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

Roel van Veldhuizen (WZB Berlin Social Science Center)

Discussion Paper No. 14

March 25, 2017

Collaborative Research Center Transregio 190 | www.rationality-and-competition.de
Ludwig-Maximilians-Universität München | Humboldt-Universität zu Berlin
Spokesperson: Prof. Dr. Klaus M. Schmidt, University of Munich, 80539 Munich, Germany
+49 (89) 2180 3405 | info@rationality-and-competition.de

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

Roel van Veldhuizen*

February 10, 2017

A large number of recent experimental studies show that women are less likely to sort into competitive environments. While part of this effect may be explained by gender differences in risk attitudes and overconfidence, previous studies have attributed the majority of the gender gap to gender differences in a separate 'competitiveness' trait. We re-examine this result using a novel experimental technique that allows us to separate competitiveness from alternative explanations by experimental design. In contrast to the literature, our results imply that the whole gender gap is driven by risk attitudes and overconfidence, which has important implications for future research.

*WZB Berlin Social Science Center, Reichpietschufer 50, 10785 Berlin, Germany (email: roel.vanveldhuizen@wzb.eu). I thank Marina Agranov, Yves Breitmoser, Thomas Buser, Marie-Pierre Dargnies, David Danz, Dirk Engelmann, Jana Friedrichsen, Uri Gneezy, Macartan Humphreys, Botond Köszegi, John List, Muriel Niederle, Eva Ranehill, Martin Schonger, Andrew Schotter, Erik Snowberg, Bertil Tungodden, Joël van der Weele, Lise Vesterlund and Melinda Vigh for valuable comments. I also thank Renke Schmacker for excellent research assistance. Financial support from the Deutsche Forschungsgemeinschaft through CRC TRR 190 is gratefully acknowledged.

I. Introduction

Increased gender parity in the labor market remains an important policy goal.[1] Women earn lower wages for similar positions, and are under-represented in positions of leadership (Goldin, 2014). In order to better understand and reduce the gender gap, it is important to know its causes. Traditional explanations focus on gender discrimination, family, preferences for certain occupations, and a number of other factors (see e.g., Goldin, 2014, for an overview).

More recently, a number of influential studies have noted that prestigious and lucrative jobs often take place in competitive environments. Starting with the seminal work of Niederle and Vesterlund (2007), these studies suggest that women have a lower willingness to seek out such environments. This idea has received support from a large number of laboratory experiments (see e.g., Almås et al., 2016; Dreber, von Essen, and Ranehill, 2014; Gillen, Snowberg, and Yariv, 2015; Reuben, Wiswall, and Zafar, 2015, and others).

But what is it about competition that attracts men but pushes women away? While part of the effect may be explained by gender differences in overconfidence and risk preferences, the current view is that the majority of the gap can be explained by gender differences in attitudes towards competition (see e.g., Niederle and Vesterlund, 2011; Niederle, 2016; Buser, Niederle, and Oosterbeek, 2014, or the literature review below). Indeed, the identification of 'competitiveness' as a separate trait – independent of risk preferences, overconfidence, and other factors – is often regarded as a key contribution of this literature (Bertrand, 2011; Gillen, Snowberg, and Yariv, 2015). This is particularly evident in recent work that examines whether laboratory mea-

1. Many organizations and governments have set up specialized task forces, including the US government, the European Union, and the World Economic Forum. Examples of affirmative action policies include the Lilly Ledbetter Fair Pay Act of 2009 in the US, and Norway's policy that there is at least a 40% minimum of female board members.

sures of 'competitiveness' are predictive of real world labor market outcomes (Buser, Niederle, and Oosterbeek, 2014; Reuben, Sapienza, and Zingales, 2015; Berge et al., 2015).

By contrast, this study presents evidence suggesting that the importance of the competitiveness trait is considerably smaller than previously thought. We use a novel experimental approach that allows us to differentiate between risk preferences, overconfidence and competitiveness by experimental design. We start by replicating Niederle and Vesterlund's (2007) laboratory experiment in which participants are tasked with solving addition problems and have to choose whether to be paid according to piece rate or tournament incentives. We then directly compare this choice to several (within-subject) control treatments where competitiveness and/or overconfidence are controlled for by design.

The results are striking. Like previous studies, we find a large gender gap in competitive choices. But unlike these studies, our results imply that the *whole* gender gap is explained by gender differences in risk attitudes and overconfidence. Women are less confident and more risk averse than men, and this is what causes them to sort out of the tournament. Taken together, the results from our four treatments imply that gender differences in overconfidence, risk attitudes, and their interaction effect explain 48%, 28%, and 37% of the gender gap respectively. In sharp contrast to the literature, competitiveness explains -13% of the gender gap in competitive choices.

Why do our results differ so strongly from the existing literature? We identify the effect of competitiveness using direct treatment comparisons. By contrast, previous studies have used an indirect approach, where risk preferences and overconfidence are controlled for using regressions, and the residual gender gap is then attributed to competitiveness. We compare the two approaches using our data and the data from seven previous experiments, and show that the regression-based method overestimates the

importance of competitiveness by approximately 50 percentage points. We therefore contribute to a recent discussion highlighting the pitfalls of controlling for important confounds using regressions (Green, Ha, and Bullock, 2010; Gillen, Snowberg, and Yariv, 2015; Westfall and Yarkoni, 2016), and propose a new way to circumvent these issues by experimental design.

Our results have important implications. The idea that women are less competitive than men has been very influential.[2] Our results imply that the gender gap in tournament choices is instead a manifestation of gender differences in risk attitudes and overconfidence, not a separate competitiveness trait.[3] This implies that future research and policies aiming to better understand or reduce gender differences in labor market outcomes would do well to focus on overconfidence, risk attitudes, and other factors not captured by these experiments, rather than competitiveness.

These results are also important because risk preferences and overconfidence (i.e., beliefs) are the main ingredients of decision theory. A long research tradition in this area has greatly expanded our knowledge of how these factors can best be modeled and measured, and has greatly increased our understanding of individual differences. This has made it possible to do structural analyses, counterfactual predictions, theoretical evaluations, policy recommendations, etc. By contrast, we are not aware of any study that attempts to model competitiveness, or indeed measures it directly. In this sense,

2. As of January 2017, Niederle and Vesterlund (2007) has been cited more than 1,500 times (Google Scholar). The idea has also received wide media coverage, for recent examples see www.washingtonpost.com/news/storyline/wp/2015/01/02/why-do-some-studies-show-that-women-are-less-competitive-then-men/
or www.economist.com/news/finance-and-economics/21692938-lesbians-tend-earn-more-heterosexual-women-girl-power.

3. The importance of separating competitiveness from competing explanations has previously been noted by Bertrand (2011), who writes that "Future research in this area should also aim to confirm that the gender differences in performance in competitive settings and willingness to enter competitive settings are more than just a reflection of already identified gender differences, such as attitudes towards risk and overconfidence." For a similar argument see Flory, Leibbrandt, and List (2015).

our results are therefore reassuring. Rather than requiring further research into a novel, not yet well-understood concept (competitiveness), they imply that policy and future research can address the gender gap using the well-understood and much-tested theory on decision under risk and uncertainty.

Hence, we contribute to the literature in three ways. First, we present a novel experimental technique to directly identify the effect of competitiveness and separate it from overconfidence and risk attitudes. Second, we show that the gender difference in tournament choices is caused by gender differences in risk preferences and overconfidence, and not a separate competitiveness trait, which has important implications for policy and future research. Third, we introduce and illustrate the general usefulness of direct treatment comparisons as a way to avoid the problems inherent to the regression-based identification strategy that has become standard in much of the experimental literature.

II. Literature Review

This study is motivated by a large literature that studies gender differences in competitive choices. Most evidence comes from laboratory experiments, typically variations of Niederle and Vesterlund's (2007) seminal design. In these experiments, participants work on real-effort tasks and are asked to choose their remuneration scheme. Typically, participants have two options, a piece rate and a winner-takes-all tournament. The tournament involves actively competing against other participants, and is therefore considered to be the more competitive environment. The standard result is that women are significantly less likely to choose the tournament. This result has been replicated in a large number of studies, discussed below.

There are at least three reasons why men and women may differ in their tendency

to sort into the tournament. First, the tournament is riskier than the piece rate. Participants who choose the winner-takes-all tournament end up with nothing if they fail to win. There is considerable evidence that women are more risk averse than men, though the exact gap varies in size and significance (Croson and Gneezy, 2009; Filippin and Crosetto, 2016). Gender differences in **risk preferences** therefore provide one potential explanation as to why women are less likely to opt for a risky tournament environment.

A second explanation is that men and women differ in their beliefs or **overconfidence**. Previous experiments investigating gender differences in competitive choices almost universally find that men are more overconfident than women. For example, Niederle and Vesterlund (2007) find that 75% of men and 43% of women think they are the best performer in a four-person group. Being more overconfident, men are more optimistic about their chances of success, which could in turn make them more likely to choose the tournament.

Finally, it has been argued that men may also be more competitive than women. In this view, **competitiveness** (i.e., a preference for competition) is seen as a trait that is independent from factors such as risk preferences or overconfidence (see e.g., Niederle, Segal, and Vesterlund, 2013; Buser, Niederle, and Oosterbeek, 2014; Gneezy, Pietrasz, and Saccardo, 2016). Thus, even in cases where men and women are equally risk tolerant and equally confident in their abilities, women may still be less likely to select into competitive environments because they lack the required taste for tournaments. This idea was first introduced by Niederle and Vesterlund (2007) and has since become very influential.[4]

The popularity of the competitiveness explanation is based on the claim that it has

4. Additional explanations for the gender gap in tournament choices include gender differences in ability, social preferences, ambiguity attitudes, and feedback aversion. We cover these explanations in more detail in the discussion.

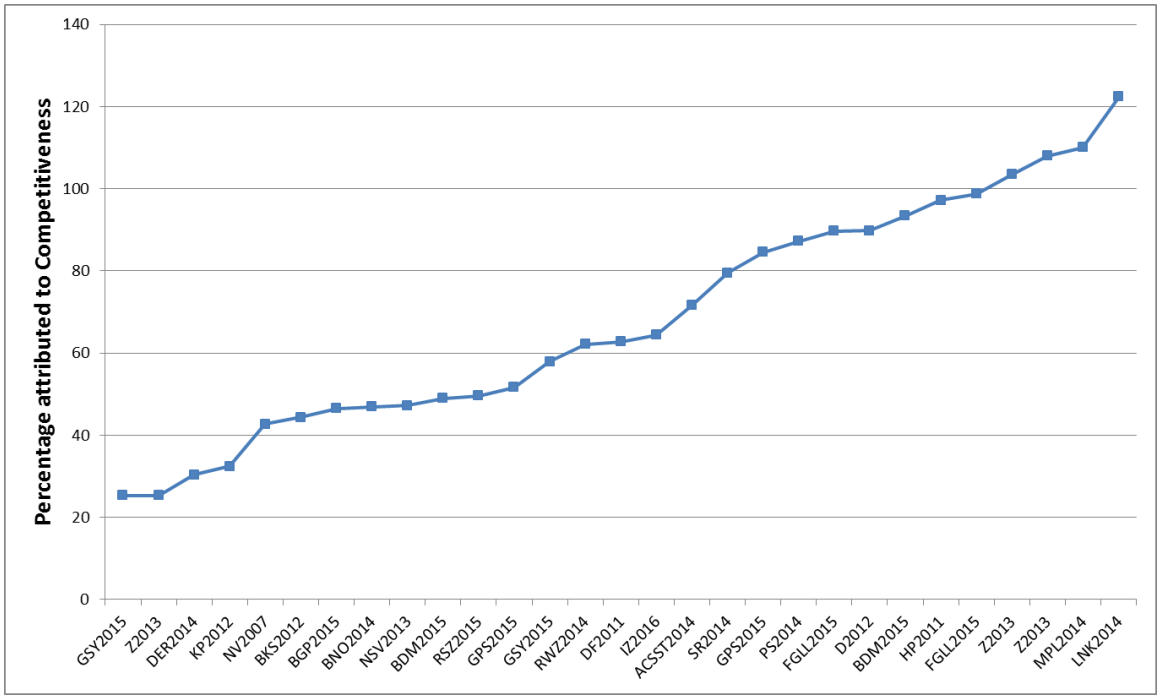

FIGURE I: PREVIOUS ESTIMATES OF THE IMPORTANCE OF COMPETITIVENESS

Notes. The figure plots the fraction of the total gender difference in tournament choices that is attributed to gender differences in competitiveness. Each dot represents the result of a single experiment. For more details concerning the individual studies, see Table VI in Appendix A2.

large explanatory power above and beyond risk attitudes and beliefs/overconfidence. Niederle and Vesterlund (2007) support this claim by econometrically controlling for laboratory measures of risk attitudes and beliefs. While these variables explain part of the gender gap, a large and significant gender gap remains. They interpret this residual gender gap as the effect of competitiveness. Specifically, they conclude that "the residual 'competitive' component is 43 percent" (p. 1096) of the original gender gap.

A similar approach is used in a large number of studies, summarized in Figure I. On average, controlling for overconfidence and risk preferences eliminates 31% of the gender difference in tournament choices. The residual gender gap (69%) is then attributed to gender differences in competitiveness.

The finding that competitiveness explains most of the gender gap in tournament choices is an important result. Risk attitudes and beliefs (and hence overconfidence) are the main ingredients of standard decision theory. Both they and their gender difference are therefore relatively well understood. By contrast, the competitiveness explanation is relatively novel, having first been introduced by Niederle and Vesterlund in 2007. Its implied importance therefore created an immediate need for follow-ups investigating its drivers and correlates, and ways in which its effect could be mitigated through policy design. It has also allowed the literature to start linking tournament choices directly to real-world outcomes (Buser, Niederle, and Oosterbeek, 2014; Reuben, Sapienza, and Zingales, 2015; Berge et al., 2015), and to interpret ensuing correlations as evidence that a 'competitiveness trait' can explain gender differences in these real-world outcomes.

To our knowledge, the regression-based method summarized in Figure I is the only method used in the literature to quantify the importance of competitiveness. However, it is not entirely without controversy. It relies heavily on the ability of regression techniques to successfully control for every relevant factor (other than competitiveness). In practice, it may not be possible to control for every relevant variable and, even when it is, measurement error and misspecification may bias the results (Hausman, 2001; Green, Ha, and Bullock, 2010; Gerber and Green, 2012; Westfall and Yarkoni, 2016). In particular, Gillen, Snowberg, and Yariv (2015) show that measurement error in laboratory measures of risk attitudes and beliefs leads to a systematic upward bias in the implied importance of competitiveness. If these effects are important, the true impact of competitiveness may therefore be substantially smaller than the 69% implied by Figure I.

Given the importance of this literature and the potential concerns with regression-based identification techniques, it therefore seems critical that the importance of com-

petitiveness is established using a different method. This is the purpose of our paper. We distinguish competitiveness from alternative explanations by directly comparing tournament choices to individual-specific control treatments. This allows us to estimate the impact of competitiveness on tournament choices, while circumventing the concerns raised by Gillen, Snowberg, and Yariv (2015) and others.

III. Design

The experiment consisted of six tasks plus a questionnaire (see Figure II). The first three tasks were a direct replication of Niederle and Vesterlund (2007). The remaining tasks were used to differentiate between the three competing explanations.

The experiment was conducted at the experimental economics laboratory of the Technical University of Berlin. 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 tasks, one of which would be randomly selected for payment. Instructions for the respective tasks were only provided after the previous task had ended, feedback on earnings and performance was only provided at the end of the experiment. All instructions can be found in Appendix B.

III.A. Replication

The first three tasks in the experiment were identical to Tasks 1–3 in Niederle and Vesterlund (2007) and many subsequent studies. In each task, participants had

FIGURE II: EXPERIMENTAL TIMELINE

Task 1: Piece Rate
- Solve exercises for 5 minutes
- Piece Rate incentives: 0.50 Euro per correct answer (x_i)

Task 2: Tournament
- Solve exercises for 5 minutes
- Tournament incentives: $2x_i$ if best performer in group of four

Task 3: Choice
- Choose between Piece Rate and Tournament incentives
- Solve exercises for 5 minutes

Task 4: Belief Elicitation
- Belief elicitation task
- Choose between Piece Rate and Tournament incentives
- Solve exercises for 5 minutes

Task 4b: Universal Feedback
- Discussed in Appendix A1

Task 5: Multiple Choice List
- Choose between $0.5x_i$ and $2x_i$ with probability p
- 20 choices, for $p \in \{0.05, 0.1, \ldots, 0.95, 1\}$

Payment Screen
- One task selected for payment, obtain feedback for this task

Questionnaire
- Demographics and risk preference elicitation

five minutes to solve addition problems consisting of five two-digit numbers. Each participant faced the same sequence of problems in each task, which was randomly determined before the first session. After participants submitted their answer, they learned whether it was correct and were simultaneously presented with the next exercise.

The three tasks differed only in their incentive schemes. In *Task 1 (Piece Rate)*, participants were paid 50 cents per correct answer. In *Task 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 *Task 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 Task 2*. This guaranteed that participants' actions in Task 3 did not impose externalities on the earnings of other participants.

III.B. Identifying Competitiveness

Our identification strategy in this experiment relies on comparing the choices made in Task 3 (the 'baseline') to three individual-specific control treatments. We start by explaining the idea behind each of these treatments, and how they allow us to differentiate competitiveness from overconfidence and risk preferences. We then describe the procedures of Task 4 and Task 5, and explain how we use these tasks to elicit each control treatment.

As a starting point, we consider the decision process of a participant i in Task 3.

Suppose she expects to solve x_i exercises, and expects to win the tournament with some subjective probability p_i^s. This implies that her choice will be as follows.

Baseline (Task 3)

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:

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

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). Competitiveness, by contrast, implies that tournament payoffs are evaluated through a different utility function $U_T()$. In this case, participant i chooses the tournament if:

$$p_i^s U_T(2x_i) > U(0.5x_i) \tag{2}$$

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

To identify the importance of competitiveness, we compare the baseline to our first control treatment, which presents participant i with the following choice:

Treatment NoComp

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

Treatment NoComp is a non-competitive version of the baseline. As with the baseline, the choice is between obtaining $0.5x_i$ with certainty and obtaining $2x_i$ with probability p_i^s. The key difference is that the second option is now a *lottery* instead of a tournament. Otherwise, the payoffs are constructed to closely approximate the payoffs of the baseline.

Note that while we present treatment NoComp here in a way that makes the similarity to the baseline somewhat obvious, this was not the case in the experiment. The exact procedure we use to elicit treatment NoComp is described in section III.F. below.

Our identification strategy assumes that competitiveness can explain choices in the baseline, but not in treatment NoComp, while the effect of risk attitudes and overconfidence is identical in the two cases. The former is achieved by transforming the right option into a lottery, which is no longer competitive. The latter is achieved by constructing treatment NoComp to have the same payoff structure as the baseline. A more formal discussion of our identifying assumption is presented in the Discussion section and Appendix A4.

Hence, irrespective of competitiveness, participant i in treatment NoComp chooses the lottery if:

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

We are interested in gender differences. If competitiveness is unimportant, the choice

(3) faced in treatment NoComp is identical to the choice (1) faced in the baseline. In this case, the gender difference in treatment NoComp and the baseline should be identical. By contrast, the literature tells us that competitiveness matters, and women W are less competitive than men M, i.e.,

$$U_{T,W}(2x_i) - U_W(2x_i) < U_{T,M}(2x_i) - U_M(2x_i)$$

It is easy to see that the gender difference should then be smaller in treatment 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 the literature), more women and fewer men should choose the lottery in treatment NoComp. Hence, we can identify the importance of competitiveness by comparing the gender gap across treatment NoComp and the baseline.

III.C. Risk Preferences and Overconfidence

To also distinguish between overconfidence (i.e., beliefs) and risk preferences, we present participants with a second alternative:

Treatment JustRisk

Fixed Amount	**Lottery**
$0.5x_i$	p_i^{o} chance of getting $2x_i$

Treatment JustRisk is similar to NoComp, except that the lottery is based on participant i's true objective probability of winning the tournament (p_i^{o}). Replacing p_i^{s} with p_i^{o} eliminates the effect of overconfidence (which is defined as $p_i^{\text{s}} - p_i^{\text{o}}$). This is important, because the literature tells us that men are more overconfident than

women (i.e., $p_M^s - p_M^o > p_W^s - p_W^o > 0$). By eliminating the effect of overconfidence, we can compare treatment JustRisk to treatment NoComp to separate overconfidence from risk preferences.

Specifically, replacing p_i^s with p_i^o makes the lottery less attractive to overconfident participants. While this makes both genders less prone to choose the lottery if overconfidence is important, the effect will be larger for men, who are likely to be more overconfident. If overconfidence is important, the gender difference will therefore be smaller in treatment JustRisk than in treatment NoComp.

By eliminating the other two explanations, treatment JustRisk also allows us to establish the importance of risk preferences. If risk preferences are important, there should still be a gender gap in treatment JustRisk. Even when both genders face equally attractive lotteries, men will be more likely to choose them if they are more tolerant to taking risks. We can therefore use the gender gap in treatment JustRisk as an estimate of the importance of risk preferences.

III.D. Interaction Effects

By using linear regressions and related techniques such as probit, the literature has implicitly assumed that (latent) tournament choices are a *linear* function of risk preferences, beliefs, and competitiveness. In practice, however, the relationship may be non-linear and characterized by interaction effects.

In fact, one such interaction effect follows naturally from standard expected utility theory (equation (1)). Utility and beliefs enter multiplicatively into the expected utility function, and utility itself is a non-linear function of risk preferences. This implies that risk preferences will matter more for participants with intermediate beliefs. Intuitively, pessimistic participants (p_i^s close to zero) will always choose the piece rate,

regardless of their risk preferences, because it has a higher expected value and is less risky. Similarly, for optimistic participants (p_i^{s} close to one) the expected utility of the tournament will exceed the piece rate's for all but the most risk averse participants. Expected utility theory therefore predicts a non-linear interaction effect between risk preferences and beliefs, where risk preferences matter most for 'intermediate' beliefs.

This discussion implies that choices in treatment JustRisk may differ from treatment NoComp for two reasons. First, treatment JustRisk eliminates the direct effect of gender differences in overconfidence (since $p_M^{\text{o}} = p_W^{\text{o}}$). Second, treatment JustRisk also eliminates overconfidence per se (since $p^{\text{o}} < p^{\text{s}}$), which makes the lotteries faced by participants less attractive. This may change the role of risk preferences, which may in turn affect the size of the gender difference in observed choices.

To separate these two effects, we run a final treatment, treatment IntEff:

Treatment IntEff

Fixed Amount	**Lottery**
$0.5x_i$	p_i^{n} chance of getting $2x_i$

Here, p_i^{n} can be thought of as the gender-neutral version of participant i's subjective belief. Specifically, we construct p_i^{n} to eliminate gender differences in overconfidence ($p_M^{\text{n}} = p_W^{\text{n}}$), without eliminating overconfidence per se ($p_i^{\text{n}} > p_i^{\text{o}}$ still). One way to do this is to replace the subjective beliefs p_i^{s} of all men with beliefs drawn from the belief distribution of the women. More details on how we constructed p_i^{n} are presented in section III.F.

Treatment IntEff allows us to separate the direct effect of overconfidence from its interaction with risk preferences. Relative to NoComp, treatment IntEff eliminates the gender difference in overconfidence. Hence, if men are more overconfident and the direct effect of overconfidence is important, the gender gap in treatment IntEff

should be smaller than in treatment NoComp.

The comparison between treatment IntEff and treatment JustRisk then allows us to identify the importance of the interaction effect. Relative to treatment IntEff, treatment JustRisk also eliminates overconfidence per se. The interaction effect predicts that this decreases the importance of risk preferences for overconfident participants who have intermediate subjective probabilities p_i^s (say .4 to .7) but are actually not very good (a p_i^o of less than .25). Given that the median participant has a probability of winning of approximately $(0.5)^3$, we expect this to be true for most participants. Hence, if the interaction effect is important, we expect the gender gap to be larger in treatment IntEff than in treatment JustRisk. To our knowledge, we are the first to examine this interaction effect in the literature on gender differences in tournament choices.

III.E. Overview and Predictions

Table I summarizes the four treatments. Comparing the gender gap in the baseline to treatment NoComp gives us the importance of competitiveness. The comparison between NoComp and IntEff gives the importance of overconfidence, and the comparison between JustRisk and IntEff gives the importance of the interaction between risk preferences and overconfidence. The residual gender gap in treatment JustRisk provides an estimate of the effect of risk preferences.

We can use the results of previous experiments to predict the gender gap in each of the four treatments (Figure III). If competitiveness indeed explains 69% of the gender gap in the baseline, as suggested by the literature, the gender gap in treatment NoComp should be 69% smaller than the baseline. Overconfidence is typically found to have some effect as well, explaining why treatment IntEff's predicted gender gap

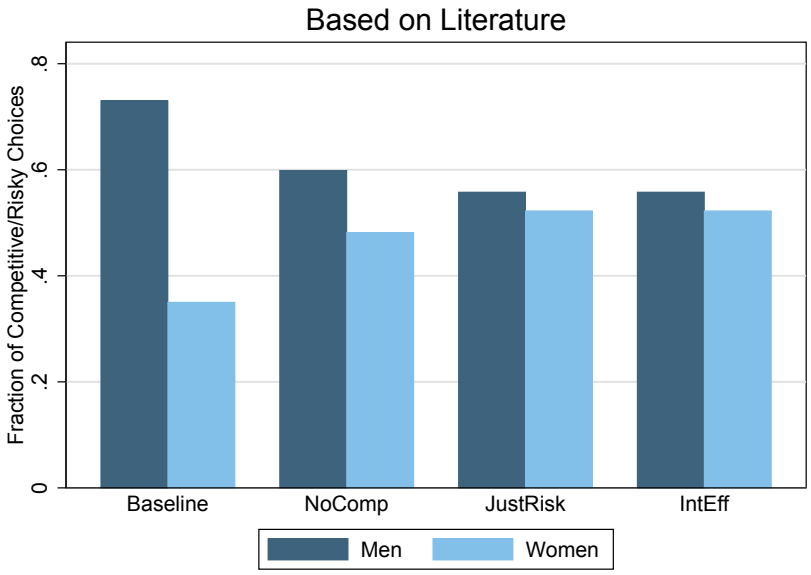

FIGURE III: PREDICTED CHOICES BY GENDER

Notes. This figure gives the predicted fraction of men and women choosing the tournament or lottery in each of the four treatments, based on the literature. The size of the baseline gender gap is taken from Niederle and Vesterlund (2007). For treatment NoComp, we reduce the baseline gender gap by the percentage attributed to competitiveness in the studies reviewed in Figure I and Table VI (69.2%). For treatment JustRisk and IntEff, we reduce the gender gap by the percentage attributed to overconfidence (21.5%, determined by an analysis similar to Table VI; full results available on request). Since the risk overconfidence interaction has not been previously studied, the gender gap in treatment IntEff is predicted to be identical to treatment JustRisk.

TABLE I: The Four Treatments

Treatment	Safe	Risky	Mechanisms
Baseline	$0.5x_i$ €	Tournament: $2x_i$ € with probability p_i^{s}	C, R, O, R*O
NoComp	$0.5x_i$ €	Lottery: $2x_i$ €with probability p_i^{s}	R, O, R*O
JustRisk	$0.5x_i$ €	Lottery: $2x_i$ €with probability p_i^{o}	R
IntEff	$0.5x_i$ €	Lottery: $2x_i$ €with probability p_i^{n}	R, R*O

Notes. This table displays the four treatments studied in the experiment. The baseline ask participants to choose between a piece rate and a tournament. The other treatments ask participants to choose between a safe prospect and a risky lottery. p_i^{s}, p_i^{o}, and p_i^{n} are the subjective, objective and gender-neutral probability of winning respectively. The final column summarizes the mechanisms that can explain the gender gap in each row. The four mechanisms are Competitiveness (C), Risk Preferences (R), Overconfidence (O), and the interaction of the latter two (R*O).

is even smaller. Risk preferences typically matter little, as reflected by the nearly non-existent gender gap in treatment JustRisk. The interaction effect has not been studied in the literature and is therefore assumed to have no effect, which implies that the gender gaps in treatment JustRisk and IntEff are identical.

III.F. Procedures

We will now move on to explain exactly how we elicited the choices in treatments, NoComp, JustRisk and IntEff. Doing so required us to acquire the values for the parameters x_i, p_i^{s}, p_i^{o} and p_i^{n} and obtain participants' choices given these parameters.

III.F.1. Belief Elicitation

We elicited the subjective probability of winning p_i^s using a belief elicitation task taken from Mobius et al. (2014). Our goal was to elicit a precise measure of the subjective probability of winning p_i^s. For this purpose, we required an incentive-compatible procedure with sufficient scope for variation. Another requirement was for the elicitation technique to not be confounded by risk preferences. The reservation probability or "crossover" method used by Mobius et al. (2014) meets both criteria. In a recent survey of the literature, Schlag, Tremewan, and van der Weele (2015) recommended the use of this method as a way to prevent bias resulting from risk aversion.

In order to minimize the effect of belief changes over the course of the experiment (as a response to learning or performance feedback), we elicited beliefs directly before participants chose between tournament and piece rate. Since the belief elicitation procedure is not trivial, a potential concern is that the elicitation task may detract participants from the choice between tournament and piece rate. To test whether such effects were relevant, we included both a choice without belief elicitation (Task 3) and one with belief elicitation (Task 4).

In *Task 4 (Belief Elicitation)*, participants were asked to choose between piece rate and tournament incentives, and then solved addition problems for five minutes, similar to Task 3. In addition, just before participants chose their preferred incentive, they were first asked to indicate their subjective belief. A comparison between choices in Task 3 and Task 4 allows us to see whether the belief elicitation procedure per se influenced participants' decisions.

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 they win the tournament (i.e., their performance exceeds the Task 2 performance of their teammates).

2. Receiving 2€ with probability $p^r \in 0, 0.01, 0.02, \ldots, 0.99, 1$.

It is incentive-compatible for participants to report a reservation probability equal to their subjective belief. The mechanism itself was carefully explained following the wording used by Mobius et al. (2014), and understanding was tested using a check-up question. After finishing the instructions, participants first reported their reservation probability and only then did they choose between piece rate and tournament.

If Task 4 was selected for payment, participants received their earnings for the exercises depending on whether they had chosen the piece rate or tournament, in a similar fashion to Task 3. In addition, a random value p was drawn for each participant. If p was above the reservation probability, the respective participant was paid according to a lottery with probability p. Otherwise, participants were paid 2€ if their performance was high enough to win the tournament.

III.F.2. Obtaining Other Parameters

What remains is to obtain values for x_i, the objective probability p_i^{o} and the gender-neutral belief p_i^{n}. For x_i, we used participants' actual performance in the forced tournament task (Task 2). Using a participant-specific value for x_i allowed us to ensure that the stakes involved in treatments NoComp, JustRisk and IntEff were similar to Task 3. We used Task 2 as a proxy for ability because performance in Task 2 could not be affected by the choice of incentives. This is a typical approach in the literature. In any case, performance across Tasks 1–4 is highly correlated ($.75 < r < .83$ for each individual correlation).

We obtained the objective probability of winning p_i^{o} in the following way. After

the last session of the experiment, we computed the empirical probability that the participant's performance would beat three competitors chosen randomly from all participants in the experiment (across all sessions). For example, for a participant who solved 12 exercises, we would check the probability that he or she would randomly be matched to only participants with a performance of 11 or less, or would win a random tiebreaker with another person (or people) with a score of 12. This represents a participant's objective probability of winning the tournament given his or her performance x_i and the performance of all other participants in the experiment.

To generate the gender-neutral belief p_i^{n}, we ranked the subjective beliefs p_i^{s} elicited over all sessions from largest to smallest, separately for each gender. For each man in the sample, we then replaced his elicited belief with the belief of the woman with the corresponding rank. For example, we replaced the beliefs of the 27th most confident man with the beliefs of the 27th most confident woman. This has the effect of imposing 'female beliefs' on men, allowing us to eliminate gender differences in overconfidence without eliminating overconfidence per se.[5]

III.F.3. Eliciting Choices

Obtaining participants' choices in treatments NoComp, JustRisk and IntEff required us to elicit three binary choices between a safe outcome and a lottery, that differed only in the probability of payment of the lottery. We elicited these choices using a separate Task, *Task 5 (Multiple Choice List)*. In Task 5, each participant made 20 choices between a fixed amount $0.5x_i$€ and a lottery with prize $2x_i$€, as per Table II. The probability of the lottery varied from 1 for the first row to .05 for the

5. By the logic of the interaction effect, the size of the estimated gender gap in treatment IntEff could depend on whether the gender-neutral beliefs p_i^{n} are generated by imposing female beliefs on men or by imposing male beliefs on women. However, in Appendix A1.2 we show that the two approaches yield very similar results, as does taking some average of the two beliefs.

TABLE II: Task 5 Choice Menu

	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 in the experiment was equal to performance in Task 2 (the forced tournament). In practice, the average value of the left option ranged from 2€ to 12.50€, with an average of 5.35€.

20th row. For the analysis, we use only the three choices made in the three rows for which the probability corresponded to treatment NoComp (p_i^{s}), JustRisk (p_i^{o}), and IntEff (p_i^{n}) respectively. These choices are what we refer to as treatment NoComp, JustRisk, and IntEff in this paper.

Presenting treatments NoComp, JustRisk and IntEff to participants in this way has several important advantages relative to separately presenting three binary choices. First, it means that the belief elicited in Task 4 did not directly affect the attractiveness of the lottery in Task 5. Second, it allowed us to compute p_i^{o} and p_i^{n} using the actual ex post distribution of performances over all sessions, rather than having to compute them during each session. Third, it gives us access to the full range of potential probabilities, which will allow us to check the robustness of our results to perturbations of p_i^{s}, p_i^{o} and p_i^{n}. Fourth, using a choice list greatly reduces the similarity between the baseline and treatments NoComp, JustRisk and IntEff, reducing the influence of order effects caused by preferences for consistency (Falk and Zimmermann,

2011; Cialdini, 1984; Cialdini, Trost, and Newsom, 1995) and similar phenomena.[6]

Finally, it is important to emphasize that the payoffs faced in Task 5 are individual-specific. For example, a participant who correctly solved 14 exercises in Task 2, received 20 choices between 7€ and 28€ with varying probability. A participant who solved 9 exercises was instead choosing between 4.50€ and a probability of getting 18€. This allowed the stake size to be similar to Task 3 (the baseline).

III.G. Other Parts

Between Task 4 and Task 5, participants went through one additional Task: *Task 4b (Universal Feedback).* Task 4b was identical to Task 4, with one exception. In Task 4, participants only received feedback on their relative performance if they chose the tournament (i.e., they found out whether they won). In Task 4b, we also told participants who chose the piece rate whether they would have won the tournament. This eliminates the effect of an alternative explanation of the gender difference in tournament choices. This explanation suggests that women may avoid tournaments in order to avoid receiving a signal about their relative ability (Niederle and Vesterlund, 2007). However, since gender differences in feedback aversion are not discussed in subsequent papers and we find no evidence for it in Task 4b, we postpone its main discussion to Appendix A1.

After the end of Task 5, one of the participants in the session was asked to roll a die to determine the task selected for payment. Participants then received feedback on their selected task, but not the other tasks. Feedback included absolute performance, total earnings and – when applicable – the outcome of the tournament

6. Note that Task 5 varies the probability in increments of .05. In cases where a probability is not a multiple of .05 (say .44), we therefore take the average of the closest rows (.4 and .45). For 98.3% of the choices in treatments NoComp, JustRisk and IntEff, the choices on the two closest rows are identical. In the remaining cases, we classify participants as indifferent between the two options.

and belief elicitation task. After receiving feedback, paticipants 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.[7]

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€. Of the participants, 98.6% 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.

IV. Results

Turning to the results, we first examine whether we replicate the gender difference in tournament choices in the baseline. We then compare the baseline to the control treatments. The results are summarized in Figures IV and V, which plot the raw data and the implied importance of the respective mechanisms. Additional figures summarizing gender differences in performance, beliefs and risk attitudes are presented in Appendix A1.5.

7. For the SOEP measure, participants were asked, on a scale from 1 to 10, whether they were persons who were willing to take risks, both in general and in five specific areas. We used the exact wording found in the German Socio-Economic Panel (SOEP). For the Holt-Laury task, participants made 10 choices between two lotteries. Lottery A resulted in either 1€ or 80 cents. Lottery B resulted in either 1.90€ or 10 cents. The probability of obtaining the higher payoff was identical for A and B, and varied from 10% to 100% across choices. One of the 10 choices was randomly chosen and paid out at the end of the experiment. For the Eckel-Grossman measure, participants chose between 6 options. Option 1 was a fixed payoff of 1.40€. Option 2–5 were progressively more risky, but with a higher expected value. Option 6 was riskier than option 5, but had the same expected value.

IV.A. Baseline

There were no gender differences in performance in the piece rate (men: 9.03 vs. women: 8.80, p=.723, t-test) and the tournament (men: 10.90 vs. women: 10.51, p=.569, t-test). Based on Task 2 performance, 24 men and 26 women would have maximized their expected payoffs by competing. Nevertheless, men (59%) were far more likely to choose the tournament than women (27%). The gender gap is comparable in size to Niederle and Vesterlund (2007), and significant (p<.001, Fisher's exact test).

Before moving to the control treatments, it is important to note that adding the belief elicitation task did not affect the gender gap in competitive choices. In Task 4, women (34%) were still significantly less likely to choose the tournament than men (67%, p<.001, Fisher's exact test). Since Task 4 is more closely connected to the beliefs used to generate treatments NoComp and IntEff, we will therefore use Task 4 as the baseline in this section. In Appendix A1, we show that all the results are robust to using Task 3 instead.

IV.B. Competitiveness

We identify the importance of competitiveness by comparing the baseline to treatment NoComp. If competitiveness is important, the gender gap should be smaller in treatment NoComp. However, this is not what we find. Instead, men (68.6%) were still significantly more likely than women (31.4%) to choose the risky option (p<.001, Fisher's exact test). The size of the gender gap (37.1 percentage points) is not significantly smaller than in the baseline (32.9pp; p=.672, one-sided t-test). If anything, it is slightly larger.[8]

8. These results are robust to removing the eight participants (5.7%) who made dominated choices or switched multiple times in Task 5. After removing these participants, the gender gap is 34.7pp.

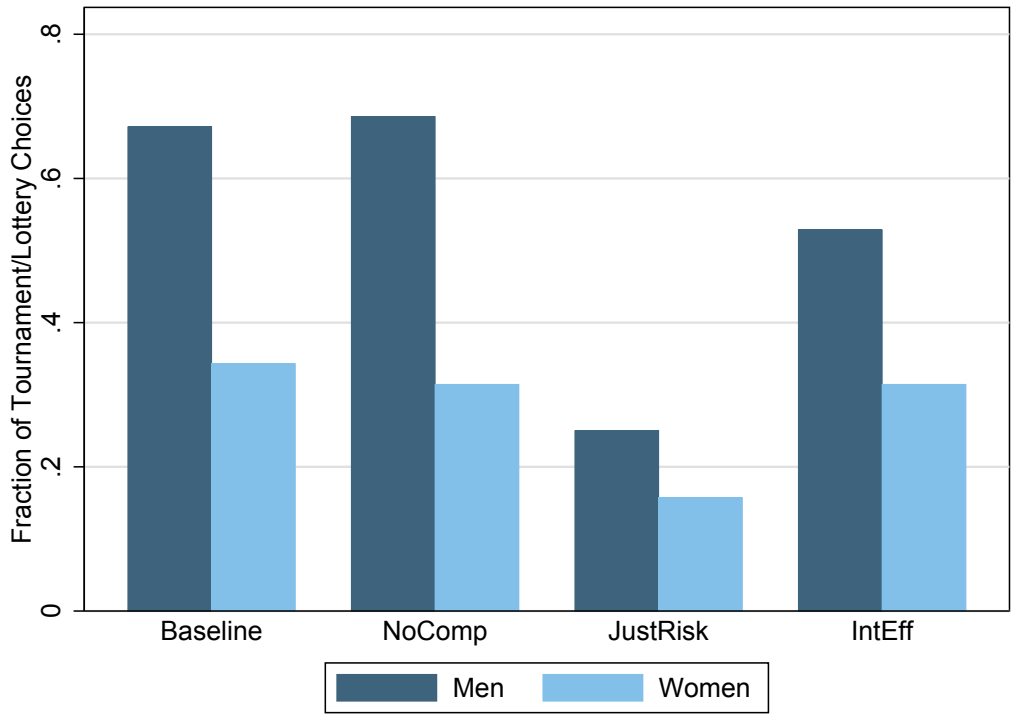

FIGURE IV: SUMMARY OF CHOICES BY GENDER.

Notes. This figure gives the fraction of participants choosing the tournament (baseline) or lottery (remaining bars) by gender.

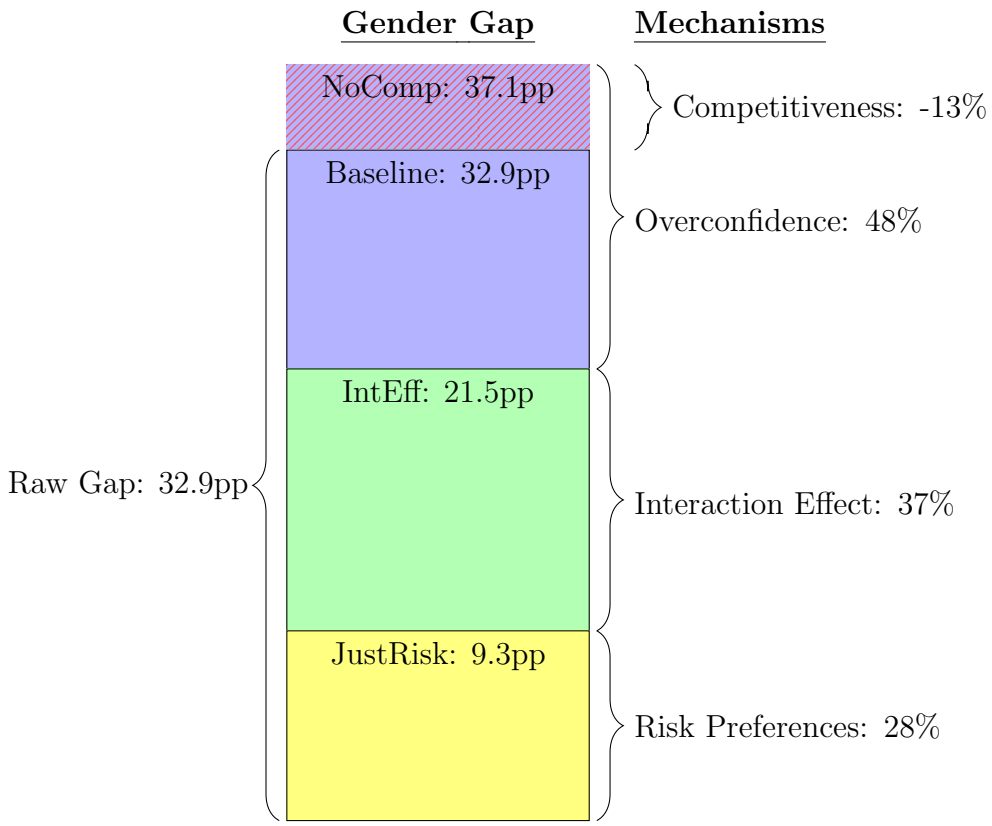

FIGURE V: THE GENDER GAP IN TOURNAMENT CHOICES IS FULLY EXPLAINED BY OVERCONFIDENCE, RISK PREFERENCES, AND THEIR INTERACTION.

Notes. This figure displays the gender gap in each treatment, and the importance of overconfidence, risk preferences, their interaction effect, and competitiveness in explaining the raw gender gap in tournament choices (baseline). The percentages are calculated by taking the difference between the gender gap in the respective treatments, and dividing the difference by the raw gender gap in tournament choices.

Given the existing literature, this is a very surprising result. Previous studies found that competitiveness, on average, explained 69.2% of the gender difference in tournament choices. By contrast, our results suggest that it explains -13.1% (i.e., $\frac{-4.3}{32.9}$). We clearly reject the hypothesis that competitiveness explains at least 69.2% of the gender difference in tournament choices (p<.001, one-sided Wald test). Indeed, our results suggest that the gender difference can be explained by risk preferences and beliefs, and competitiveness is unimportant.

IV.C. Risk Preferences and Overconfidence

Having established that competitiveness does not explain the gender difference in tournament choices, we next investigate the role of risk preferences and overconfidence. For these factors to explain the gender gap, it needs to be true that women are respectively more risk averse and less overconfident. We find both to be the case: women are indeed more risk averse (Holt-Laury: p=.009; Eckel-Grossman: p<.001; SOEP: p<.001, t-tests) and less overconfident (p=.007, t-test).

We identify the effect of overconfidence by comparing treatment NoComp to treatment IntEff. If overconfidence is important, the gender gap should be smaller in treatment IntEff. This is indeed what we find: 52.9% of the men and 31.4% of the women chose the lottery (p=.015, Fisher's exact test). The resulting gender difference (21.5pp) is significantly smaller than in treatment NoComp (37.1pp; p=.001, one-sided t-test). As illustrated by Figure V, these results imply that overconfidence explains 48% ($\frac{37.1-21.5}{32.9}$) of the gender difference in tournament choices.

To identify the importance of risk attitudes, we examine the gender gap in treatment JustRisk. Even though this treatment eliminates the effect of both overconfidence and competitiveness, we still find that men (25.0%) were more likely to choose

the lottery than women (15.7%). However, this difference is no longer significant (p=.197, Fisher's exact test), which implies that risk attitudes – by themselves – do not generate a significant gender gap for our sample size. At the same time, the observed gender difference does imply that risk attitudes can explain 28% ($\frac{9.3}{32.9}$) of the gender gap in tournament choices.

Similar to the literature, our results therefore suggest a greater role for overconfidence than for risk attitudes. Importantly, however, both the effect of overconfidence and the effect of risk attitudes in our study are more than twice as large as the effect implied by previous studies (see Figure III).

IV.D. The Interaction Effect

Finally, we investigate the interaction effect between risk attitudes and overconfidence. For this effect to be important, it needs to be true that men and women are both overconfident. This is indeed the case. Men on average, think they have a 58.9% chance of winning the tournament, versus 48.8% for women. Both numbers are significantly greater than the true probability of around 25% (p<.001, t-test, for each gender individually or combined).[9]

Given that participants are highly overconfident, the interaction effect implies that the gender gap should be larger in treatment IntEff than in treatment JustRisk. To see this, note that the lottery faced by the median participant in treatment IntEff had a 48.5% chance of payment – as implied by the median female subjective probability of winning. By contrast, the median probability in treatment JustRisk was 9.6% – as implied by the median objective probability. As a result, risk preferences had less

9. Finding that both genders are significantly overconfident is not uncommon. For example, Niederle and Vesterlund (2007) found that 75% of men and 43% of women thought they ranked first in a four-person tournament.

scope to explain a gender gap in the JustRisk treatment.

In line with this reasoning, we indeed find a smaller gender gap in the JustRisk treatment (9.3pp vs. 21.5pp), see Figure V. The treatment difference is marginally significant (p=.107, one-sided t-test) for our sample size. Hence, our results imply that the interaction effect explains 37% (i.e., $\frac{21.5-9.3}{32.9}$) of the gender difference in tournament choices.

These results illustrate the importance of accounting for interaction effects. For example, our results imply that – conditional on beliefs – removing the gender difference in risk attitudes would eliminate 65% of the gender difference in tournament choices. Without investigating the interaction effect, we would have estimated it to be 28%. To our knowledge, we are the first to study the interaction effect between risk attitudes and overconfidence in this literature.

V. Discussion

The key result of the previous section is that competitiveness does not explain the gender difference in tournament choices. In fact, the importance of competitiveness (-13%) is 82 percentage points smaller than implied by the literature (69%). This raises the question of why our results differ so starkly from the existing literature. Could our data be merely a large outlier? Or is there something more systematic about our experiment and the way we analyze the data that causes our results to be so different? And are there any potential confounds that could have interfered with our identification strategy?

V.A. Identification Strategy

We start our discussion by taking a closer look at our identification strategy. The identifying assumption for our treatment comparisons is that one treatment eliminates the effect of the variable of interest, but does not change the effect of other relevant variables. In this section, we examine whether this identifying assumption is reasonable by separately investigating the implications of these assumptions for the three main variables and potential confounds. For a more formal discussion we refer the interested reader to Appendix A4.

V.A.1. Identifying Competitiveness

We identify the effect of competitiveness by comparing the gender gap across the baseline and treatment NoComp. For this comparison, the identifying assumption requires that treatment NoComp eliminates the influence of competitiveness, but does not change the effect of risk preferences, overconfidence, and any other variables that could explain the gender difference.

Is this assumption reasonable? The literature treats competitiveness as a preference for being in a competitive environment, such as a tournament. Lotteries are not typically considered to be competitive. Hence, it seems reasonable that competitiveness cannot explain participants' choices in treatment NoComp.

The assumption also requires the effect of risk preferences and beliefs to be constant across these two treatments. For risk preferences, this holds if individual risk attitudes are constant across treatments, as seems reasonable. It also holds under weaker conditions, for example if individual risk attitudes vary across treatments, but the distribution of risk attitudes in the population stays constant. More details are presented in Appendix A4.

For beliefs, the assumption requires the distribution of elicited beliefs – which are used to generate treatment NoComp – to be a sufficiently accurate representation of actual latent beliefs. Perfectly measured beliefs are not required. Intuitively, we are only interested in comparing gender differences. Although inaccurately measured beliefs may distort the choices of individuals in treatment NoComp, they need not affect the gender difference since, e.g., mistakes may cancel out on average. We explore potential distortions caused by inaccurate beliefs in Appendix A4 and A5, where we show that such distortions are unlikely and, if anything, would lead us to overestimate the importance of competitiveness.

V.A.2. Identifying Other Variables

The comparisons we use to identify the effect of the other variables assume that treatment IntEff eliminates gender differences in overconfidence and treatment JustRisk eliminates the effect of overconfidence per se. Both are true by construction. To identify risk preferences, we also need to assume that choices in treatment JustRisk can only be driven by risk attitudes. We see no reason for these assumptions to be violated.

V.A.3. Potential Confounds

One potential confound may arise when treatment comparisons vary more than just the variable of interest. For example, treatment NoComp is not only less competitive than the baseline. It is also less ambiguous, and removes some social preferences, rank effects, performance feedback and the real-effort task per se. This could affect our results if one of these variables both (1) explains tournament choices and (2) is correlated with gender. We are, however, unaware of consistent evidence of any

variable that meets both criteria.[10]

Moreover, to obscure a positive effect of competitiveness and explain its null effect in our data, these variables would have to contribute to an increased gender difference in treatment NoComp. This would require women to be more feedback-seeking, more ambiguity-seeking, less altruistic, and/or have a lower cost of effort in the real-effort task, all of which seem unlikely.

Gender differences in performance are another potential confound. Previous studies (e.g., Niederle, Segal, and Vesterlund, 2013) have occasionally found that men perform better at solving addition problems. Although there is no evidence of a systematic performance difference in our experiment, even small and unsystematic differences could impact our identification strategy. We therefore explore performance differences in greater detail in Appendix A1.4. There, we present versions of treatment JustRisk and treatment IntEff that control for performance differences by design. This increases the gender gap in both treatments by approximately 2.5 percentage points, which implies a slightly greater importance for risk attitudes and a slightly lower importance for overconfidence. At the same time, these differences are small and do not affect any of our main conclusions.

Another potential concern are order effects caused by preferences for consistency and related phenomena. Since we have a within-subject design (as is custom in the literature), choices made in earlier parts of the experiment could conceivably have affected later choices. While we cannot fully exclude these effects, we sought to minimize them by presenting treatments NoComp, JustRisk and IntEff using a multiple choice list. This greatly reduced the similarity between the baseline (a

10. There is little or no evidence that ability, social preferences, ambiguity attitudes or feedback aversion differ systematically by gender. In addition, previous studies investigating social preferences (e.g., Almås et al., 2016; Kamas and Preston, 2012) and ambiguity attitudes (e.g., Gneezy, Pietrasz, and Saccardo, 2016) find that these cannot explain tournament choices; an exception is Balafoutas, Kerschbamer, and Sutter (2012).

single binary choice between tournament and piece rate) and the other treatments (20 choices between lotteries and certain amounts of money).

Finally, to separate overconfidence from risk attitudes and competitiveness, it is also important that elicited beliefs are not confounded by risk attitudes or competitiveness. If, for example, the elicitation method were to systematically overestimate the beliefs of risk tolerant or competitive individuals, part of the effects of these variables would incorrectly be attributed to overconfidence.

For risk attitudes, we can almost certainly exclude this possibility, since neither the utility-maximizing belief nor the actual reported belief depends on risk attitudes ($-.26 < r < .22$ depending on gender and the measure used). For competitiveness, we lack the theory and direct independent measure of competitiveness required to make similar claims. However, the beliefs we elicit are similar to the literature: both genders are overconfident and men are more overconfident than women. This is not consistent with competitiveness confounding our belief estimates, unless a similar confound is also present in earlier studies.[11]

V.A.4. Additional Considerations

Finally, it is important to highlight that our identification strategy does not impose any requirements on the consistency of the responses of individual participants. The reason is that our treatment comparisons aim to explain gender differences, not individual behavior. Provided that the ways individuals change their choices across treatments is not correlated with gender, such changes do not affect the gender difference in choices. Hence, our treatment comparisons are unaffected.

11. Further, our results are robust to even a large competitiveness confound that would lead us to overestimate the gender difference in beliefs by 50%. Specifically, artificially reducing the gender gap in beliefs by 50% only lowers the gender gap in treatment NoComp to 32.9pp. The implied importance of competitiveness would then be 0% (instead of -13%).

However, this also implies that we cannot pinpoint the exact reason why any individual may have changed their choice across treatments. For example, we cannot explain exactly why 43 participants (31%) changed their choice (e.g., from piece rate to lottery) from the baseline to treatment NoComp. Some participants may have switched because they were ambiguity averse, some because they were feedback averse, and others may have switched by mistake. Still others may actually have switched because of a competitiveness trait, or for a combination of reasons. Our design does not allow us to identify the exact reason (or reasons) why each individual switched, nor is it our goal to do so. Instead, the fact that switches are not correlated with gender tells us that these factors do not explain the gender difference in tournament choices.

V.B. Regression Analysis

If our identifying assumption holds, it is natural to ask why our results differ so starkly from the existing literature. As we previously noted, a key difference between our study and previous work is that we use treatment comparisons rather than regression analysis to differentiate between mechanisms. It seems conceivable that this may at least in part explain the lesser importance attached to competitiveness in our study.

To investigate whether this is the case, we replicate the regression analysis of previous studies. Comparing the results of this analysis to the results of our treatment comparisons allows us to establish whether the smaller importance attached to competitiveness in our study is driven by our treatment-based identification strategy, or by some idiosyncratic characteristics of our data.

Table III presents the results. Column (1) presents the gender gap after control-

TABLE III: Regression-Based Approach in our Data

	Coefficient (p-value)		
	(1)	(2)	(3)
Dependent Variable: Tournament Choice (Task 4)			
Female	-0.303***	-0.223**	-0.160*
	(0.081)	(0.089)	(0.093)
Confidence		0.228	
		(0.230)	
Eckel-Grossman		0.036	
		(0.025)	
SOEP		0.018	
		(0.017)	
Treatment NoComp Choice			0.190
			(0.134)
All Risk Measures			F=.84
			p=.500
All Conf. Measures			F=.01
			p=.987
Constant	0.065	-0.199	-0.328
	(0.214)	(0.224)	(0.257)
Ability Controls	yes	yes	yes
Observations	140	140	140

Notes. OLS Estimates, robust standard errors in parentheses. Dependent variable is the Task-4 choice of compensation scheme (1-tournament, 0-piece rate). Ability controls include performance in Task 2, the difference between performance in Task 2 and Task 1, and the objective probability of winning p^o, conditional on Task 2 performance. Confidence is the elicited probability of winning from Task 4. Eckel-Grossman and SOEP are the Eckel and Grossman (2002) and SOEP measures of risk preferences, respectively. Treatment NoComp choice is the choice made in treatment NoComp. In column 3, the confidence measures include the elicited beliefs from Task 4 and Task 4b. The risk measures include the Eckel-Grossman, Holt-Laury, and SOEP measures plus the number of risky choices taken in Task 5.
*** p<0.01, ** p<0.05, * p<0.1

ling for gender differences in ability (the raw gender gap is 32.9 percentage points). Following Buser, Niederle, and Oosterbeek (2014), column (2) adds the belief elicited in Task 4 and two measures of risk preferences as controls. Even after controlling for risk preferences and overconfidence, a 22.3 percentage point gender gap remains. Following the standard approach in the literature, these results would imply that 73.6% ($\frac{.223}{.303}$) of the gender gap in tournament choices can be attributed to competitiveness. Even when we add several additional measures of risk attitudes and confidence to our regressions in column 3, the residual gender gap is still 16 percentage points, which would imply that competitiveness explains 52.8% of the gender gap in tournament choices.[12]

The 53–74% attributed to competitiveness in Table III is considerably more than the -13% we obtained using treatment comparisons, and is similar to the 69% obtained by the literature using similar regressions. In other words, when we analyze our data the conventional way, our results closely replicate the standard result in the literature. This implies that the smaller importance of competitiveness in our data is driven by our treatment-based identification strategy, not by some idiosyncratic characteristics of our data.

V.C. Regression Analysis versus Treatments

Why are the results in Table III so different from the ones obtained using treatment comparisons? Table III uses regressions to control for risk attitudes and overconfidence, and attributes the residual gender coefficient to competitiveness. Among other things, this implicitly assumes that tournament choices are a linear function of com-

12. Following Niederle and Vesterlund (2007), we use the choice taken in treatment NoComp as a proxy for risk attitudes and overconfidence. Other new variables in this column are the Holt-Laury measure of risk preferences, the total number of risky choices in Task 5 and the beliefs elicited in Task 4b.

petitiveness, risk attitudes and beliefs – and no other variables. The experimental results already show that the linearity assumption is unrealistic, since risk attitudes and beliefs interact in a non-linear way, as predicted by expected utility theory.

Another problem arises when overconfidence and/or risk attitudes are measured with error.[13] Specifically, it is well known (see e.g., Hausman, 2001) that in such cases the coefficients for risk attitudes and overconfidence are downward biased and inconsistently estimated. Importantly, Gillen, Snowberg, and Yariv (2015) show that this in turn implies that the effect of competitiveness is overestimated. In Appendix A6 and A7 we present evidence that our control variables are subject to substantial measurement error, in part by showing that the estimated importance of competitiveness in Table III is substantially reduced after adjusting for measurement error econometrically.[14]

These results have important implications. Using an unbiased method (treatment comparisons) to identify competitiveness, it appears to explain -13% of the gender gap in tournament choices. Using regressions, we instead attribute 53-74% to competitiveness. The difference between these numbers suggests that the regression results may have overestimated the importance of competitiveness by as much as 87 percentage points. Since other studies have relied exclusively on regressions, this therefore implies that these studies may also have strongly overestimated the importance of competitiveness in explaining gender differences in tournament choices.

13. Measurement error in these variables is intuitive. Participants make mistakes, may not perfectly know their own preferences, preferences may vary across contexts or over time, etc. Formally, Gillen, Snowberg, and Yariv (2015); Holt and Laury (2014); Beauchamp, Cesarini, and Johannesson (2015) and Kimball, Sahm, and Shapiro (2008) show that measurement error is indeed a substantial part of standard measures of risk preferences, typically representing more than half the total variance in the elicited variable. We present similar results in Appendix A7.

14. Appendix A8 uses simulations to present an intuitive illustration of the effects of measurement error on treatment comparisons and regressions such as Table III.

V.D. Literature Comparison

Taken together, our results also generate a prediction: regressions such as Table III will overestimate the importance of competitiveness, relative to direct treatment comparisons. In this section, we test the prediction using the data of Niederle and Vesterlund (2007) and six other studies.

These studies include a control treatment where participants are asked to choose whether they want to submit their past performance (from Task 1) to piece rate or tournament incentives. The argument is that this choice is similar to the baseline, except that participants no longer have to face the stress or thrill of actually solving a task in a competitive environment. Thus, the control treatment is argued to eliminate the competitive element, but retain the effect of overconfidence and risk preferences.

Each of these studies then use the control treatment as a proxy for risk preferences and overconfidence in a regression. The idea is that the control treatment is more similar to the baseline than standard measures of risk preferences and beliefs, and may therefore serve as a better proxy for these variables. Otherwise, these studies follow the approach of Table III and assume that after controlling for this variable, any remaining gender difference is due to competitiveness.

However, the data from the control treatment also allow for a direct treatment comparison. If competitiveness is important, the gender gap should be significantly smaller in the control treatment. The difference between treatments can then serve as an estimate of the importance of competitiveness, similar to the comparison between treatment NoComp and the baseline in our experiment. Interestingly, none of these studies make this comparison themselves.

Figure VI presents the results of the seven studies that allow for a direct comparison. For more details on each individual study, we refer the reader to Table VII and

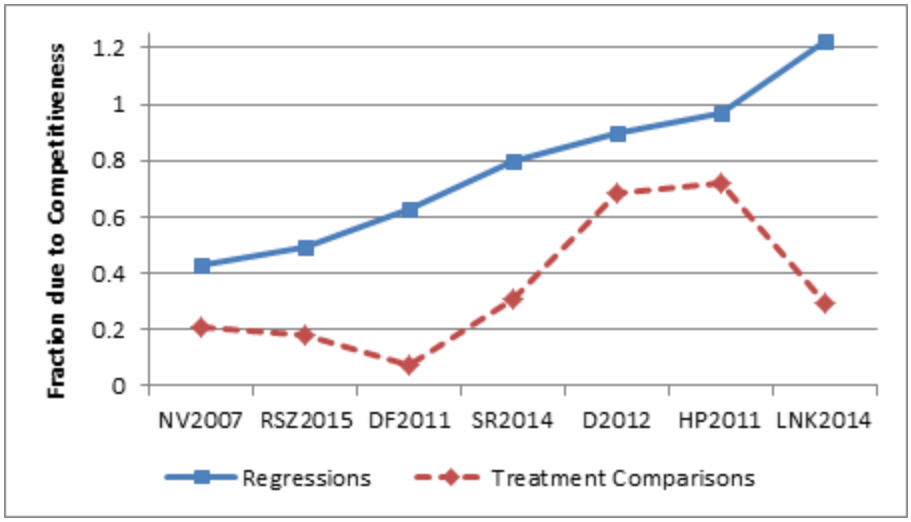

FIGURE VI: REGRESSIONS AND TREATMENT COMPARISONS IN THE LITERATURE

Notes. The figure plots the fraction of the total gender difference in tournament choices that is attributed to gender differences in competitiveness in seven different studies, using either regressions or direct treatment comparisons. For more details concerning the individual studies, see Table VII in Appendix A3

Appendix A3. The results are line with our prediction. Removing the competitive element decreases the gender gap by 35% on average. Using the approach taken in this paper, this would imply that around 35% of the gender gap is driven by competitiveness. The regressions presented in these studies instead imply that 78% of the gender gap is due to competitiveness. The difference between the two approaches is large (43 percentage points) and present in the predicted direction in all seven studies.[15]

The message is clear. In eight separate experiments, including our own, regressions attribute a considerably larger fraction of the total gender gap to competitiveness. Assuming that the treatment comparisons are unbiased, this implies that previous studies relying on regressions have substantially overestimated the importance of

15. It is worth noting that the fraction attributed to competitiveness in these studies (35%) is larger than in our study (-13%). One possible reason is that the control treatment also eliminates the effect of optimism about future performance. If men are more optimistic than women, the 35% may therefore reflect gender differences in optimism (i.e., overconfidence) as well as competitiveness.

competitiveness.[16]

VI. Conclusion

There is ample evidence from laboratory experiments that women are less likely to sort into competitive environments. While part of the gap may be explained by gender differences in overconfidence and risk preferences, the current view is that the majority can be explained by gender differences in a separate competitiveness trait. Indeed, the idea that women are less competitive than men has grown to be very influential.

We replicate the basic result, but provide new evidence that the gender gap in these studies reflects not competitiveness, but is the result of gender differences in risk preferences, and overconfidence. Our ability to distinguish between competitiveness and these alternative mechanisms comes from a powerful, novel, experimental design. Our design allows us to directly and cleanly identify the importance of competitiveness, while avoiding the criticisms raised against the techniques used in earlier work.

Our message is clear. Gender differences in tournament choices in our experiment are fully explained by risk preferences, overconfidence and their interaction effect. This has important implications for policy and future research. Rather than competitiveness, our results suggests that attempts at understanding and fighting gender differences in labor market outcomes would be better served by targeting overconfidence, risk attitudes, and other factors not captured by these experiments. Our

16. In Appendix A3, we discuss additional studies that can be used to identify the effect of overconfidence or risk preferences using treatment comparisons. The results are in line with our experiment, and suggest a greater role for overconfidence and risk preferences than implied by studies relying on regression-based identification.

results are reassuring, in the sense that rather than requiring us to further study a novel, not yet well-understood concept (competitiveness), they imply that policies and future research can address the gender gap using the well-understood and much-tested theory on decision under risk and uncertainty.

Our results also illustrate the importance of accounting for the interaction of risk attitudes and overconfidence. Policies reducing gender differences in overconfidence may be less effective or even counterproductive if, by changing the level of overconfidence, they simultaneously increase the importance of risk attitudes. In a similar vein, policies that change the appeal of competitive environments may change the gender gap by lowering the impact of risk preferences. An interesting recent example is Petrie and Segal (2015), who show that gender differences are largely eliminated in situations where the tournament is very attractive.

On a broader level, our results also contribute to the recent discussion on the potential pitfalls involved in controlling for risk preferences, beliefs, and other elicited variables using regressions (most notably, Gillen, Snowberg, and Yariv, 2015). We are able to obtain a direct measure of the bias involved, and show that it is large: the importance of competitiveness is overestimated by upwards of 50 percentage points. Importantly, the method we use circumvents these concerns by controlling for relevant variables by experimental design. Indeed, we hope that our results illustrate the usefulness of controlling for important confounds experimentally, and that they will encourage others to adopt a similar approach in the future.

References

Almås, Ingvild, Alexander W. Cappelen, Kjell G. Salvanes, Erik Ø. Sørensen, and Bertil Tungodden. 2016. "Willingness to Compete: Family Matters." *Management*

Science 62 (8):2149–2162.

Ambuehl, Sandro and Shengwu Li. 2016. "Belief Updating and the Demand for Information." *Working Paper* .

Andersen, Steffen, Glenn W. Harrison, Morten I. Lau, and E. Elisabet Rutström. 2008. "Lost in State Space: are Preferences Stable?" *International Economic Review* 49 (3):1091–1112.

Balafoutas, Loukas, Rudolf Kerschbamer, and Matthias Sutter. 2012. "Distributional preferences and competitive behavior." *Journal of Economic Behavior & Organization* 83 (1):125–135.

Baron, Reuben M. and David A. Kenny. 1986. "The moderatormediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations." *Journal of Personality and Social Psychology* 51 (6):1173–1182.

Beauchamp, Jonathan, David Cesarini, and Magnus Johannesson. 2015. "The Psychometric and Empirical Properties of Measures of Risk Preferences." *SSRN Working Paper* .

Berge, Lars Ivar Oppedal, Kjetil Bjorvatn, Armando Jose Garcia Pires, and Bertil Tungodden. 2015. "Competitive in the lab, successful in the field?" *Journal of Economic Behavior and Organization* 118:303–317.

Bertrand, Marianne. 2011. "New Perspectives on Gender." In *Handbook of Labor Economics Volume 4B*, edited by Orley Ashenfelter and David Card. 1543–1590.

Bullock, John G., Donald P. Green, and Shang E. Ha. 2010. "Yes, but whats the mechanism? (dont expect an easy answer)." *Journal of Personality and Social Psychology* 98 (4):550–558.

Buser, T., M. Niederle, and H. Oosterbeek. 2014. "Gender, Competitiveness, and Career Choices." *The Quarterly Journal of Economics* 129 (3):1409–1447.

Buser, Thomas, Anna Dreber, and Johanna Mollerstrom. 2016. "The impact of stress on tournament entry." *Experimental Economics* (forthcoming).

Buser, Thomas, Lydia Geijtenbeek, and Erik Plug. 2015. "Do Gays Shy Away from Competition? Do Lesbians Compete Too Much?" *IZA Discussion Paper* (9382).

Cadsby, C. Bram, Maroš Servátka, and Fei Song. 2013. "How competitive are female professionals? A tale of identity conflict." *Journal of Economic Behavior & Organization* 92:284–303.

Cialdini, Robert. 1984. *Influence, the Psychology of Persuasion*. New York: Harper Collins.

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 (2):318–328.

Croson, Rachel and Uri Gneezy. 2009. "Gender Differences in Preferences." *Journal of Economic Literature* 47 (2):448–474.

Dargnies, Marie-Pierre. 2012. "Men Too Sometimes Shy Away from Competition: The Case of Team Competition." *Management Science* 58 (11):1982–2000.

Dohmen, Thomas and Armin Falk. 2011. "Performance Pay and Multidimensional Sorting: Productivity, Preferences, and Gender." *American Economic Review* 101 (2):556–590.

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 (3):522–550.

Dreber, Anna, Emma von Essen, and Eva Ranehill. 2014. "Gender and competition in adolescence: task matters." *Experimental Economics* 17 (1):154–172.

Eckel, Catherine C. and Philip J. Grossman. 2002. "Sex differences and statistical stereotyping in attitudes toward financial risk." *Evolution and Human Behavior* 23 (4):281–295.

Ertac, Seda and Balazs Szentes. 2011. "The Effect of Information on Gender Differences in Competitiveness: Experimental Evidence." *Working Paper* .

Falk, Armin and Florian Zimmermann. 2011. "Preferences for Consistency." *IZA Discussion Paper* (5840).

Filippin, Antonio and Paolo Crosetto. 2016. "A Reconsideration of Gender Differences in Risk Attitudes." *Management Science* (forthcoming).

Flory, J. A., A. Leibbrandt, and J. 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 (1):122–155.

Flory, Jeffrey A., Kenneth L. Leonard, Uri Gneezy, and John A. List. 2016. "Gender, Age, and Competition: the Disappearing Gap." *Working Paper* :1–49.

Gerber, Alan S. and Donald P. Green. 2012. *Field Experiments*. London: W. W. Norton & Company.

Gillen, Ben, Erik Snowberg, and Leeat Yariv. 2015. "Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study." *NBER Working Paper* (21517):1–43.

Gneezy, Uri, Aniela Pietrasz, and Silvia Saccardo. 2016. "On the Size of the Gender Difference in Competitiveness." *Management Science* (forthcoming).

Gneezy, Uri and Jan Potters. 1997. "An Experiment on Risk Taking and Evaluation Periods." *The Quarterly Journal of Economics* 112 (2):631–645.

Goldin, Claudia. 2014. "A Grand Gender Convergence: Its Last Chapter." *American Economic Review* 104 (4):1091–1119.

Green, D. P., S. E. Ha, and J. G. Bullock. 2010. "Enough Already about "Black Box" Experiments: Studying Mediation Is More Difficult than Most Scholars Suppose." *The ANNALS of the American Academy of Political and Social Science* 628 (1):200–208.

Greiner, Ben. 2015. "Subject pool recruitment procedures: organizing experiments with ORSEE." *Journal of the Economic Science Association* 1 (1):114–125.

Grosse, Niels D., Gerhard Riener, and Markus Dertwinkel-Kalt. 2014. "Explaining Gender Differences in Competitiveness: Testing a Theory on Gender-Task Stereotypes." *Mimeo* :1–35.

Hausman, Jerry. 2001. "Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left." *Journal of Economic Perspectives* 15 (4):57–67.

Healy, Andrew and Jennifer Pate. 2011. "Can Teams Help to Close the Gender Competition Gap?" *The Economic Journal* 121 (555):1192–1204.

Holt, Charles A. and Susan K. Laury. 2002. "Risk aversion and incentive effects." *American Economic Review* 92 (5):1644–1655.

———. 2014. "Assessment and Estimation of Risk Preferences." In *Handbook of Economics of Risk and Uncertainty*, edited by Mark J. Machina and W. Kip Viscusi. Amsterdam: Elsevier North Holland, 135–202.

Kamas, Linda and Anne Preston. 2012. "The importance of being confident; gender, career choice, and willingness to compete." *Journal of Economic Behavior & Organization* 83 (1):82–97.

Kimball, Miles S, Claudia R Sahm, and Matthew D Shapiro. 2008. "Imputing Risk Tolerance From Survey Responses." *Journal of the American Statistical Association* 103 (483):1028–1038.

Kline, Rex B. 2005. *Principles and Practice of Structural Equation Modeling*. New York/London: The Guilford Press.

Krashinsky, Harry A. 2004. "Do Marital Status and Computer Usage Really Change the Wage Structure?" *The Journal of Human Resources* 39 (3):774.

Lee, Soohyung, Muriel Niederle, and Namwook Kang. 2014. "Do single-sex schools make girls more competitive?" *Economics Letters* 124 (3):474–477.

Masclet, David, Emmanuel Peterle, and Sophie Larribeau. 2015. "Gender differences in tournament and flat-wage schemes: An experimental study." *Journal of Economic Psychology* 47:103–115.

Mobius, Markus, Muriel Niederle, Paul Niehaus, and Tanya Rosenblat. 2014. "Managing Self-Confidence: Theory and Experimental Evidence." *Working Paper* .

Niederle, Muriel. 2016. "Gender." In *Handbook of Experimental Economics (Forthcoming)*, edited by John Kagel and Alvin E. Roth. Second edition ed.

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):1–16.

Niederle, Muriel and Lise Vesterlund. 2007. "Do Women Shy Away From Competition? Do Men Compete Too Much?" *Quarterly Journal of Economics* 122 (3):1067–1101.

———. 2011. "Gender and Competition." *Annual Review of Economics* 3 (1):601–630.

Petrie, Ragan and Carmit Segal. 2015. "Gender Differences in Competitiveness: The Role of Prizes." *Working Paper* :1–32.

Reuben, Ernesto, Paola Sapienza, and Luigi Zingales. 2015. "Taste for competition and the gender gap among young business professionals." *Mimeo* .

Reuben, Ernesto, Matthew Wiswall, and Basit Zafar. 2015. "Preferences and Biases in Educational Choices and Labour Market Expectations: Shrinking the Black Box of Gender." *The Economic Journal* (forthcoming).

Schlag, Karl, James Tremewan, and Joël van der Weele. 2015. "A penny for your thoughts: a survey of methods for eliciting beliefs." *Experimental Economics* 18 (3):457–490.

Shurchkov, Olga. 2012. "Under Pressure: Gender Differences in Output Quality and Quantity under Competition and Time Constraints." *Journal of the European Economic Association* 10 (5):1189–1213.

Sutter, Matthias and Daniela Glätzle-Rützler. 2015. "Gender Differences in the Willingness to Compete Emerge Early in Life and Persist." *Management Science* 61 (10):2339–23354.

Westfall, Jacob and Tal Yarkoni. 2016. "Statistically Controlling for Confounding Constructs Is Harder than You Think." *PLOS ONE* 11 (3):e0152719.

Wozniak, David, William T. Harbaugh, and Ulrich Mayr. 2014. "The Menstrual Cycle and Performance Feedback Alter Gender Differences in Competitive Choices." *Journal of Labor Economics* 32 (1):161–198.

Zhang, Jane. 2013. "Culture, Institutions, and the Gender Gap in Competitive Inclination: Evidence from The Communist Experiment in China." *Mimeo* .

Appendix A – For Online Publication

A1. Additional Results

In this section, we present five sets of additional results. First, we discuss the outcome of Task 4b and the implications for feedback aversion. Second, we examine the robustness of treatment IntEff to imposing male beliefs on women, rather than vice versa. Third, we discuss the robustness of our results with respect to using Task 3, rather than Task 4, as our baseline. Fourth, we examine the robustness of JustRisk to directly controlling for gender differences in performance. Finally, we present additional figures summarizing gender differences in performance, probability of winning, and risk preferences.

A1.1 Feedback Aversion

In addition to being less competitive, more risk averse and less confident, women may also be more averse to receiving relative performance feedback. Niederle and Vesterlund (2007) refer to this as feedback aversion, and argue that it could be a fourth reason why women may want to avoid the tournament. To investigate its importance, we included another Task (4b) in which participants had to choose between tournament and piece rate. In contrast with the baseline, however, even participants who chose the piece rate were now told whether they would have won the tournament. This treatment therefore eliminated the influence of feedback aversion, since choosing the piece rate was no longer a way for women to avoid relative performance feedback.

If feedback aversion is an important driver of choices in the tournament, the gender difference should be smaller in Task 4b. However, this is not what we find. Instead, we find that 62.9% of men and 28.6% of women choose the tournament in Task 4b. The gender difference is 34.3 percentage points, which is significantly different from

zero (p<.001, Fisher's exact test), and very similar to the gender difference in Task 4 (32.9 percentage points; p=.853, t-test). Hence, we find no evidence that feedback aversion contributes to the gender difference in tournament choices.

A1.2 Treatment IntEff with Male Beliefs

For treatment IntEff, we eliminated gender differences in overconfidence by replacing the beliefs of men with the beliefs of the women of corresponding confidence rank. However, we could just as easily have replaced the beliefs of the women with the beliefs of the men. This is important, because by the logic of the interaction effect, the effect of risk preferences may differ depending on whether choices are evaluated at the beliefs of men ($p_i^s = .59$ on average) or women ($p_i^s = .49$ on average).

When we replace the beliefs of women by the beliefs of men instead, 68.6% of men and 50.7% of women choose the lottery. The resulting gender difference is 18.1 percentage points (p=.039, Fisher's exact test). This is slightly smaller than the one reported in treatment IntEff using female beliefs, but the difference-in-difference is small (3.3 percentage points) and not significant.

These results do not affect any of our main conclusions: both overconfidence and the interaction term still explain a non-negligible part of the gender gap in tournament choices, and overconfidence is the most important factor. Imposing instead a mix of female and male beliefs leads to a gender difference somewhere between these two cases.

A1.3 Task 3 and Task 4 Tournament Choices

Since Task 4 is more closely connected to the beliefs used to generate treatments NoComp and IntEff, we used the tournament choice in Task 4 as our main dependent variable of interest. However, our results are robust to using Task 3 as the baseline

TABLE IV: Task 3 as Dependent Variable

	Coefficient (Std. Error)		
	(1)	(2)	(3)
Dependent Variable: Tournament Choice (Task 3)			
Female	-0.313***	-0.160**	-0.157*
	(0.078)	(0.081)	(0.087)
Confidence		0.419**	
		(0.227)	
Eckel-Grossman		0.039*	
		(0.023)	
SOEP		0.057***	
		(0.016)	
Treatment NoComp Choice			0.109
			(0.136)
All Risk Measures			F=3.01**
			p=.021
All Conf. Measures			F=.44
			p=.647
Constant	0.422**	0.290	-0.001
	(0.198)	(0.208)	(0.244)
Ability Controls	yes	yes	yes
Observations	140	140	140

Notes. OLS Estimates, robust standard errors in parentheses. Dependent variable is the Task-3 choice of compensation scheme (1-tournament, 0-piece rate). Ability controls include performance in Task 2, the difference between performance in Task 2 and Task 1, and the objective probability of winning p^o, conditional on Task 2 performance. Confidence is the elicited probability of winning from Task 4. Eckel-Grossman and SOEP are the Eckel and Grossman (2002) and SOEP measures of risk preferences respectively. Treatment NoComp choice is the choice made in treatment NoComp. In column 3, the confidence measures include the elicited beliefs from Task 4 and Task 4b. The risk measures include the Eckel-Grossman, Holt-Laury and SOEP measures plus the number of risky choices taken in Task 5.
*** p<0.01, ** p<0.05, * p<0.1

instead. We already saw that the raw gender gap for Task 3 (31.4%) and Task 4 (32.9%) are very similar. In Table IV we also replicate the regression analysis from Table III using Task 3 choices. The results are also very similar. After controlling for beliefs and risk attitudes, the residual gender coefficient in column 3 is .157, nearly identical to the .160 reported in Table III.

A1.4 Controlling for Performance

Our identification strategy requires there to be no gender difference in performance. This is important, because previous studies have occasionally found that men perform better at solving addition problems. Such a performance difference would affect our identification strategy in two ways. First, it would imply that men faced more attractive lotteries in treatment JustRisk, leading us to overestimate the gender gap in this treatment. Second, it would imply that treatment IntEff underestimates male overconfidence, leading us to underestimate the gender gap in this treatment.[17] Hence, we would overestimate the importance of risk preferences and overconfidence, and underestimate the interaction effect.

However, we showed in the results section that gender differences in performance in both Task 1 and Task 2 are small, and far from significant (see also Figure VII below). Indeed, the average objective probability of winning is identical across genders ($\bar{p}^{\text{o}}_W = \bar{p}^{\text{o}}_M = .25$; p=.979, t-test; see also Figure VIII below). Nevertheless, even small and unsystematic differences in the performance distribution could impact our identification strategy. In this section, we therefore control for gender differences in performance directly, by design.

For treatment JustRisk, this involves applying the approach of treatment IntEff

17. To see this, note that overconfidence is defined as $p^{\text{s}}_i - p^{\text{o}}_i$. Treatment IntEff eliminates gender differences in p^{s}_i by construction, which eliminates gender differences in overconfidence if and only if $p^{\text{o}}_M = p^{\text{o}}_W$.

to differences in performance. Specifically, we rank the objective probabilities p_i^{o} over all sessions from largest to smallest, separately for each gender. For each man in the sample, we then replace his p_i^{o} with the objective probability of the woman of corresponding rank. This allows us to compare male and female choices in a version of treatment JustRisk where gender differences in performance are eliminated by construction (treatment JustRiskNP).

For treatment IntEff, we control for differences in performance in the following way. First, we rank participants' overconfidence ($p_i^{\text{s}} - p_i^{\text{o}}$) over all sessions from largest to smallest, separately for each gender. For each man in the sample, we then replace his $p_i^{\text{s}} - p_i^{\text{o}}$ with the overconfidence of the woman of corresponding rank. We then use the 'female' overconfidence to compute an adjusted gender-neutral belief $p_i^{\text{n,adj}} = p_{i,W}^{\text{s}} + (p_i^{\text{o}} - p_{i,W}^{\text{o}})$. $p_i^{\text{n,adj}}$ allows us to compare male and female choices in a version of treatment IntEff that eliminates gender differences in overconfidence while controlling for performance differences (treatment IntEffNP).

In treatment JustRiskNP, men (27.1%) are still more likely to choose the lottery than women (15.7%, p=.095, Fischer's exact test). The corresponding percentages in treatment JustRisk were 25.0% and 15.7%. Controlling for performance therefore *increases* the gender gap by 2.1 percentage points. Similarly, in treatment IntEffNP, 55.7% of men and 31.4% of women choose the lottery (p=.005, Fischer's exact test). This is similar to treatment IntEff, where the corresponding percentages were 52.9% and 31.4% respectively. In this case, controlling for performance *increases* the gender gap by 2.7 percentage points.

Overall, controlling directly for performance increases the size of the gender gap by approximately 2.5 percentage points in both treatments. This implies that, if anything, the analysis reported in the main text slightly underestimates the importance of risk preferences, and slightly overestimates the direct effect of overconfidence.

However, these differences are small, and do not affect any of our main conclusions.

A1.5 Additional Figures

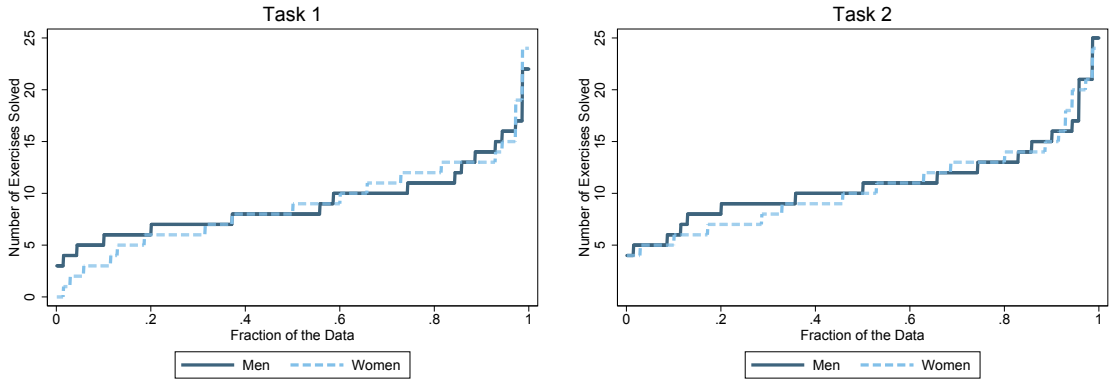

FIGURE VII: PERFORMANCE IN TASK 1 AND TASK 2

Notes. The figure presents cumulative density plots for the number of exercises solved in Task 1 (left) and Task 2 (right), by gender.

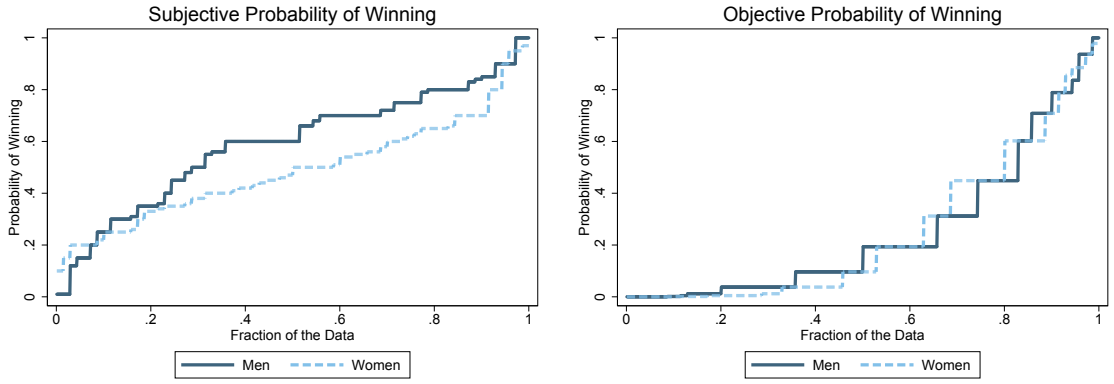

FIGURE VIII: PROBABILITY OF WINNING THE TOURNAMENT

Notes. The figure presents cumulative density plots for the subjective (left) and objective (right) probability of winning, by gender. The subjective probability of winning is the belief elicited in Task 4. The objective probability of winning is computed by comparing each participant's performance with the performance of every other participant in the experiment, more details are presented in section III.F.2.

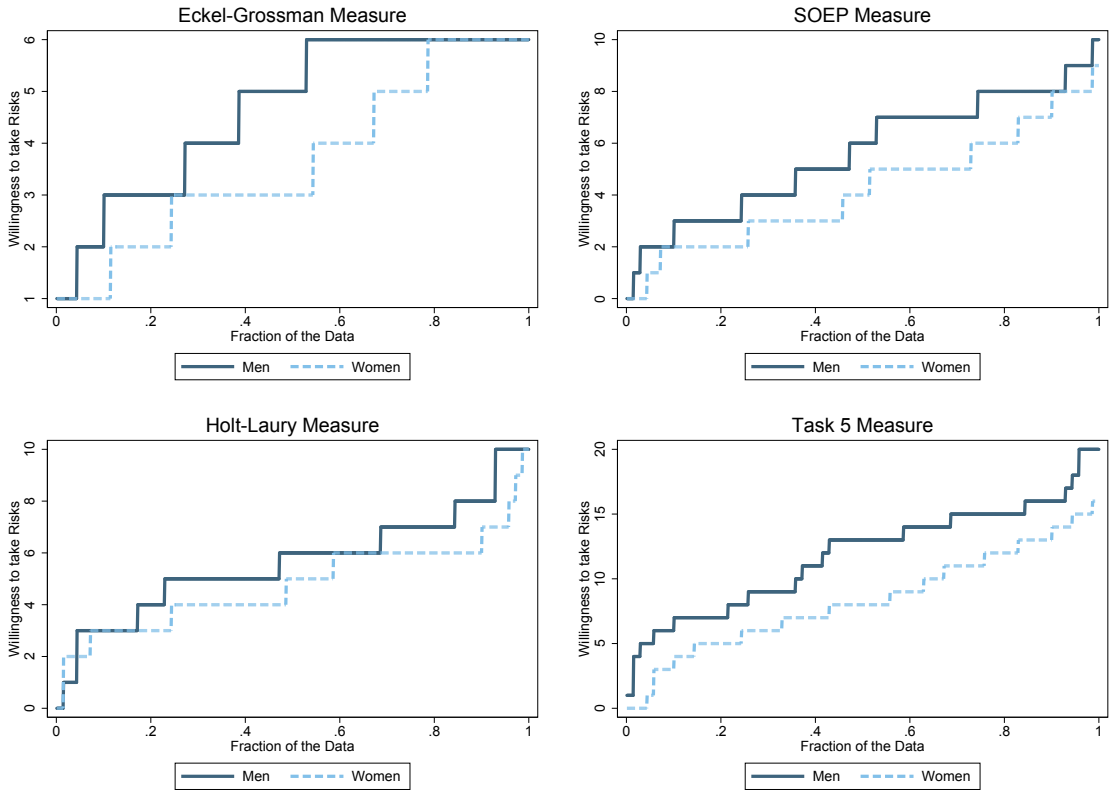

FIGURE IX: ELICITED RISK PREFERENCES

Notes. The figure presents cumulative density plots of the willingness to take risks for each of the four measures of risk preferences, by gender. For the Eckel-Grossman measure, the willingness to take risks is the riskiness of the chosen lottery. For the SOEP measure, it is the answer to the general risk taking question. For the other two measures, it is the number of times the risky lottery was selected over the safer alternative.

A2. Regression Approach in the Literature

In the literature section we discussed the results of a number of previous studies that used regressions to separate the effect of competitiveness from the effect of other factors. In this section, we briefly explain this approach using an example. We then discuss how we selected the studies featured in Figure I and present some additional details on their designs.

A2.1 Method

The studies we review in this section attempt to distinguish competitiveness from competing explanations in the following way. First, they obtain separate measures of risk attitudes, overconfidence, and any other factors of interest. They then use these measures as control variables in a regression. The idea is that after controlling for other relevant factors in a regression, any residual gender gap must then be driven by competitiveness. In the social sciences and statistics, this approach is typically referred to as mediation analysis (Baron and Kenny, 1986; Bullock, Green, and Ha, 2010).

Table V illustrates this approach using the results of a prominent recent study by Buser, Niederle, and Oosterbeek (2014). Column (1) shows that women are 23 percentage points less likely to choose the tournament, even after controlling for ability (the raw gender gap is 26 percentage points). Column (2) adds controls for overconfidence by including participants' guessed rank, elicited following the main experiment. This lowers the gender gap from 23.3 to 15.8 percentage points. Column (3) adds an incentivized lottery task (Eckel and Grossman, 2002) and a non-incentivized survey question (Dohmen et al., 2011) to control for risk attitudes. This further reduces the gender gap from 15.8 to 12.2 percentage points.

TABLE V: Regression Approach in Buser et al., (2014)

	Coefficient (Std. Error)		
	(1)	(2)	(3)
Female	-.233***	-.158***	-.122***
	(.047)	(.045)	(.044)
Tournament	.037**	.011	.011
	(.015)	(.014)	(.014)
Tournament−Piece Rate	-.027***	-.022	-.019
	(.011)	(.010)	(.010)
Win Prob	.263	.119	.072
	(.169)	(.157)	(.153)
Guessed Tournament Rank		-.205***	-.182***
		(.027)	(.027)
Lottery			.042*
			(.024)
Risk-Taking			.102***
			(.021)
Observations	362	362	362

Notes. This table reprints the results of Buser, Niederle, and Oosterbeek (2014), Table 7, columns 1, 2 and 4. It presents coefficients from OLS regressions, where the dependent variable is the compensation scheme (1, tournament, and 0, piece rate). Each regression also includes school fixed effects and test version fixed effects. Standard errors are in parentheses; *p<.10, **p<.05, ***p<.01.

They then compare the gender coefficients in column (1) and (2), and interpret the difference as evidence that "slightly over 30% of the gender gap can be explained by gender differences in overconfidence" (p. 1430). Similarly, comparing columns (2) and (3) suggests that gender differences in risk attitudes can only explain a small fraction of the gender gap. More importantly, the ratio of the gender coefficient in columns (1) and (3) suggests that competitiveness can explain 52.4% of the gender gap in tournament choices.

A2.2 Summary of Studies

Table VI summarizes the studies we incorporated in Figure I. Our main requirement was for these studies to report both (1) the raw gender difference in tournament choices and (2) the residual gender coefficient after controlling for both risk preferences and overconfidence in a regression. Figure I plots the ratio of these two numbers, which is what is typically attributed to a competitiveness trait.

The exact specification used differs considerably across studies. Different studies use different proxies for risk preferences and beliefs. Most studies also include at least one control for ability or past performance, and several studies also control for one or more additional factors. Whenever possible, we therefore use the regression that controls for risk preferences and overconfidence, but otherwise controls for as few variables as possible. The relevant control variables used in each study are listed in Table VI.

Several types of risk measures are used. Most common are a non-incentivized question (RQ) concerning a general tendency to take risks, and one of several versions of a multiple price list (MPL). These are similar to the SOEP question and Holt-Laury measure used in this study. One or two studies use variations of the Eckel and Grossman (2002) measure (EG) or the investment game (IG, Gneezy and Potters,

TABLE VI: Regression Approach in the Literature

Study	Gender Gap Raw	Gender Gap Controls	Attributed to Competitiveness	Controls	Comments
Gillen, Snowberg, and Yariv (2015)	19***	4.8	25.3	GR,TP,PD,IG,CRT1	Controls set 1
Zhang (2013)	14.6***	3.7	25.3	OC,EG,PW	Han sample
Dreber, von Essen, and Ranehill (2014)	19.1***	5.8	30.4	TP,GR,MPL	
Kamas and Preston (2012)	17.6***	5.7	32.4	MPL,RQ,GR,EP,PC	
Niederle and Vesterlund (2007)	37.9***	16.2**	42.7	TP,PD,GR,CT	
Balafoutas, Kerschbamer, and Sutter (2012)	26.1***	11.6	44.4	TP,MPL,GR,OC	Raw measure already controls for TP
Buser, Geijtenbeek, and Plug (2015)	8.6*	4	46.5	RQ,BO	Mathematical Matrix Task
Buser, Niederle, and Oosterbeek (2014)	26***	12.2***	46.9	TP,PD,PW,GR,EG,RQ	
Niederle, Segal, and Vesterlund (2013)	36***	17**	47.2	TP,PD,GR,CT	
Buser, Dreber, and Mollerstrom (2016)	23.9***	11.7	49.0	PP,TP,GR,GR1,RQ	Experiment 1
Reuben, Sapienza, and Zingales (2015)	26.8***	13.3**	49.7	TP,PD,GR,MPL	
Gillen, Snowberg, and Yariv (2015)	19***	11***	57.9	GR,TP,PD,MPL,CRT2	Controls set 2
Reuben, Wiswall, and Zafar (2015)	29***	18**	62.1	PW,SPW,MPL	
Dohmen and Falk (2011)	25***	15.7	62.8	EP,BO,RQ,SP	
Almås et al. (2016)	19.4***	13.9***	71.6	TP,BO,MPL,SP,PM	
Sutter and Glätzle-Rützler (2015)	21***	16.7***	79.5	A,TP,PD,M,GR,MPL	Raw measure already controls for age
Gneezy, Pietrasz, and Saccardo (2016)	36.4***	30.8***	84.6	EP,BO,MPL,RQ,AA	Ball task
Flory, Leonard, Gneezy, and List (2016)	15.6	14.0	89.7	PP,PD,GR,CT	American sample
Dargnies (2012)	33.3***	29.9***	89.8	CT	Own calculation
Buser, Dreber, and Mollerstrom (2016)	28.9***	27***	93.4	PP,TP,GR,GR1,RQ	Experiment 2
Healy and Pate (2011)	52.7***	51.2***	97.2	TP,PD,GR,CT	
Flory, Leonard, Gneezy, and List (2016)	7.9**	7.8**	98.7	PP,PD,GR,CT	Malawi sample
Zhang (2013)	22.9***	23.7***	103.5	OC,EG,PW	Yi sample
Zhang (2013)	27.5***	29.7***	108.0	OC,EG,PW	Mosuo sample
Masclet, Peterle, and Larribeau (2015)	37.6**	41.4**	110.1	TP,EC,MPL,SP,GR	Raw measure already controls for TP and EC
Lee, Niederle, and Kang (2014)	7.6***	9.3***	122.4	TP,PD,GR,CT	OLS results
Average			69.2		

Notes. The raw gender gap is the raw gender difference in tournament choices. The gender gap with controls is the gender coefficient in a regression of tournament choices on gender, risk preferences, ability, overconfidence, and potentially other factors. The fraction attributed to competitiveness is the ratio between the first two columns. The controls column specifies the exact control variables used in the regression. More details on these variables are in the main text.
*** p<0.01, ** p<0.05, * p<0.1

1997). One study also controlled for ambiguity aversion (AA) using an MPL.

Many measures are considered for overconfidence. A popular measure is the guessed rank (GR), typically elicited for the forced tournament (Task 2). One study includes two guessed rank measures, one for the forced tournament and one for the Task 1 piece rate (GR1). Several studies subtract the actual rank from the guessed rank to get a measure of overconfidence (OC). Other studies ask participants to guess the probability with which they would beat another person in their session (BO). One study elicits the subjective probability of winning (SPW). A few studies ask participants for their expected performance on the task (EP), and one study also asks participants how certain they are about their prediction (PC). One study also includes two measures of overconfidence in an unrelated cognitive reflection test (CRT1 and CRT2).

In addition, a few studies also included the choice in a control treatment (CT). This variable is argued to pick up the effect of overconfidence, risk preferences, and possibly other variables including ability or feedback aversion. It should also be noted that even studies that are classified as using similar controls (e.g., MPL) typically use different versions, e.g., with differently sized incentives.

A large majority of studies also control for ability. Common measures include performance in a forced tournament (TP), performance in a forced piece rate (PP), and the difference between the two (PD). Several studies also include a measure for the true probability of winning the tournament (PW). Some studies also control for mathematical ability (M). Finally, several studies also control for additional variables, including age (A), education (EC), various social preferences (SP) and patience (PM).

Several other remarks are in order. First, we include four studies that do not use the addition problem task, but use a similar design with similar outcomes. Buser,

Geijtenbeek, and Plug (2015) use a matrix addition task in which participants have to find two numbers in a 3x3 matrix that jointly add up to 10. Gneezy, Pietrasz, and Saccardo (2016) use a ball-throwing task, where participants have to throw a ball into a basket from several meters away. Flory et al. (2016) let participants sort six blocks from smallest to largest. Masclet, Peterle, and Larribeau (2015) let them decode numbers into letters. However, we do not incorporate studies that use different tasks with the aim of finding a smaller gender difference. Instead, we briefly discuss these studies in the next section.

Second, several studies are included in the sample more than once. Buser, Dreber, and Mollerstrom (2016), Zhang (2013), and Flory et al. (2016) carried out multiple experiments within the same study that all fit our criteria. We treated each of these experiments as a single observation. Gillen, Snowberg, and Yariv (2015) ran a single experiment, but present many estimates that fit our criteria. We select two representative ones for Table VI; each of these estimates only receive half the normal weight in computing the total average effect.

Third, in Dohmen and Falk (2011) and Masclet, Peterle, and Larribeau (2015), participants choose between a tournament and a fixed amount of money (as opposed to a piece rate). Risk preferences, overconfidence, and competitiveness can still explain competitive choices in both cases, and hence we include these studies in our survey as well. Fourth, Dargnies (2012) does not report a relevant regression, but sent us her raw data, allowing us to run the regression ourselves.

Fifth, several studies never report the raw gender gap in tournament choices, but always control for at least one variable (typically ability). Sixth, several studies report probit coefficients, which cannot directly be compared to the raw gender gap. In both cases, we compare coefficients between the regression controlling for risk preferences and overconfidence, and the regression controlling for the smallest number of other

variables.

A2.3 Results

The key result in Table VI is that the average fraction attributed to competitiveness is 69%. We discuss this result extensively in the main text.

It is also interesting that the raw gender gap differs considerably across studies, ranging from 8 to 53 percentage points. This could be due to a number of factors, including changes in the design (e.g., shorter tasks, internet versus laboratory experiment), as well as the population studied in the paper (students vs. non-students, Western vs. non-Western). Nevertheless, the raw gender gap is significant at the 10% level in all but one of the 26 experiments.

The fraction attributed to competitiveness also differs strongly across studies. Individual estimates fall anywhere between 25% and 122%. Part of this variation can likely be explained by differences in the sample and design of the experiments. Another explanation is that certain control variables may be better than others at filtering out the effect of risk preferences and overconfidence. Indeed, the two estimates we include from Gillen, Snowberg, and Yariv (2015) use different controls in the same data, and find very different results. A third potential reason is measurement error, which increases the variation in estimates that may be obtained.

A3. Literature Comparison: Regressions versus Design

In the discussion, we examined the results of several earlier studies that allow us to compare regression results to direct treatment comparisons. Here, we discuss each of the included studies in greater detail.

TABLE VII: Regressions and Treatment Comparisons in the Literature

Study	Gender Gap		Attributed to Competitiveness	
	Tournament	Control	Treatment Comps	Regressions
Niederle and Vesterlund (2007)	38	30	21.1	42.7
Reuben, Sapienza, and Zingales (2015)	26.8	22.0	17.9	49.6
Dohmen and Falk (2011)	25	23.1	7.6	62.8
Sutter and Glätzle-Rützler (2015)	21	14.5	31.0	79.5
Dargnies (2012)	33.3	10.6	68.2	89.8
Healy and Pate (2011)	53	15	71.7	97.0
Lee, Niederle, and Kang (2014)	7.6	5.3	29.7	122.4
Average			35.0	77.7

Notes. The first column presents the gender difference in the percentage of participants who chose the tournament. The second column presents the gender difference in the percentage of participants who chose the risky option in the non-competitive control treatment. The percentage attributed to competitiveness in treatment comparisons equals 1 minus the ratio of the first two columns. For more details on the regression approach, we refer the reader to Table VI.

A3.1 Method

Table VII presents more details on the studies used to generate Figure VI. These studies were selected if they (1) reported the raw gender difference in tournament choices, (2) included a regression controlling for overconfidence and risk preferences, and (3) reported the raw gender difference in a control treatment that could be used to isolate competitiveness from other factors.

Six of the studies included in the Table have a similar setup. First, participants solve addition problems and decide between piece rate and tournament incentives, as in Task 3 in this study. Second, they go through a control treatment in which they are asked to submit their performance from an earlier part of the experiment (typically Task 1, the forced piece rate) to either tournament or piece rate incentives. Starting with Niederle and Vesterlund (2007), it has been argued that this additional treatment is similar to the standard tournament choice in terms of risk preferences

and overconfidence, but is no longer competitive.

The other study (Dohmen and Falk, 2011) uses a different design. There, one-third of the participants choose between tournament incentives and fixed pay. The other participants instead choose between fixed pay and either a piece rate or revenue sharing. Where the tournament is competitive, the piece rate and revenue sharing are not. However, all three options are riskier than fixed pay, and are also more likely to attract confident participants. If one is willing to assume that risk preferences and overconfidence play a similar role across the three treatments (a strong assumption), then the piece rate and revenue sharing treatments can serve as a control for the tournament treatment.

The data reported in these studies allow for direct treatment comparisons between the baseline and control treatments. If competitiveness is important, the gender gap should be smaller in the non-competitive control treatment. Hence, the ratio between the gender gap in the control treatment and the gender gap in the tournament reflects the importance of competitiveness in these studies. Interestingly, none of these studies appear to make this comparison themselves, instead using regressions to control for risk preferences and overconfidence as per the previous section.

A3.2 Results

The key result in Table VII is the difference between the regressions and treatment comparisons. Using regressions, these studies estimate that approximately 78% of the gender difference in tournament choices can be attributed to gender differences in competitiveness. By contrast, treatment comparisons estimate the corresponding percentage to be around 35%. A more extensive discussion of this result can be found in the main text.

Note that our analysis compares gender differences in levels (percentage points).

An alternative approach would be to compare gender differences in percentages or ratios. The two approaches may give different results if people are overall more or less likely to compete in the control treatment as opposed to the baseline. For example, the former approach would treat a 40/20 and 30/10 gap as equal, while the latter would consider the second gap to be 50 percent larger. However, for our sample of studies the percentage/ratio approach attributes 32% of the gender gap to competitiveness, which is very similar to the level estimate (35%).

A3.3 Other Related Work

While the studies in Figure VI allow us to investigate the importance of competitiveness, there are also studies that allow us to investigate overconfidence. These studies decrease or eliminate gender differences in overconfidence by design, either by using a less stereotypically male task (Grosse, Riener, and Dertwinkel-Kalt, 2014; Shurchkov, 2012; and Dreber, von Essen, and Ranehill, 2014) or by providing performance feedback (Ertac and Szentes, 2011; and Wozniak, Harbaugh, and Mayr, 2014). Assuming competitiveness and risk preferences are unaffected by these changes (a strong assumption), these studies allow us to use treatment comparisons to isolate overconfidence from other factors.

The results of the direct treatment comparisons imply that respectively 62, 64, 70, 87, and 74% of the gender gap are due to gender differences in overconfidence.[18] Note that these numbers may capture both the direct effect of overconfidence and its interaction with risk attitudes, and are hence in line with our results (48-85%). These studies therefore suggest a substantially larger effect of overconfidence than studies relying on regressions.

18. The numbers are calculated by taking one minus the ratio between the gender gap in the control treatment and the gender gap in the baseline. For Shurchkov (2012) we take the data from the high pressure treatment only.

A direct comparison to the regression-based method is more difficult, since only Shurchkov (2012) and Wozniak, Harbaugh, and Mayr (2014) report a regression that controls for overconfidence without also controlling for risk preferences. However, the results of these two studies are in line with the previous section. Regressions attribute respectively -5% and 39.1% of the gender gap to overconfidence, considerably less than the 64% and 74% estimated through treatment comparisons. As before, the regression approach therefore appears to underestimate the importance of overconfidence (and overestimate the importance of the residual component), relative to a direct treatment comparison.

There is also one study that allows us to investigate risk preferences in a similar way. Cadsby, Servátka, and Song (2013) ask participants to choose between a piece rate and obtaining tournament payoffs with 25% chance. This treatment eliminates gender differences in overconfidence and competitiveness, and imposes an objective probability of .25 on every participant. Relative to the baseline, this reduces the gender gap in tournament choices from 27 to 9 percentage points. Assuming that the true distribution of the objective probability of winning is well approximated by a constant probability of .25 (a strong assumption), this implies that risk preferences explain 33% of the gap in tournament choices, similar to the 28% obtained in this study.[19]

Finally, Flory, Leibbrandt, and List (2015) run a field experiment that compares gender differences in application rates for real jobs with different incentive schemes. In line with the results from the lab, they find that women disproportionally shy away from competitive jobs. More importantly for our purposes, they find a near-

19. This study does not control for risk preferences using regressions. Grosse, Riener, and Dertwinkel-Kalt (2014) run a similar control treatment and use it to control for risk preferences in a regression. They do not report the raw gender gap in the control treatment, but find that controlling for risk preferences and many other variables, if anything, increases the gender gap.

identical gender gap for a non-competitive job with similar wage uncertainty. Similar to our results, their study therefore suggests that removing the competitive element by design does not affect the gender difference in people's choices.

A4. Assumption of the Treatment Comparisons

In this section, we discuss the identifying assumption of our treatment comparisons using a formal framework. We do so by splitting the assumption into distinct parts for the treatment and control variables, and discussing several sets of sufficient (but not necessary) conditions under which the identifying assumption holds.

A4.1 A Formal Framework

Let r^i, o^i, c^i be the risk preferences, overconfidence and competitiveness of individual i. Let x^i be a potential confounding variable that may affect gender differences in tournament choices, such as ambiguity aversion or social preferences. Further, let R, O, C and X be the empirical distribution of the respective variables over the whole population. For our purposes, it is convenient to separate the population into two parts – men and women – e.g., $R = \{R_M, R_F\}$. Finally, let

$$(R, O, C, X) = (\{R_M, R_F\}, \{O_M, O_F\}, \{C_M, C_F\}, \{X_M, X_F\})$$

be the joint empirical distribution of the four variables across the population.

We are interested in explaining the gender difference $G_T(R_T, O_T, C_T, X_T)$ observed in tournament choices T. Our goal in the experiment is to investigate whether G_T changes when we eliminate the effect of gender differences in one of the three explanatory variables.

A4.2 Treatment Variables

This can be done in two distinct ways. One approach is to prevent the variable of interest from affecting tournament choices. We use this approach to identify competitiveness. Specifically, the goal is to compare the tournament T to an otherwise identical environment T' where competitiveness can no longer explain gender differences in choices, i.e., $G_{T'}(R_T, O_T, X_T)$. In the experiment, we approximate $G_{T'}$ using the gender gap G_N in treatment NoComp. Hence, our assumption is the following:

ASSUMPTION 1. $G_N(R_N, O_N, C_N, X_N)$=$G_{T'}(R_T, O_T, X_T)$

In words, the gender difference in treatment NoComp (G_N) is assumed to be identical to the difference that would have been observed in the tournament had competitiveness not been able to explain tournament choices.

A second approach is to eliminate the gender difference in the variable of interest. We could, in theory, also have eliminated gender differences in competitiveness, and then examined $G_T(R_T, O_T, C_0, X_T)$, where C_0 is such that $C_M = C_F$. In that case, the identifying assumption would have reduced to:

ASSUMPTION 2. $G_N(R_N, O_N, C_N, X_N)$=$G_T(R_T, O_T, C_0, X_T)$

In words, this assumes that the gender difference in treatment NoComp is identical to the difference that would have been observed in the tournament in the absence of gender differences in competitiveness. In the experiment, we had no way to eliminate gender differences in competitiveness, and therefore used the first approach to identify competitiveness. For the remaining variables, we used the second approach.

A4.3 Control Variables

In the remainder of this section, we examine several special cases of sufficient conditions for the control variables under which assumption 1 and 2 hold. This allows

us to formalize the intuition – expressed in the discussion – that potential confounds and measurement error in beliefs need not affect our results. For illustrative purposes, we focus on competitiveness as the treatment variable of interest, the conditions for the other treatment comparisons are analogous. We start by looking at a strong set of sufficient conditions.

CONDITION 3. For the treatment variable C at least one of the following holds

1. $G_N() = G_T(.,.,C_0,.)$ (No gender differences)
2. $G_N() = G_{T'}()$ (No effect on choices)

CONDITION 4. For the control variables R, O and X:

$(r_N^i, o_N^i, x_N^i) = (r_T^i, o_T^i, x_T^i)\ \forall i.$ (1-to-1 correspondence)

Imposing condition 3 allows us to focus on the control variables. In words, it states that the treatment either eliminates the gender difference in the treatment variable, or eliminates the effect of this variable on choices.

Condition 4 requires the risk attitudes, overconfidence, and other control variables governing choices to be identical in the tournament and treatment NoComp for all individuals. Since treatment NoComp is constructed using elicited beliefs, in practice this would also require that beliefs are perfectly measured. However, condition 4 is unnecessarily strong and the following (weaker) condition is sufficient as well.

CONDITION 5. For the control variables R, O and X:

$(R_T, O_T, X_T) = (R_N, O_N, X_N)$ (Identical distribution)

Condition 5 allows individual risk attitudes, beliefs, and other factors to vary across treatments, as long as the joint distribution is the same in both cases. This allows elicited beliefs to be noisy, provided that errors cancel out, on average, in a way

that leaves the distribution unaffected. If we assume that the three variables are independent conditional on gender, we can also rewrite the condition to apply to the distribution of each variable individually.

Condition 5 may be violated if, for example, the distribution of elicited beliefs has a higher variance than the latent belief distribution. However, we can weaken the condition even further:

CONDITION 6. $G_T(R_T, B_T, C, X_T) = G_N(R_N, B_N, C, X_N)$

(Changes in control variables do not affect the gender difference)

Condition 6 says that assumptions 1 and 2 will also hold if changes in the control variables do not affect the gender difference G. For example, beliefs may change in a way that makes men and women adjust their behavior in a similar direction, leaving the gender gap unaffected. As discussed in the discussion, we see no reason for this condition to be violated for risk attitudes or potential confounds. Whether this condition is realistic for beliefs is an empirical question that will be addressed in the next section.

A5. Accuracy of Elicited Beliefs

Treatment NoComp is constructed using the beliefs elicited in Task 4. As we saw in the previous section, it is therefore important that elicited beliefs are a sufficiently accurate representation of the latent belief distribution.

We took several steps to maximize the accuracy of our elicited belief measure. First, we used a continuous measure, allowing participants to express a wide range of beliefs. Second, we used monetary incentives, which were carefully explained to participants using instructions taken from Mobius et al. (2014). Third, our belief measure captures the expected probability of winning the tournament, which is precisely

the belief that expected utility theory says determines participants' entry decisions. Fourth, we elicited the expected probability of winning directly *before* the tournament choice. This ensures that participants had the same information (past performance and expected future performance) for both the belief elicitation task and the tournament choice.

Several indicators suggest that elicited beliefs are indeed similar to the latent belief distribution. We have already shown that both genders are overconfident, and men are more overconfident than women, which is in line with the literature. In addition, there is a sizeable correlation between elicited beliefs and prior (Task 2) tournament performance (r=.49 for women and r=.48 for men, p<.001 in both cases).

Nevertheless, beliefs in the experiment are elicited, and therefore measured imperfectly. Measurement error increases the variance of elicited beliefs, which may impact results. Specifically, as our discussion of the interaction effect illustrates, gender differences in choices are only going to appear for 'intermediate' subjective beliefs. Increased variance could, in principle, imply both a larger and smaller frequency of intermediate beliefs. If the case is the latter, the gender gap in treatment NoComp will be downward biased and will overestimate the importance of competitiveness; the converse is true if elicited beliefs are more intermediate.

We investigate the empirical effect of measurement error in two steps. First, we attempt to specify what qualifies as an 'intermediate' belief. Naturally, any definition is somewhat arbitrary, but one important component is that intermediate beliefs imply substantial variation in the choices made by participants (e.g., due to variation in risk attitudes). Task 5 elicited choices for the full range of probabilities, giving us access to the degree of agreement for the whole distribution. For probabilities in the (0.25,0.75) range, at least 25% of participants still chose the least popular option. If we define an intermediate belief as a belief that lies in this range, 68% of participants

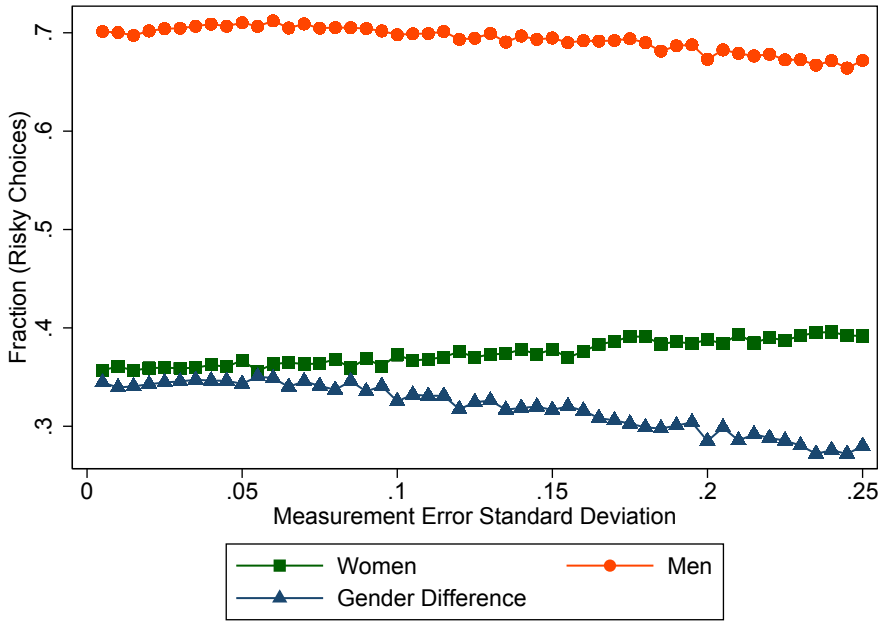

FIGURE X: CHOICES AFTER ADJUSTING OVERCONFIDENCE

Notes. The figure uses simulated data, where the elicited beliefs are perturbed 100 times for every participant using a mean zero truncated normally distributed disturbance, with 50 different values for the standard deviation of the disturbance. The disturbance term is truncated such that only beliefs in the interval [0,100] are generated. The figure plots the average fraction of risky choices made by each gender as a function of the standard deviation of the disturbance, averaged over all simulations and participants.

have intermediate elicited beliefs.

Second, we investigate the effect of additional measurement error by perturbing elicited beliefs with a truncated normally distributed noise term. Figure X presents the results. The added noise term makes beliefs less intermediate which, as expected, decreases the gender difference. The implications are twofold. First, measurement error decreases the gender gap. This implies that, if anything, the results of treatment NoComp may underestimate the gender difference and hence overestimate the importance of competitiveness. Second, and more importantly, even substantial measurement error does not greatly affect our results. For example, even when the additional noise term quadruples the variance of the elicited belief distribution (which happens

when the standard deviation term equals .22), the gender gap is still close to .30, and hence very similar to the baseline.

In summary, our beliefs appear reasonably accurate in the sense that they strongly correlate with performance. The results of our simulation suggest that even high levels of measurement error have only a limited impact on the gender difference observed in treatment NoComp. Importantly, any bias that does occur would actually decrease the gender difference in this treatment, which would imply that we are overestimating the importance of competitiveness. The fact that the gender difference in treatment NoComp is, in fact, larger than in the baseline therefore suggests that measurement error did not contribute to our results on competitiveness.

A6. Measurement Error: Econometric Adjustment

Measurement error may explain why previous studies have attached much greater importance to competitiveness in explaining the gender gap in tournament choices. We are able to circumvent the measurement error problem using our experimental design. However, it is also possible to adjust for measurement error through better measurement or by using statistical techniques. In this section, we apply several of these techniques to our data. This presents us with additional evidence on the importance of measurement error and allows us to compare statistical and design-based ways of adjusting for measurement error.

A6.1 Better Measurement

Perhaps the most intuitive way to reduce the bias resulting from measurement error is to use better measures. This could involve using techniques that are thought to be more reliable in general, as well as using techniques that fit more closely to the

TABLE VIII: Regressions with more direct proxies

	Coefficient (p-value)			
	(1)	(2)	(3)	(4)
Female	-0.329***	-0.293***	-0.228**	-0.226**
	(0.080)	(0.083)	(0.090)	(0.089)
Holt-Laury		0.081**		
		(0.040)		
Task 5 Risk			0.125***	
			(0.045)	
Treatment NoComp Choice				0.277***
				(.089)
Constant	0.671***	0.653***	0.621***	0.481***
	(0.057)	(0.057)	(0.060)	(0.086)
Observations	140	140	140	140

Notes. OLS Estimates, robust standard errors in parentheses. Dependent variable is a dummy for tournament choice in Task 4. Holt-Laury is the number of risky choices in the Holt-Laury task. Task 5 Risk is the number of risky choices in Task 5. These variables are both standardized to have mean zero and standard deviation 1. The remaining variable is the choice in treatment NoComp. *** p<0.01, ** p<0.05, * p<0.1

specific experiment. For example, rather than using a standard multiple price list such as Holt and Laury (2002), one might adapt the payoffs of the price list to more closely approximate the payoffs involved in the tournament choice.

We do this in Task 5, which was constructed to closely approximate the payoffs in the tournament choice. In the main text, we used three individual rows from Task 5 to construct our control treatments. Here, we instead use the number of risky choices taken in Task 5 as a proxy for risk preferences in a regression. In Table VIII, we compare the results for this measure to Holt-Laury. Note that in the table and the rest of this section, we will no longer control for ability, allowing us to focus on the effect of overconfidence (beliefs) and risk preferences. We also standardize all non-binary explanatory variables, in order to facilitate comparisons.

Controlling for the Holt-Laury measure reduces the gender coefficient from -.329

to -.293, an 11% decrease (we observe a similar effect for the Eckel-Grossman and SOEP measure). Controlling for the Task 5 risk measure instead reduces the gender coefficient to -.228, which is a decrease of 31%. In other words, using a specifically tailored control variable increases the implied importance of risk preferences by nearly a factor of three.

Another way to improve measurement is presented by Niederle and Vesterlund (2007), whose measure of risk preferences and overconfidence (choice in a control treatment) has a very similar payoff structure to the main experiment. This measure is more directly tailored to the main experiment, and could therefore be expected to better reflect the relevant risk preferences and overconfidence than more general measures. While we did not run the same control treatment, we can use treatment NoComp as a similar sort of proxy for risk preferences and overconfidence.

The results are presented in column (4) of Table VIII. Relative to the Task 5 risk measure, the choice in treatment NoComp also controls for overconfidence and could therefore be expected to decrease the gender difference even further. At the same time, being a binary variable decreases precision in measurement. Interestingly, the gender coefficient after controlling for treatment NoComp is nearly identical to the one obtained using the Task 5 risk proxy. This suggests that the effects of decreased precision and also controlling for overconfidence cancel out on average.

Though directly tailored measures of risk preference may outperform standard measures, they are still imperfect. While they may reduce one source of measurement error (misfit between experiment and proxy), other sources of error (e.g., mistakes, binary/ordinal scale) still remain. Indeed, the 31% attributed to risk preferences and overconfidence in column (4) is considerably less than the 113% attributed to these variables using treatment comparisons.

A6.2 Multiple Measures

An alternative approach is to measure risk preferences and/or overconfidence several times. Multiple measures can then be used to filter out some of the noise from the regression estimates. One way to use these measures is to include all of them simultaneously in a single regression. The idea is that these controls may jointly capture a greater fraction of the total effect than any single control variable on its own. Following Niederle and Vesterlund (2007) and Buser, Niederle, and Oosterbeek (2014), we already applied this technique in Table III, where we showed that adding additional control variables further reduced the gender coefficient.

Column (1) of Table IX does a similar analysis for risk preferences only, omitting the coefficients for ability and overconfidence. Relative to a regression with just a single control (columns 2 and 3 of Table VIII), the gender coefficient is smaller. Controlling for additional proxies of risk preferences therefore does increase the fraction of the gender gap that is attributed to risk preferences. However, the added benefit relative to controlling for just the Task 5 risk measure is a relatively pedestrian 2 percentage points.

Including all four measures of risk preferences into a single regression may lead to multicollinearity and overfitting issues. A different approach is to combine the multiple measures into a single variable. In column (2), we construct a risk index by taking the average of the four standardized risk measures. In column (3), we instead use factor analysis to create a weighted sum. The weights assigned to each variable in the latter case are nearly identical, which explains why we obtain very similar results with both approaches. The results suggest that, relative to including each variable separately, using an index of risk preferences does not affect the size of the gender coefficient.

TABLE IX: Regressions with multiple measures

	Coefficient (p-value)				
	(1)	(2)	(3)	(4)	(5)
Female	-0.204**	-0.217**	-0.217**	-0.221**	-0.144
	(0.094)	(0.091)	(0.091)	(0.096)	(0.124)
Holt-Laury	0.018			0.244***	
	(0.046)			(0.088)	
Eckel-Grossman	0.042				0.300**
	(0.046)				(.120)
SOEP	0.031				
	(0.042)				
Task 5 risk	0.090*				
	(0.052)				
Risk Index		0.133***			
		(0.043)			
Risk Factor			0.134***		
			(0.043)		
Constant	0.609***	0.616***	0.616***	0.617***	0.579***
	(0.062)	(0.057)	(0.060)	(0.065)	(0.074)
Method	OLS	OLS	OLS	IV	IV
Observations	140	140	140	140	140

Notes. Robust standard errors in parentheses. Dependent variable is a dummy for tournament choice in Task 4. Holt-Laury, Eckel-Grossman, SOEP and Task 5 risk are the standardized scores on the respective elicitation method. The risk index is the standardized sum of the four methods. The risk factor is a standardized weighted sum of the four methods, with the weights determined by a factor analysis. For the IV regressions, the instruments include the female dummy, the SOEP question, Task 5 risk, and the other omitted risk preference measures.
*** p<0.01, ** p<0.05, * p<0.1

Repeated measures can also be used to filter out the noise using instrumental variables. The IV estimator filters out measurement error by using one or more proxies of risk preferences as instruments for another. A key assumption here is that there is no correlation between measurement error in the two measures. This assumption may not hold in practice, for example, if participants make similar mistakes in several elicitation procedures. To our knowledge, this method has only been used in the experimental literature on gender differences by Gillen, Snowberg, and Yariv (2015).

Columns (4) and (5) of Table IX present the results of an application of this method to our data. In these columns, we respectively instrument the Holt-Laury and Eckel-Grossman measure using the other three measures of risk preferences. Using instrumental variables triples the coefficient of Holt-Laury (compare column 2 of Table VIII), and has an even larger effect on the Eckel-Grossman coefficient (the OLS coefficient is .080). Thus, adjusting for measurement error using IV greatly increases the coefficient estimates for risk preferences in our study.

In column (5), the gender coefficient also decreases strongly, implying that risk preferences by themselves explain 56% of the gender difference in tournament choices. This number is no longer very different from the approximately 65% implied by our data (including the interaction term). However, it is important to note the difference between column 5 and column 4. When Holt-Laury is the instrumented variable (column 4), the gender coefficient is similar to the previous columns (-.221). When Eckel-Grossman is used, the gender coefficient drops to -.144 (column 5). Had we used one of the other two measures as the instrumented variable, the gender coefficient would have been either -.154 or -.163. Hence, the conclusions of the IV approach may differ based on which variables are selected as instruments.

It is not unusual that the results of IV regressions differ somewhat depending on which variable(s) are used as instruments. However, this does complicate the

interpretation of the results. Gillen, Snowberg, and Yariv (2015) address this problem (for the case of two proxies) by separately using each measure as an instrument for the other. They then take the average of the two coefficients as their estimate of the true effect, an approach they refer to as obviously related instrumental variables (ORIV). In our case, taking the average of the four IV regressions gives us a gender coefficient of .171, implying that 48% of the gender difference in tournament choices is driven by gender differences in risk preferences. This is a far cry from the 11% we obtained by using a single general control question for risk preferences, as is common practice in the literature.

In summary, measuring a noisy explanatory variable several times may alleviate if not eliminate the measurement error problem, especially through the use of instrumental variables. However, in the latter case the results vary considerably depending on which variables are used as instruments.

A6.3 Errors-in-Variables Regressions

A disadvantage of the techniques in the previous section is that they require multiple measures. For beliefs, we have only two measures of overconfidence, one for Task 4 and one for Task 4b. These two measures are highly correlated (.87), implying either near-perfect measurement or highly correlated measurement error. In the latter case, which seems more plausible, the IV method would no longer be able to filter out all measurement error, and would not sufficiently adjust the estimated coefficient upwards.

However, when the reliability of the independent variables is known, it is possible to correct the coefficient estimates directly without the use of instrumental variables.

Following Krashinsky (2004), the vector of adjusted coefficients can be estimated as

$$\beta_{\text{ADJ}} = (X'X - n\Sigma_\eta)^{-1}X'Y \tag{4}$$

We change our notation such that β_{ADJ} is now a vector of adjusted coefficients, Σ_η is the variance-covariance matrix of measurement errors, and $X = (F, R, B)$ is a matrix containing gender, a proxy for risk preferences and a proxy for overconfidence (i.e., beliefs). We assume that gender is perfectly measured and measurement error in risk preferences is not correlated with measurement error in beliefs. Further, let σ_* be the variance of the latent variable, and let σ_η the measurement error of the elicited proxy. For our regressions, we will vary the reliability ratio $\frac{\sigma_*}{\sigma_* + \sigma_\eta}$ and adjust Σ_η correspondingly. Assuming that measurement error is not correlated across two measures of the same variable, the reliability ratio equals the correlation coefficient between two measures and can thus be estimated in our data.

The regression results are presented in Table X. We use the Task 5 measure of risk preferences, since it appears to be the least noisy measure based on the results of Table VIII. Column 1 presents the standard unadjusted OLS results. Both beliefs and risk preferences are significant, and jointly explain approximately 43% of the gender difference in tournament choices. For column (2), we assume that both measures have a reliability of .7. Given the literature, which we review below, this represents a fairly optimistic estimate that implies that only 30% of the variation in the elicited measures is due to noise. Nevertheless, the coefficient for gender already falls to -.104, and is no longer significant.

In columns (3) and (4), we lower the reliability to .6 and .5 respectively, for both variables. This pushes the gender coefficient even closer to zero; in the second case the coefficient for gender is even positive, though not significant. Columns (5) and

TABLE X: Errors in Variables Regressions

	Coefficient (p-value)					
	(1)	(2)	(3)	(4)	(5)	(6)
Female	-0.187**	-0.104	-0.050	0.039	-0.025	-0.140
	(0.087)	(0.097)	(0.105)	(0.118)	(0.112)	(0.093)
Task 5 Risk	0.131***	0.200***	0.260***	0.344***	0.326***	0.137***
	(0.042)	(0.064)	(0.078)	(0.099)	(0.100)	(0.042)
Belief	0.081**	0.126**	0.154**	0.200**	0.093**	0.173**
	(0.040)	(0.057)	(0.067)	(0.080)	(0.038)	(0.083)
Constant	0.600***	0.559***	0.532***	0.488***	0.520***	0.577***
	(0.058)	(0.061)	(0.064)	(0.069)	(0.067)	(0.060)
Reliability Risk	1	.7	.6	.5	.5	1
Reliability Belief	1	.7	.6	.5	1	.5
Observations	140	140	140	140	140	140

Notes. Robust standard errors in parentheses. Dependent variable is a dummy for tournament choice in Task 4. Task 5 risk is the standardized number of risky choices in Task 5. Belief is the probability of winning elicited in Task 4. The reliability ratio is $\frac{\sigma_*}{\sigma_*+\sigma_\eta}$, the ratio between the assumed variance of the latent variable, and the variance of the measured variable.
*** p<0.01, ** p<0.05, * p<0.1

(6) show that a comparable reduction in the assumed reliability has a larger effect for risk preferences than for beliefs. This suggests that either risk preferences are more important in explaining the gender difference in tournament choices or that the risk preference measure is less noisy than the belief measure (or both).

Do these results imply that the gender difference in tournament choices can be fully attributed to gender differences in risk preferences and overconfidence? Possibly, but only if the reliability ratios in these columns are accurate. Though the evidence presented in the literature and below suggests that risk preference measures may have considerably smaller reliability ratios than .5, the particular measure we use here was constructed to be very similar to the tournament choice and may therefore be an unusually good predictor of behavior in the experiment. Similarly, we have no direct way of estimating the reliability of the belief variable, and it may therefore very well

be less noisy (or noisier) than assumed in the table.

A6.4 Structural Equation Modelling

Structural Equation Models (SEMs) are a popular tool in the social sciences for estimating relationships between latent (i.e., imperfectly measured) variables. We are interested in the effect of two latent variables – risk preferences and overconfidence – neither of which is perfectly observed. Using an SEM, it is possible to simultaneously estimate these latent variables through a measurement model, and estimate their effect on the variable of interest (tournament choices). Naturally, the method relies on several strong assumptions, for an overview see, for example, Kline (2005).

For our estimations, we assume that the first latent variable (risk preferences) is imperfectly measured by four proxy variables: Holt-Laury, Eckel-Grossman, the SOEP question and the Task 5 risk measure. Similarly, we assume that beliefs are imperfectly measured by the elicited beliefs from Task 4 and Task 4b. Following Niederle and Vesterlund (2007), we further assume that the choice made in treatment NoComp can serve as a proxy for both risk preferences and beliefs. This gives us a measurement model where the two latent variables are measured by five and three proxies respectively.

In addition to these measurement models, we also simultaneously estimate two additional equations. First, we assume that tournament choices are a function of risk preferences, beliefs, and gender, similar to our earlier regressions. Second, we assume that differences in beliefs are caused by differences in ability, as reflected by the objective probability of winning p_o (as computed using performance in Task 2) and the difference between Task 2 and Task 1 performance. In other words, we allow ability to influence choices only indirectly, through beliefs. Using standard SEM terminology, we will refer to these two equations as structural models.

TABLE XI: Structural Equation Model Estimates

Parameter	Estimate	Std. Error
Measurement Model: Risk Preferences		
SOEP	1	
Eckel-Grossman	0.989***	(0.229)
Holt-Laury	1.130***	(0.235)
Task 5 Risk	1.843***	(0.326)
Treatment NoComp Choice	1.091***	(0.208)
Measurement Model: Beliefs		
Belief Task 4	1	
Belief Task 4b	0.877***	(0.056)
Treatment NoComp Choice	0.591***	(0.057)
Structural Model: Beliefs		
p_o	0.406***	(0.085)
T-PR	0.014	(0.031)
Structural Model: Tournament Choice		
Female	-0.144	(0.090)
Risk Tolerance	0.330***	(0.109)
Belief	0.081**	(0.039)
Constant	0.577***	(0.061)
Observations	140	

Notes. Holt-Laury, Eckel-Grossman, SOEP and Task 5 risk are the standardized scores on the respective elicitation method. Beliefs are the probability of winning elicited in Task 4 and Task 4b, respectively. Treatment NoComp Choice is a dummy for the choice in treatment NoComp. p_o is the objective probability of winning, based on Task 2 performance. T-PR is the difference between performance in the tournament (Task 2) and piece rate (Task 1). In the final rows, risk tolerance and belief are the latent variables for risk preferences and beliefs, respectively. In total, the results are based on seven simultaneously estimated equations, the two structural models plus one equation per proxy variable in the measurement models.
*** p<0.01, ** p<0.05, * p<0.1

The results are estimated using maximum likelihood and are presented in Table XI. With the exception of gender, all variables are standardized to have a mean of zero and a standard deviation of 1. The first two parts of the table present the results of the measurement equations. For risk preferences, Task 5 has a considerably larger weight than the other measures, in line with the idea that it is a better proxy of risk preferences in this setting. For beliefs, the beliefs elicited in each task have a similar weight. The dummy variable for choices in treatment NoComp has a smaller weight than the elicited measures of beliefs.

The first structural model reaffirms the result of the previous section that beliefs and ability are highly correlated. Better participants (as reflected by their probability of winning) have more optimistic beliefs. At the same time, the difference between performance in the tournament and piece rate (which reflects the effect of competitive incentives on performance) does not significantly correlate with beliefs.

The second structural model shows that latent beliefs and, in particular, risk preferences have a strong effect on tournament choices. A one standard deviation increase in risk tolerance increases the probability of entering the tournament by 33 percentage points. The estimated effect for risk preferences greatly exceeds the estimate obtained with standard regressions, and is comparable only to the Instrumental Variables approach and errors-in-variables regressions with a low assumed reliability.

By contrast, the belief variable has a relatively modest effect. A likely explanation is that measurement error was correlated across the two elicitations, which implies that our measurement model for beliefs was unable to filter out all measurement error. This would then imply that the gender coefficient, which is equal to -.144, may still be overestimated.

We only discuss the results of a single SEM for reasons of space. Naturally, we could have estimated a number of different models which, for example, did not

directly incorporate a causal relationship between ability and beliefs, did not use treatment NoComp as a proxy for risk preferences or overconfidence, etc. While the results of these alternative models differ slightly, all give qualitatively similar results. Specifically, the gender coefficient tends to vary very little, and always remains in the -.15 to -.12 range, and is never significant.

A6.5 Final Remarks

These results have several implications. First, they point toward the existence of substantial measurement error in the proxy variables we used in Table III, particularly risk preferences. Adjusting estimates for measurement error by using a larger number of proxies, better measurement or statistical techniques all increased the implied importance of risk preferences and, to some extent, overconfidence. As shown by Gillen, Snowberg, and Yariv (2015), substantial measurement error implies that previous studies – which did not adjust their estimates for measurement error – underestimated the importance of risk preferences and overconfidence in their data.

Second, the greater the adjustment in the risk and belief coefficients, the smaller the residual gender coefficient. In many of our specifications, the residual gender coefficient was no longer significant, in some it was even a precisely estimated zero. This suggests that regressions that do not account for measurement error overestimate the residual gender coefficient and hence the importance of competitiveness. By contrast, the adjusted estimates of this section are more consistent with our experimental comparisons and suggest that competitiveness explains at most a small part of the gender gap in tournament choices.

Third, our results illustrate that good measurement and statistical techniques can serve as complements. The risk preference coefficient was largest in regressions that combined both statistical techniques (IV or SEM) and a specifically tailored measure

of risk attitudes (from Task 5). In contrast, we did not have enough measures of beliefs to filter out the noise using statistical techniques. Indeed, we interpret the relatively small coefficient estimates for beliefs in e.g., Table XI as evidence of an inability to fully adjust our estimates for measurement error.

A7. Measurement Error: Evidence

In this section, we review and present evidence that is strongly suggestive of measurement error in elicited measures of beliefs and risk attitudes. We also directly estimate the signal-to-noise ratio using our data.

A7.1 Evidence

Why would we expect proxies for risk preferences and overconfidence to be noisy? One reason is that participants make mistakes when responding to elicitation tasks. They may not exactly know their risk preferences or beliefs. They may not understand the incentive scheme, etc. A second reason is that the elicitation procedure may not capture the latent variable that determines behavior in the experiment. For example, the risks taken in the Holt and Laury (2002) task involve different payoffs and probabilities than the risk taken in tournament choices. Risk preferences may also be context-specific. Third, both risk preferences and beliefs (e.g., guessed rank) are typically elicited using ordinal scales. Unless the latent variables actually follow the same ordinal scale (which seems unlikely), these proxies will therefore be unable to perfectly measure the latent variable.

These intuitive ideas are supported by studies that estimate the amount of measurement error in the data. One approach relies on the idea that perfectly measured proxies that reflect the same underlying construct should be perfectly correlated. In-

stead, the literature typically finds rather small correlations between different proxies, or between repeated elicitations of the same proxy. Kimball, Sahm, and Shapiro (2008); Beauchamp, Cesarini, and Johannesson (2015), and Gillen, Snowberg, and Yariv (2015) suggest that these small correlations may be due in large part to measurement error, and indeed find substantially larger correlations after adjusting for measurement error using statistical techniques. Similarly, Gillen, Snowberg, and Yariv (2015) and Ambuehl and Li (2016) find substantially larger correlations between beliefs and actions after adjusting for measurement error using statistical techniques.

We can apply a similar logic to our data. For risk preferences, we can compare correlations across three measures of risk preferences (the SOEP question, Eckel-Grossman and Holt-Laury), and can use the number of risky choices in Task 5 as a fourth proxy of risk preferences. Pairwise correlations range from .21 for the SOEP question and Eckel-Grossman to .51 for Task 5 and Holt-Laury. All six correlations are significantly smaller than 1 and imply substantial measurement error (see below).

For overconfidence, we can examine the correlation between the beliefs elicited in Task 4 and Task 4b. This correlation is very high (.87). While this could reflect highly precise measurement, a more likely explanation is that measurement error is also correlated across the two measures. For example, a misunderstanding of the incentives is likely to affect responses to the two elicitations in the same way.

Three other pieces of evidence point to the existence of measurement error. The first is the large variation in the estimates of the regression approach across studies (Table VI). Second, different proxies give different results even when applied to the same data. Third, Table III showed that the more proxies that are included, the lower the residual gender coefficient. Without measurement error, different proxies are perfectly correlated, give identical results in the same data, and adding additional proxies could not improve the fit of the regression model.

A7.2 Size

The evidence summarized in the previous section suggests that measurement error forms a substantial part of our proxies for risk preferences and overconfidence. In this section, we estimate exactly how large the measurement error component is likely to be in our data.

One approach is to assume that the results of the treatment comparisons reflect the true data-generating process (DGP). If this is true, we can combine knowledge of the true DGP with the results of the regression analysis to back out an estimate for the amount of noise in the explanatory variables. In so doing, we impose the strong assumption that the true DGP is a well-behaved linear equation with classical measurement error in risk preferences and overconfidence.

This assumption allows us to investigate the effect of measurement error using simulations, which are presented in greater detail in Figure XI below. As a first step, we regress actual tournament choices on risk preferences, beliefs, and their interaction in our data. These regressions are similar to Table III, except that we no longer control for ability and include an interaction term. Depending on the proxies we use, we obtain residual gender differences ranging from 58% to 80% of the raw gender gap.

As a second step, we use simulated data from Figure XI to investigate the amount of noise required to obtain similar results in the simulated data. Retaining a residual gender coefficient of .58 to .8 in the simulated samples requires imposing a signal-to-noise ratio between .5 and .83. This is consistent with anywhere between 55% and 67% of the total variance in our explanatory variables being due to measurement error.

A different approach is to investigate correlations between multiple measures of the same variable. Any deviation from a unit correlation between multiple measures

can be thought of as a reflection of measurement error. Under strong parametric assumptions, we can use the empirical correlation coefficients to back out the measurement error variance.

For risk preferences, we can estimate the signal-to-noise ratio by looking at correlations between the four elicited measures. Assuming that each of these measures reflects the same latent 'true' risk preference variable R^*, we get:

$$R_1 = R^* + \eta_1 \tag{5}$$

$$R_2 = R^* + \eta_2 \tag{6}$$

The correlation coefficient between the two risk measures $r_{R1,R2}$ equals:

$$r_{R1,R2} = \frac{COV(R_1, R_2)}{\sqrt{Var(R_1)Var(R_2)}} \tag{7}$$

Let us now assume that measurement error is classical, which implies that measurement error is not correlated with the latent variable $COV(R^*, \eta_{R1}) = COV(R^*, \eta_{R2}) = 0$. We also impose the additional strong assumptions that the two variables in question have the same variance and that the error terms of the two proxies are not correlated, such that $\eta_{R1}, \eta_{R2} \sim N(0, \sigma_\eta^2)$. Equation (7) then reduces to:

$$r_{R1,R2} = \frac{\text{VAR}(R^*)}{\text{VAR}(R_1)} = \frac{\text{VAR}(R^*)}{\text{VAR}(R^* + \eta)} \tag{8}$$

Under these assumptions, the correlation coefficient therefore directly reflects the ratio between the variance of the latent variable risk preferences and the observed

variable (which includes measurement error). Even the highest correlation obtained in our data (.507 for Holt-Laury and the Task 5 risk measure) implies that half the variance in the elicited proxy is due to measurement error. In other words, the signal-to-noise ratio is approximately 1. The average correlation (.38) implies a signal-to-noise ratio of .62, with the smallest correlation (.212) even implying a signal-to-noise ratio as low as .27.[20]

While both approaches rely on strong assumptions, it is encouraging that they yield estimates that are in the same ballpark. Both approaches suggest that at least half, and probably more, of the variance of the control variables is due to measurement error. Our results of the second method yield estimates that are similar to Kimball, Sahm, and Shapiro (2008), Andersen et al. (2008), and Gillen, Snowberg, and Yariv (2015).

A8. Measurement Error: an Example

This section illustrates the consequences of classical measurement error in beliefs and risk preferences on regressions and treatment comparisons using a simulated example based on our data. We use a well-behaved linear data-generating process that assumes that the variables of interests are subject to classical measurement error. Specifically, the data-generating process we use to generate tournament choices Y is the following.

$$Y = \alpha_0 + \beta_1 R^* + \beta_2 B^* + \beta_3 C^* + \beta_4 B^* R^* + \epsilon^* \tag{9}$$

20. In principle, we could have done a similar analysis for overconfidence. However, this would almost certainly violate the assumption that measurement error is not correlated across measures.

$$R^* = \alpha_1 + \gamma_1 M + \nu_R^* \tag{10}$$

$$B^* = \alpha_2 + \gamma_2 M + \nu_B^* \tag{11}$$

$$C^* = \alpha_3 + \gamma_3 M + \nu_C^* \tag{12}$$

Here, M is equal to one if the participant is a man, and zero otherwise. We assume that $\beta_1 = .25$, $\beta_2 = .42$, $\beta_3 = 0$ and $\beta_4 = .33$. These parameters correspond to the results of our treatment comparisons, except that we round the effect of competitiveness to zero and adjust the other coefficients to sum to 1. We assume that beliefs (B^*), risk preferences (R^*), and competitiveness (C^*) are partially determined by gender, and we normalize the effect of gender to be equal to one, i.e., $\gamma_1 = \gamma_2 = \gamma_3 = 1$. The explanatory variables also reflect an idiosyncratic term $\nu_R^*, \nu_B^*, \nu_C^* \sim N(0, 1)$. Without loss of generality, we further normalize all constants α to zero.

To investigate the effect of measurement error, we assume that both risk preferences and beliefs are measured with classical measurement error, i.e., $R = R^* + \eta_R$ and $B = B^* + \eta_B$, where $\eta_R, \eta_B \sim N(0, \sigma_\eta^2)$. In our simulations, we vary the measurement error variance to examine its impact on the estimated effect of risk preferences, beliefs, and competitiveness. For the regression approach, we will estimate the following equation.

$$Y = \hat{\alpha_0} + \hat{\beta}_1 R + \hat{\beta}_2 B + \hat{\beta}_4 BR + \hat{\theta} M + e \tag{13}$$

This equation is similar to Table III, except that it also includes an interaction

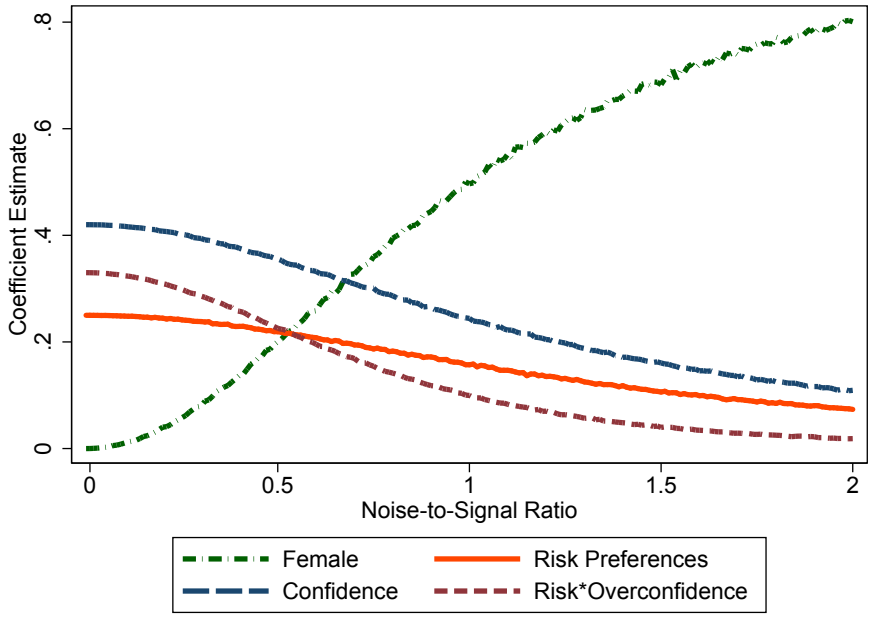

FIGURE XI: SIMULATIONS ILLUSTRATING MEASUREMENT ERROR IN REGRESSIONS

Notes. The figure plots the average coefficient estimates for gender, risk preferences, overconfidence and the interaction of risk preferences and overconfidence. The coefficient estimates represent the average coefficient over 1,000 simulated samples for 200 different values of the noise-to-signal ratio σ_η.

term in line with the results of the experiment. As in Table III, there is no variable that measures competitiveness. Instead, competitiveness is proxied for using the gender dummy. We generate 1,000 random samples for 200 different values of the measurement error variance and estimate equation (13) on each of the simulated samples. In line with our experiment, each sample consists of 70 men and 70 women.

The results are presented in Figure XI. When all variables are perfectly measured, the regression results are in line with the DGP. The gender coefficient (reflecting competitiveness) is zero and all other coefficients correspond to the DGP. However, as σ_η increases, the coefficient estimates for risk preferences, overconfidence, and the interaction term are attenuated and converge to zero. By contrast, the coefficient for

gender – zero in the DGP – grows larger and converges to one.

These results are intuitive. When risk preferences and overconfidence are measured with high levels of noise, their coefficients will only be able to capture part of the true underlying effect. After controlling for the noisy proxies of risk preferences and overconfidence, the residual gender coefficient is therefore positive and significant. Rather than reflecting competitiveness, however, it reflects whichever part of the effect of risk preferences and overconfidence the noisy proxies were unable to control for.

Next, we investigate the effect of measurement error on the results of our treatment comparisons. We focus on the comparison between the baseline and treatment NoComp, which isolates the effect of competitiveness. We assume that behavior in treatment NoComp is governed by the true DGP and true risk preferences R^*, but by noisy beliefs (B). Hence, choices in treatment NoComp (Y^1) are constructed as follows.

$$Y^1 = \alpha_0 + \beta_1 R^* + \beta_2 B + \beta_3 B R^* \tag{14}$$

The effect of competitiveness can be obtained by subtracting the size of the gender gap in Y^1 from the size of the gender gap in tournament choices Y within each simulated sample. We do this comparison using the same 200,000 random samples we used for Figure XI.

Figure XII gives the results of the comparison between the experimental and regression approaches. While measurement error increases the implied importance of competitiveness in the regression approach, there is no such effect for treatment comparisons. Independent of the amount of measurement error, the treatment com-

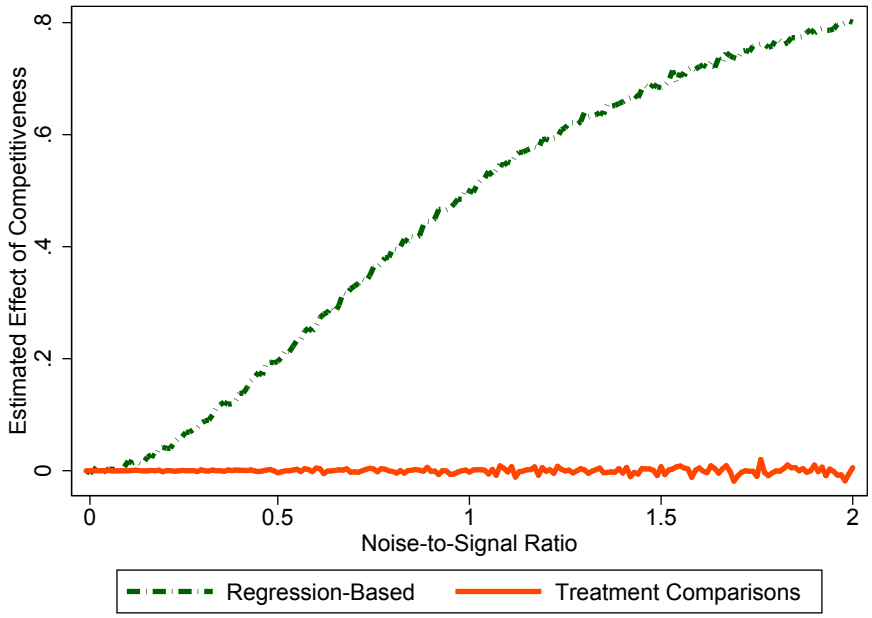

FIGURE XII: SIMULATIONS COMPARING THE EFFECT OF MEASUREMENT ERROR

Notes. The regression-based line displays the average estimated coefficient for gender in a regression of competitive choices on gender, risk preferences, overconfidence, and the interaction of risk preferences and overconfidence (over 1,000 samples). The treatment-based approach is the average gender gap in the DGP minus the gender gap in the alternative choice of equation (14) in the same simulated samples.

parisons correctly attributes none of the gender gap to competitiveness. Increased measurement error only affects the treatment-based estimator through an increase in its variance.

All in all, these results illustrate that classical measurement error systematically biases the results of the regression approach, but does not affect the average estimate of the treatment-based approach.

Appendix B: Experimental Instructions – For Online Publication

This is the English version of the instructions used for the experiment. The original German version is available upon request. The welcome section is required for all experiments in the lab at the Technical University of Berlin. The instructions for Task 1 to Task 3 closely followed Niederle and Vesterlund (2007). The instructions for the belief elicitation task were based on Mobius et al. (2014). In the experiment, we referred to Task 4b and Task 5 as Task 5 and Task 6, respectively.

Welcome

Welcome to our experiment. During the experiment, it is not allowed to use electronic devices or to communicate with other participants. On your computer, please only use the experimental software. Please do not communicate with other participants. If you have a question, please raise your hand, and we will come and answer your question in private. Please do not ask the question in a way that everyone can hear it. If the question is relevant for all participants, we will repeat and answer it for everyone. If you break one of these rules, you will be excluded from the remainder of the experiment and will not receive any earnings.

Instructions

In the experiment today you will be asked to complete six different tasks. None of these will take more than 5 minutes. At the end of the experiment you will receive €3

for having completed the six tasks, in addition we will randomly select one of the six tasks and pay you based on your performance in that particular task. Once you have completed the six tasks we will determine which task counts for payment by asking one of you to roll a six-sided die. The method we use to determine your earnings varies across tasks. Before each task we will describe in detail how your payment is determined.

Your total earnings from the experiment are the sum of your payment for the randomly selected task, your €3 payment for completing the tasks, and a €5 show up fee. At the end of the experiment you will be asked to come to the side room where you will be paid in private.

Task 1: Piece Rate

For Task 1 you will be asked to calculate the sum of five randomly chosen two-digit numbers. You will be given 5 minutes to calculate the correct sum of a series of these problems. You cannot use a calculator to determine this sum, however, you are welcome to write the numbers down and make use of the provided scratch paper. You submit an answer by clicking the submit button with your mouse. When you enter an answer the computer will immediately tell you whether your answer is correct or not. Your answers to the problems are anonymous.

If Task 1 is the one randomly selected for payment, then you will get 50 cents per problem you solve correctly in the 5 minutes. Your payment does not decrease if you provide an incorrect answer to a problem. We refer to this payment as the piece rate payment.

Please do not talk to one another for the duration of the experiment. If you have any questions, please raise your hand.

[*Participants were brought to a wait screen, and waited until everyone had finished reading these instructions. On the wait screen, they could re-read the instructions printed above. After everyone had finished reading the instructions, they then had five minutes to solve addition problems. At the end of the task, they received feedback on the number of exercises they had solved and were notified that the next task would start in 20 seconds.*]

Task 2: Tournament

As in Task 1 you will be given 5 minutes to calculate the correct sum of a series of five two-digit numbers. However for this task your payment depends on your performance relative to that of a group of other participants. Each group consists of four people. If Task 2 is the one randomly selected for payment then your earnings depend on the number of problems you solve compared to the three other people in your group. The individual who correctly solves the largest number of problems will receive €2 per correct problem, while the other participants will not receive any

payment. We refer to this as the tournament payment. You will not be informed of how you did in the tournament until all six tasks have been completed. If there are ties, the winner will be randomly determined.

Please do not talk to one another. If you have any questions, please raise your hand.

[*The wait screen, task, and feedback were identical to Task 1. Notably, participants only received feedback on their absolute performance, not on their relative.*]

Task 3: Choice

As in the previous two tasks you will be given 5 minutes to calculate the correct sum of a series of five two-digit numbers. However, you will now get to choose which of the two previous payment schemes you would prefer to apply to your performance in the third task.

If Task 3 is the one randomly selected for payment, then your earnings for this task are determined as follows. If you choose the piece rate you will receive 50 cents per problem you solve correctly. If you choose the tournament your performance will be evaluated relative to the performance of the other three participants of your group in the Task 2-tournament. The Task 2-tournament is the one you just completed. If you correctly solve more problems than they did in Task 2, then you will receive four times the payment from the piece rate, which is €2 per correct problem. You will receive no earnings for this task if you choose the tournament and do not solve more problems correctly now, than the others in your group did in the Task-2 tournament. You will not be informed of how you did in the tournament until all four tasks have been completed. If there are ties the winner will be randomly determined.

The next computer screen will ask you to choose whether you want the piece rate or the tournament applied to your performance. You will then be given 5 minutes to calculate the correct sum of a series of five randomly chosen two-digit numbers.

Please do not talk to one another. If you have any questions, please raise your hand.

[*Wait screen with Task 3 instructions; subsequent feedback and task identical to Task 1 and Task 2.*]

Task 4: Choice 2

As in the three previous tasks you will be given 5 minutes to calculate the correct sum of a series of five two-digit numbers. As in the previous task, you can choose which payment scheme you would prefer to apply to your performance. There will also be an additional part of the task that will be explained on the next page.

If Task 4 is the one randomly selected for payment then your earnings for this task are determined the same way as in Task 3. In particular, you can choose between:

- Piece Rate: 50 cents per problem you solve correctly.

- Tournament: €2 per correct answer if you correctly solve more problems than your group members did in Task 2, €0 otherwise.

You will not be informed of how you did in the tournament until all six tasks have been completed. If there are ties, the winner will be randomly determined.

Task 4: Robots

Imagine that you live in a world full not only of TU students, but also full of robots. This is Bob the Robot. Bob is going to solve exercises too, along with all his clones – 100 robots in all. On average the robots are about as good at the exercises as TU students, but some are much better than others. In fact, they have been programmed so that

- Bob 1 has a 1% chance of scoring better than your three group members in Task 2.
- Bob 2 has a 2% chance of scoring better than your three group members in Task 2.
- ...etc...
- Bob 100 has a 100% chance of scoring better than your three group members in Task 2.

One of these robots will be assigned to be your robot. But we aren't going to tell you which robot it is until the end of the experiment – it could be any of the 100 models.[21]

For this part of Task 4 (the Robot part) you can earn €2 (in addition to the payment discussed on the previous page). For this, you can use either your performance on this task or the performance of your Bob. Whichever you use, you will earn €2 if that performance is larger than the number of exercises solved by your group members in Task 2.

Task 4: Robots (2)

You will thus have to help us decide whether to use your score or the robot's score to determine your payment. We are going to ask you which robot you think you are most like. That means, which of the 100 Bob clones is as likely as you are to score better than your group members in Task 2. Based on the decision you have made, we will pick either your score or the robot's, depending on who is most likely to have a better score than your group members in Task 2.

21. These instructions, and the picture of the robot, were adapted or taken from (Mobius et al., 2014)

Figure XIII: Bob the Robot (Mobius et al., 2014)

For example, suppose that you say you are as good as Bob 60. If your actual robot is Bob 34, we would base your payoff on your score, since you are then more likely to solve more exercises. But if your actual robot is Bob 97 we will use the robot's score, since the robot is then more likely to solve more exercises.

Note also that since Bob X has an X% chance of having a higher score than your group members in Task 2, you are in effect estimating the probability that you will have a higher score than your group members in Task 2. The bottom line is that you are most likely to win €2 if you are as accurate as possible when you estimate your probability of solving more exercises than your group members did in Task 2.

Check-up Question

Suppose you think that you have a 44% chance of getting a higher score than your group members in Task 2. Given that you estimate your chance of winning at 44%, which Bob should you select to have the highest chance of winning the prize of €2?

Task 4: Robots (3)

The next computer screen will ask you to choose a robot. After that, you will be asked to choose between the piece rate and the tournament. You will then be given 5 minutes to calculate the correct sum of a series of five randomly chosen two-digit numbers.

Please do not talk to one another. If you have any further questions, please raise your hand.

[*Participants went to a wait screen until all participants had finished reading these instructions. While on the wait screen, participants were able to reread the instructions for this task if they so wished. Once all participants finished the instructions, each participant then decided on his robot:*]

Please state which Bob you think you are most like. Remember, Bob X has an X% chance of having a higher score than your group members in Task 2, so you are in effect estimating the probability that you will have a higher score than your group members in Task 2. You are most likely to win the 2 if you are as accurate as possible.

I am as likely to have a higher score than my group members in Task 2 as Bob...

[*They then worked on addition problems for five minutes, after which they received feedback on their absolute performance. They did not receive any feedback on their relative performance, nor on whether the computer used a robot or their own performance to determine their earnings for the Robot task.*]

Task 5: Choice 3 and Robot 2

As in the previous tasks you will be given 5 minutes to calculate the correct sum of a series of five two-digit numbers. As in the previous task, you can choose which

payment scheme you would prefer to apply to your performance. As in the previous exercise, you will be asked which Robot you are most similar to.

If Task 5 is the one randomly selected for payment, then your earnings for this task are determined the same way as in Task 4. You can choose between:

- Piece Rate: 50 cents per problem you solve correctly.
- Tournament: €2 per correct answer if you correctly solve more problems than your group members did in Task 2, €0 otherwise.

In addition, you may earn €2 for the Robot part. For this part, we will ask you which Robot is most similar to you in your opinion. In other words, which of the 100 Bob clones has the same probability as you to solve more exercises than your group members in Task 2.

The only new part about Task 5 is that you will find out whether your score was higher than the score of your team members in Task 2, even if you choose the piece rate.

You will not be informed of how you did in the tournament until all six tasks have been completed. If there are ties the winner will be randomly determined.

Please do not talk to one another. If you have any questions, please raise your hand.

[*Wait screen with Task 5 instructions; subsequent feedback and task identical to Task 4.*]

Task 6: Table

In Task 6 you will make 20 decisions. For each of these decisions, you will be choosing between a certain payment (option A) and a lottery (option B). Option A is identical for every decision problem: you will receive X Euro for certain. For option B you will receive either 4X Euro or 0 Euro. The probability with which you will receive 4X Euro will differ for every decision problem.[22]

If Task 6 is the one randomly selected for payment, then your earnings for this task are determined in the following way:

- First, one of your 20 decisions will be chosen at random. For this purpose, one participant in the experiment will be asked to roll a 20-sided die. If, for example, the number is 14, your 14th decision will be chosen for payment.
- In case you chose the certain payment (option A) on the selected decision problem, you will receive X Euro. In case you chose the lottery (option B), the computer will draw a random number from 0 to 100. If the randomly chosen number is smaller or equal to the probability of receiving 4X Euro in the selected decision problem, you will receive the 4X Euro. Otherwise, you will receive 0 Euro. Please have a look at the two examples below.

22. X equals 50 cents times the performance of the participant in Task 2.

Example 1:

Decision problem 17 was selected for payment. You chose the lottery (option B).

Option A	Option B
X Euro	20% chance of obtaining 4X Euro 80% chance of obtaining 0 Euro

In this example, a random number between 1 and 20 would yield a payment of 4X Euro, a higher number would give you a payment of 0 Euro.

Example 2:

Decision problem 12 was selected for payment. You chose the lottery (option B).

Option A	Option B
X Euro	40% chance of obtaining 4X Euro 60% chance of obtaining 0 Euro

In this example, a random number between 1 and 40 would yield a payment of 4X Euro, a higher number would give you a payment of 0 Euro.

Please do not talk to one another. If you have any questions, please raise your hand.

Für jede der unteren Zeilen wählen Sie bitte zwischen Option A und B.

Entscheidungsproblem	Option A		Option B	
1	○	4,00 € zu erhalten	**100%** Chance 16,00 € zu erhalten; **0%** Chance 0 € zu erhalten	○
2	○	4,00 € zu erhalten	**95%** Chance 16,00 € zu erhalten; **5%** Chance 0 € zu erhalten	○
3	○	4,00 € zu erhalten	**90%** Chance 16,00 € zu erhalten; **10%** Chance 0 € zu erhalten	○
4	○	4,00 € zu erhalten	**85%** Chance 16,00 € zu erhalten; **15%** Chance 0 € zu erhalten	○
5	○	4,00 € zu erhalten	**80%** Chance 16,00 € zu erhalten; **20%** Chance 0 € zu erhalten	○
6	○	4,00 € zu erhalten	**75%** Chance 16,00 € zu erhalten; **25%** Chance 0 € zu erhalten	○
7	○	4,00 € zu erhalten	**70%** Chance 16,00 € zu erhalten; **30%** Chance 0 € zu erhalten	○
8	○	4,00 € zu erhalten	**65%** Chance 16,00 € zu erhalten; **35%** Chance 0 € zu erhalten	○
9	○	4,00 € zu erhalten	**60%** Chance 16,00 € zu erhalten; **40%** Chance 0 € zu erhalten	○
10	○	4,00 € zu erhalten	**55%** Chance 16,00 € zu erhalten; **45%** Chance 0 € zu erhalten	○
11	○	4,00 € zu erhalten	**50%** Chance 16,00 € zu erhalten; **50%** Chance 0 € zu erhalten	○
12	○	4,00 € zu erhalten	**45%** Chance 16,00 € zu erhalten; **55%** Chance 0 € zu erhalten	○
13	○	4,00 € zu erhalten	**40%** Chance 16,00 € zu erhalten; **60%** Chance 0 € zu erhalten	○
14	○	4,00 € zu erhalten	**35%** Chance 16,00 € zu erhalten; **65%** Chance 0 € zu erhalten	○
15	○	4,00 € zu erhalten	**30%** Chance 16,00 € zu erhalten; **70%** Chance 0 € zu erhalten	○
16	○	4,00 € zu erhalten	**25%** Chance 16,00 € zu erhalten; **75%** Chance 0 € zu erhalten	○
17	○	4,00 € zu erhalten	**20%** Chance 16,00 € zu erhalten; **80%** Chance 0 € zu erhalten	○
18	○	4,00 € zu erhalten	**15%** Chance 16,00 € zu erhalten; **85%** Chance 0 € zu erhalten	○
19	○	4,00 € zu erhalten	**10%** Chance 16,00 € zu erhalten; **90%** Chance 0 € zu erhalten	○
20	○	4,00 € zu erhalten	**5%** Chance 16,00 € zu erhalten; **95%** Chance 0 € zu erhalten	○
Abschicken				

FIGURE XIV: SCREEN SHOT OF TASK 6: CHOICE TABLE (GERMAN)

Overview of Payment

Task 4 was randomly chosen for payment.During this Task, you solved 10 exercises and chose the tournament, and won it. You therefore receive 20€.

In addition, in Task 4 you chose Bob 52. Your randomly chosen Bob was 34. We therefore used your performance to determine your earnings.

You therefore earned 2€ for the robot part of Task 4.

In addition to the show-up fee of 5€ and the 3€ participation fee, you therefore earned a total of 30€.[23]

Questionnaire

Please answer the following questions:

- What is your gender?
- What is your age?
- What is your major?
- Which finger on your right hand is longer, the index-finger or ring-finger?
- Which finger on your left hand is longer, the index-finger or ring-finger?

[*Page 2*]

Please answer the following questions:

- How do you see yourself: are you in general a person who is fully prepared to take risks or do you try to avoid taking risks?
- How do you see yourself: are you in driving a person who is fully prepared to take risks or do you try to avoid taking risks?
- How do you see yourself: are you in financial matters a person who is fully prepared to take risks or do you try to avoid taking risks?
- How do you see yourself: are you in your free time and in sports a person who is fully prepared to take risks or do you try to avoid taking risks?
- How do you see yourself: are you in your professional career a person who is fully prepared to take risks or do you try to avoid taking risks?
- How do you see yourself: are you in terms of your personal health a person who is fully prepared to take risks or do you try to avoid taking risks?

23. The choices, payment amounts and selected task presented here are an example.

[*Page 3*]

For each of the following items, please choose either A or B. One of the items will be randomly selected for payment at the end of the experiment.

Item	Option A	Option B
1	1/10 Chance of obtaining 1.00€ 9/10 Chance of obtaining 0.80€	1/10 Chance of obtaining 1.90€ 9/10 Chance of obtaining 0.10€
2	2/10 Chance of obtaining 1.00€ 8/10 Chance of obtaining 0.80€	2/10 Chance of obtaining 1.90€ 8/10 Chance of obtaining 0.10€
3	3/10 Chance of obtaining 1.00€ 7/10 Chance of obtaining 0.80€	3/10 Chance of obtaining 1.90€ 7/10 Chance of obtaining 0.10€
4	4/10 Chance of obtaining 1.00€ 6/10 Chance of obtaining 0.80€	4/10 Chance of obtaining 1.90€ 6/10 Chance of obtaining 0.10€
5	5/10 Chance of obtaining 1.00€ 5/10 Chance of obtaining 0.80€	5/10 Chance of obtaining 1.90€ 5/10 Chance of obtaining 0.10€
6	6/10 Chance of obtaining 1.00€ 4/10 Chance of obtaining 0.80€	6/10 Chance of obtaining 1.90€ 4/10 Chance of obtaining 0.10€
7	7/10 Chance of obtaining 1.00€ 3/10 Chance of obtaining 0.80€	7/10 Chance of obtaining 1.90€ 3/10 Chance of obtaining 0.10€
8	8/10 Chance of obtaining 1.00€ 2/10 Chance of obtaining 0.80€	8/10 Chance of obtaining 1.90€ 2/10 Chance of obtaining 0.10€
9	9/10 Chance of obtaining 1.00€ 1/10 Chance of obtaining 0.80€	9/10 Chance of obtaining 1.90€ 1/10 Chance of obtaining 0.10€
10	10/10 Chance of obtaining 1.00€ 0/10 Chance of obtaining 0.80€	10/10 Chance of obtaining 1.90€ 1/10 Chance of obtaining 0.10€

[*Page 4*]

Please choose one of the following lotteries. Your chosen lottery will be paid out to you at the end of the experiment.

Item	Option
1	50% Chance of obtaining 1.40€
	50% Chance of obtaining 1.40€
2	50% Chance of obtaining 1.20€
	50% Chance of obtaining 1.80€
3	50% Chance of obtaining 1.00€
	50% Chance of obtaining 2.20€
4	50% Chance of obtaining 0.80€
	50% Chance of obtaining 2.60€
5	50% Chance of obtaining 0.60€
	50% Chance of obtaining 3.00€
6	50% Chance of obtaining 0.10€
	50% Chance of obtaining 3.50€

Overview

In the experiment you already earned 22€. In the questionnaire, you earned an additional 1.80€.

Including the show-up fee of 5€ and the participation fee of 3€, you have therefore earned a total of 31.80€

Thanks for taking part in this experiment. Please enter your total payment on the receipt on your desk, and raise your hand when ready. The experimenter will then come to your desk to check the total amount, after which payment will be done in the office next door.