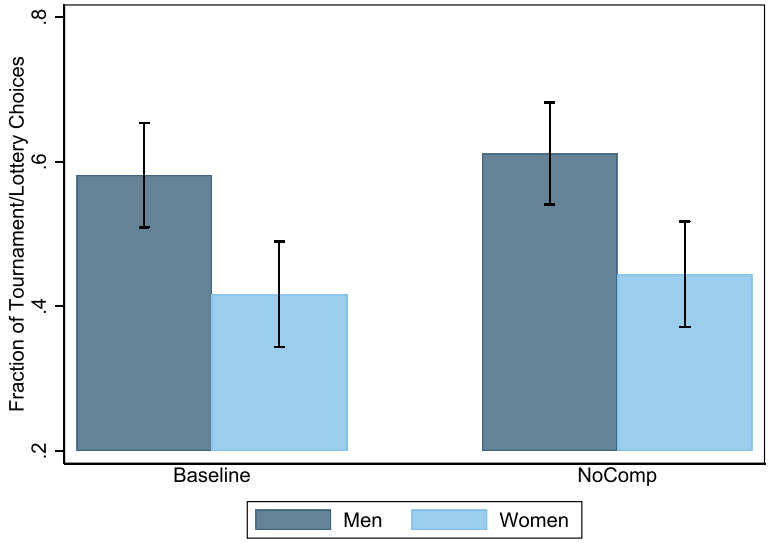

FIGURE 4. Summary of choices in experiment 2 by gender. This figure gives the fraction of participants choosing the tournament (baseline) or lottery (NOCOMP) by gender in experiment 2, using Stage 4 as the baseline. The error bars represent 95% confidence intervals. The figure excludes 60 participants who violated expected utility or displayed extreme levels of overconfidence or risk aversion.

Overall, the key result in experiment 2 is similar to experiment 1: the gender gap in tournament entry does not appear to be driven by a gender difference in a competitiveness trait. The results of experiment 2 further demonstrate that this result is robust to randomizing the order of stages, which implies that the limited importance attributed to competitiveness in experiment 1 is not due to order effects.[12] A more detailed comparison of the results of the two experiments is presented in Online Appendix A1.3.

5. Discussion

The main identifying assumption of my identification strategy is that treatment NOCOMP removes the effect of competitiveness but keeps the riskiness of the environment and the subjective probability of winning the same as in the tournament entry decision. Is this assumption reasonable? The literature treats competitiveness as a preference for being in a competitive environment, such as a tournament.

12. A direct test of order effects in experiment 2 demonstrates that whether Stage 6 came before or after Stage 4 did not have a significant impact on the estimated importance of competitiveness ($p = 0.305$, difference-in-difference test); the same is true for the order of Stages 3 and 4 ($p = 0.391$, difference-in-difference test). The full results of these tests are presented in Online Appendix A1.4.

Lotteries are not typically considered to be competitive. Hence, it seems reasonable that competitiveness cannot explain participants' choices in treatment NoComp.

In the remainder of this section, I will more closely examine the two other elements of the identifying assumption: the subjective probability of winning and the riskiness of the environment. I will then combine the data from experiments 1 and 2 and several robustness checks to provide a pooled estimate of the importance of competitiveness. Finally, I will further decompose the gender gap in tournament entry into the effects of overconfidence, performance, and risk preferences.

5.1. Measurement Error in Elicited Beliefs

The elicited subjective win probabilities I use to construct treatment NoComp are measured with error. Gillen, Snowberg, and Yariv (2019) show that measurement error generates an upward bias in the importance of competitiveness when its effect is identified as the residual gender gap in a regression that controls for risk attitudes, performance, and confidence. By contrast, my treatment-based estimate is based on a difference-in-difference (treatment*gender) test, where the tournament entry and treatment NoComp decisions serve as the y-variable, and gender and treatment indicators (and their interaction) serve as the x-variables. Importantly, whereas classical measurement error in the x-variables biases the coefficient estimates (e.g. Gillen, Snowberg, and Yariv 2019), classical measurement error in the y-variable increases the variance of the coefficient estimates but does not generate a bias (e.g. Hausman 2001). Assuming that neither gender nor treatment are measured with error, classical measurement error will therefore not bias my results. I present a more extensive version of this argument in Online Appendix A2.1 and support it with simulations in Online Appendix A2.2. In Online Appendix A2.3, I further show that adding additional measurement error to elicited beliefs does not change the estimated importance of competitiveness.

It is also possible to bypass these issues entirely by using an identification strategy that does not rely on eliciting subjective beliefs. For this purpose, I included a new stage in experiment 2 (Stage 7) in which, prior to making a tournament entry decision, participants were informed about their true "objective" probability of winning the tournament given their performance in Stage 2. Assuming that participants base their entry decisions on this objective win probability, I can then identify competitiveness by comparing the gender gap in Stage 7 to a version of treatment NoComp based on the same objective win probability (treatment NoComp7).[13] If measurement error in beliefs was a key determinant of my main results, then I should observe a significantly larger estimate for competitiveness in this alternative test, which does not rely on eliciting beliefs. Instead, however, the point estimate for competitiveness is still small (4.1%) and not significant ($p = 0.456$, one-sided difference-in-difference test).

13. The only difference between treatment NoComp7 and treatment NoComp is that I select the relevant row in Stage 6 using the "objective" probability that a participant wins the tournament given their Stage 7 performance, instead of using the subjective probability of winning elicited in Stage 4.

This implies that measurement error in elicited beliefs does not explain the limited importance attributed to competitiveness in my main analysis. I present a more detailed discussion of this comparison and its results in Online Appendix A2.4.

5.2. Risk and Tournament Entry

My identification strategy also requires that treatment NoComp and the tournament entry decision contain the same amount of risk. To achieve this, I calibrated the stake size and lottery win probability in treatment NoComp to closely approximate the stakes and subjective win probability in the tournament entry decision, as explained in Section 3.1.3. Nevertheless, some differences remain between the two choices. In particular, the tournament entry decision is a binary choice, in which winning is determined by performance in a real effort task (a "social risk"). By contrast, treatment NoComp uses a different elicitation method (a price list) whereby the winner is determined using a computerized random draw (a "nature risk"). These differences in the elicitation method and the type of risk could conceivably impact my results if they affect men and women in different ways.

To test whether this was the case, I conducted two robustness checks that harmonized the elicitation method and the source of risk across the tournament entry and treatment NoComp decisions. The first robustness check elicited both decisions using a binary choice, whereas the second elicited both decisions using a price list. The second robustness check also included a real effort task in both treatments and further harmonized the way uncertainty was resolved by using a computerized random draw in both treatments. The first robustness check uses data from experiment 2 (Stages 5 and 7), whereas the second uses data from a follow-up experiment (Bartos and Van Veldhuizen 2022).

Whereas the results of the first robustness check are inconclusive due to the win probability not being equal in both treatments (see Online Appendix 3.1), the second robustness check shows that harmonizing the source of risk and elicitation method does not increase the estimated importance of competitiveness. The point estimate is -28.8% $(-6.2/21.6)$, which is not significantly greater than zero ($p = 0.908$, one-sided difference-in-difference test). I present a more detailed overview of the design and results of these robustness checks in Online Appendix A3, where I also show that my estimates are robust to a potential "midpoint bias" in price lists (whereby choices in price lists are biased toward the midpoint; see e.g. Andersson et al. 2016) and remain valid under non-expected utility.

5.3. Collecting Evidence

The data from both experiments can also be combined to further increase the power of the test for the importance of competitiveness. The first row in Table 6 shows that the resulting point estimate is -2.9% with a one-sided 95% confidence interval of $(-\infty, 37.4\%)$. To further increase the sample size, the second row adds the data from the robustness check discussed in Section 5.1, whereas the third row also adds the

TABLE 6. Estimates of competitiveness using pooled data.

	Estimate	Sample size
(1) Main comparison from experiments 1 and 2	−2.9%	487 (487)
	$(-\infty, 37.4\%)$	
(2) All comparisons from experiments 1 and 2	−0.6%	851 (487)
	$(-\infty, 35.8\%)$	
(3) All data	−8.6%	1491 (1127)
	$(-\infty, 16.2\%)$	

Notes: The first column presents point estimates for the importance of competitiveness, with one-sided 95% confidence intervals clustered at the participant level in parentheses. The second column presents the number of relevant observations (within-subject treatment comparisons); the term in brackets presents the number of independent observations (i.e. participants). The first row includes only the baseline versus treatment NOCOMP comparison in the present study, pooled across the two experiments, removing participants who violated EU in Stage 5 (experiment 2) or 6 (both experiments) or displayed extreme risk preferences or extreme overconfidence. The second row adds the data from the "beliefless" robustness check in experiment 2, described in Section 5.1 and Online Appendix A2.4. The third row adds the data from Bartos and Van Veldhuizen (2022) and Kessel, Mollerstrom, and Van Veldhuizen (2021).

results from two follow-up experiments (Bartos and Van Veldhuizen 2022; Kessel, Mollerstrom, and Van Veldhuizen 2021) that are discussed in greater detail in Online Appendix A4. When including all available data, the point estimate for competitiveness is −8.6% with a confidence interval of $(-\infty, 16.2\%)$. Hence, these estimates allow me to rule out all but the smallest effects of competitiveness with 95% certainty. In Online Appendix A4, I show that these results are robust to including the comparison between the two binary choices in Stages 5 and 7 discussed in the previous section, and are also consistent with other types of treatment effects reported in previous work.

5.4. Decomposing the Gender Gap in Tournament Entry

What explains the gender gap in tournament entry if not competitiveness? Previous work and the theoretical framework in Section 2 point to two potential mechanisms: gender differences in confidence (the subjective probability of winning p_i^s) and gender differences in risk attitudes (the curvature of the utility function). To distinguish between these mechanisms, I use a second control treatment that is similar to treatment NOCOMP but also removes the role of gender differences in confidence (treatment JUSTRISK). Intuitively, men had more optimistic elicited beliefs than women (57.6% versus 50.2% win chance; $p < 0.0001$, t-test). Since elicited beliefs directly determine the win probability for the lottery in treatment NOCOMP, men also faced lotteries with a higher win probability in treatment NOCOMP. Treatment JUSTRISK eliminates this gender difference by using equally attractive lotteries for both genders. This implies that I can attribute any residual gender gap in treatment JUSTRISK to gender differences in risk preferences. This in turn allows me to identify the importance of confidence by comparing the gender gap across treatments JUSTRISK and NOCOMP. To maximize the sample size, I use the pooled data from both experiments. I present a more detailed description of the design of treatment JUSTRISK in Online Appendix A5.1.

When only risk preferences can explain the gender gap (treatment JUSTRISK), 52.7% of men and 41.7% of women chose the lottery (Figure A10 in Online Appendix A5, $p = 0.014$, t-test). The resulting gender gap (10.9pp) is significantly smaller than in treatment NOCOMP (20.9pp; $p < 0.0001$, one-sided difference-in-difference test). Given that the gender gap in tournament entry is 20.3 percentage points, these point estimates imply that 49.2% $((20.9 - 10.9)/20.3)$ of the gender difference in tournament entry is driven by confidence, and 53.7% $(10.9/20.3)$ by risk attitudes (and −2.9% by competitiveness, as per the first row in Table 6). In other words, gender differences in confidence and risk attitudes each explain approximately half of the gender gap in tournament entry; competitiveness appears to play no role. Further analysis demonstrates that of the total effect of confidence (49.2%), 29.3 percentage points are driven by gender differences in performance; the remaining effect (19.9pp) can be attributed to gender differences in overconfidence (see Online Appendix A5.2 for more details).

6. Conclusion

Starting with Niederle and Vesterlund (2007), a long literature has interpreted the existence of large gender differences in tournament entry as evidence of a gender difference in a competitiveness trait. However, previous estimates of the importance of this trait are based on an identification strategy that is known to be susceptible to measurement error. My main contribution lies in developing a new identification strategy that avoids the critiques raised against previous work and provides an unbiased estimate of the importance of competitiveness under reasonable assumptions. Consistent with Gillen, Snowberg, and Yariv (2019), I find that the gender gap in tournament entry can be explained without invoking a competitiveness trait. Instead, I show that all of the gender gap in tournament entry can be explained by gender differences in risk attitudes, overconfidence, and performance. My results are consistent across two separate experiments and a number of robustness checks.

The limited importance of competitiveness has implications for policy and future research. In terms of policy, it suggests that institutional changes that reduce the role of competitiveness may not be as effective as previously thought (this is consistent with Flory, Leibbrandt, and List 2015). Instead, my results suggest that institutional changes that limit the role of factors such as risk attitudes and confidence, by reducing payment uncertainty or uncertainty about (relative) ability, for example, are more likely to reduce the gender gap. In terms of future research, my results imply that attempts to better understand or reduce gender differences in labor market outcomes would do well to focus on confidence and risk attitudes rather than competitiveness.

My results may also be of interest to researchers concerned about the potential impact of measurement error in experiments. In particular, they demonstrate that causal treatments can be used to directly identify the importance of key variables by exogenously changing the role played by these variables. When such treatments are feasible, they may serve as a powerful alternative to the econometric adjustments

proposed by Gillen, Snowberg, and Yariv (2019). A comparison of the comparative advantages of the two approaches would be an intriguing topic for future research.

References

Andersson, Ola, Håkan J. Holm, Jean-Robert Tyran, and Erik Wengström (2016). "Risk Aversion Relates to Cognitive Ability: Preferences or Noise?" *Journal of the European Economic Association*, 14, 1129–1154.

Balafoutas, Loukas and Mattias Sutter (2012). "Affirmative Action Policies Promote Women and Do Not Harm Efficiency in the Laboratory." *Science*, 335, 579–582.

Bartos, Vojtech and Roel Van Veldhuizen (2022). "Non-Parametric Identification of Preferences through Experiments." Working paper, Lund University.

Buser, Thomas, Muriel Niederle, and Hessel Oosterbeek (2014). "Gender, Competitiveness, and Career Choices." *The Quarterly Journal of Economics*, 129, 1409–1447.

Buser, Thomas, Eva Ranehill, and Roel Van Veldhuizen (2021). "Gender Differences in Willingness to Compete: The Role of Public Observability." *Journal of Economic Psychology*, 83, 102366.

Cialdini, Robert (1984). *Influence, the Psychology of Persuasion*. Harper Collins, New York, NY.

Cialdini, Robert B., Melanie R. Trost, and Jason T. Newsom (1995). "Preference for Consistency: The Development of a Valid Measure and the Discovery of Surprising Behavioral Implications." *Journal of Personality and Social Psychology*, 69, 318–328.

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G. Wagner (2011). "Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences." *Journal of the European Economic Association*, 9, 522–550.

Eckel, Catherine C. and Philip J. Grossman (2002). "Sex Differences and Statistical Stereotyping in Attitudes toward Financial Risk." *Evolution and Human Behavior*, 23, 281–295.

Enke, Benjamin and Thomas Graeber (2021). "Cognitive Uncertainty." NBER Working Paper No. 26518.

Falk, Armin and Florian Zimmermann (2011). "Preferences for Consistency." IZA Discussion Paper No. 5840.

Flory, Jeff A., Andreas Leibbrandt, and John A. List (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.

Gillen, Ben, Erik Snowberg, and Leeat Yariv (2019). "Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study." *Journal of Political Economy*, 127, 1826–1863.

Gneezy, Uri, Kenneth L. Leonard, and John A. List (2009). "Gender Differences in Competition: Evidence from a Matrilineal and a Patriarchal Society." *Econometrica*, 77, 1637–1664.

Greiner, Ben (2015). "Subject Pool Recruitment Procedures: Organizing Experiments with ORSEE." *Journal of the Economic Science Association*, 1, 114–125.

Harrison, Glenn W., Morten I. Lau, Elisabet E. Rutström, and Melonie B. Sullivan (2005). "Eliciting Risk and Time Preferences Using Field Experiments: Some Methodological Issues." In *Field Experiments in Economics*, edited by, G. W. Harrison, J. Carpenter, and J. A. List. Emerald, Greenwich, CT, pp. 17–50.

Hausman, Jerry (2001). "Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left." *Journal of Economic Perspectives*, 15, 57–67.

Holt, Charles A. and Susan K. Laury (2002). "Risk Aversion and Incentive Effects." *American Economic Review*, 92(5), 1644–1655.

Karni, Edi (2009). "A Mechanism for Eliciting Probabilities." *Econometrica*, 77, 603–606.

Kessel, Dany, Johanna Mollerstrom, and Roel Van Veldhuizen (2021). "Can Simple Advice Eliminate the Gender Gap in Willingness to Compete?" *European Economic Review*, 138, 103777.

Mobius, Markus, Muriel Niederle, Paul Niehaus, and Tanya Rosenblat (2014). "Managing Self-Confidence: Theory and Experimental Evidence." NBER Working Paper No. 17014.
Niederle, Muriel (2017). "A Gender Agenda: A Progress Report on Competitiveness." *American Economic Review*, 107(5), 115–119.
Niederle, Muriel, Carmit Segal, and Lise Vesterlund (2013). "How Costly Is Diversity? Affirmative Action in Light of Gender Differences in Competitiveness." *Management Science*, 59, 1–16.
Niederle, Muriel and Lise Vesterlund (2007). "Do Women Shy Away from Competition? Do Men Compete Too Much?" *Quarterly Journal of Economics*, 122, 1067–1101.

Supplementary data

Supplementary data are available at *JEEA* online.